ISSN 1847-2001 (Print)
ISSN 1848-2295 (Online)

UNIVERSITY OF ZAGREB
FACULTY OF ORGANIZATION AND INFORMATICS
VARAŽDIN

Central European CONFERENCE on Information and Intelligent Systems

Proceedings

37th International Scientific Conference

CECIIS

Organized by:
University of Zagreb
Faculty of Organization and Informatics

September 16-18/2026
Varaždin, Croatia

ISSN 1847-2001 (Print)
ISSN 1848-2295 (Online)

UDC 004:007

37th International Scientific Conference

Central European Conference on Information and Intelligent Systems

CECIIS

Proceedings

September 16th – 18th, 2026
Varaždin, Croatia

Organized by:
University of Zagreb
Faculty of Organization and Informatics

Publisher
University of Zagreb Faculty of Organization and Informatics,
Pavlinska 2, HR-42000 Varaždin, Croatia

Contact
E-mail: ceciis@foi.unizg.hr
Phone: +385 42 390 835
Fax: +385 42 213 413
Web: ceciis.foi.hr
Fb: facebook.com/ceciis.foi

For Publisher
Marina Klačmer Čalopa, University of Zagreb Faculty of Organization and Informatics, Croatia

Editors
Nina Begičević Ređep, University of Zagreb Faculty of Organization and Informatics, Croatia
Bogdan Okreša Đurić, University of Zagreb Faculty of Organization and Informatics, Croatia

Editorial Board
Victoria Appatova, University of Cincinnati, USA
Igor Balaban, University of Zagreb, Croatia
Neven Vrček, University of Zagreb, Croatia
Marko Bohanec, Jozef Stefan Institute, Slovenia
Aruna Chandra, Indiana State University, USA
Luis de Marcos Ortega, University of Alcalá, Spain
Blaženka Divjak, University of Zagreb, Croatia
Jasminka Dobša, University of Zagreb, Croatia
Juraj Dud'ak, Slovak University of Technology in Bratislava, Slovakia
Matjaž Garms, Jozef Stefan Institute, Slovenia
Gabriel Gaspar, Slovak University of Technology in Bratislava, Slovakia
Gordan Gledec, University of Zagreb, Croatia
Marcelo Fabian Maina, Open University of Catalonia, Spain
Marjan Heričko, University of Maribor, Slovenia
Dieter Hertweck, Herman Hollerith Research Centre, Germany
Nikola Kadoić, University of Zagreb, Croatia
Marina Klačmer Čalopa, University of Zagreb, Croatia
Michal Kohani, University of Žilina, Slovakia
Mario Konecki, University of Zagreb, Croatia
Vladimir Kovšca, University of Zagreb, Croatia
Marjan Krašna, University of Maribor, Slovenia
Sandra Lovrenčić, University of Zagreb, Croatia
Ivan Luković, University of Belgrade, Serbia
Ivan Magdalenić, University of Zagreb, Croatia
Aleksandar Marković, University of Belgrade, Serbia
Peter Marton, University of Žilina, Slovakia
Marta Mas Machuca, UIC Barcelona, Spain
Marko Mijač, University of Zagreb, Croatia
Dunja Mladenić, Jožef Stefan Institute, Slovenia
Oliver Moravčik, Slovak University of Technology, Slovakia
Enrique Mu, Carlow University, USA
Karlo Nyarko, University of Osijek, Croatia
Jan Paralič, Technical University of Košice, Slovakia
Patrizia Poščić, University of Rijeka, Croatia
Elisabeth Pergler, CAMPUS 02 - University of Applied Sciences, Austria
Sonja Ristić, University of Novi Sad, Serbia
Markus Schatten, University of Zagreb, Croatia
Zlatko Stapić, University of Zagreb, Croatia
William Steingartner, Technical University of Kosice, Slovakia
Vjeran Strahonja, University of Zagreb, Croatia
Diana Šimić, University of Zagreb, Croatia
Martina Tomičić Furjan, University of Zagreb, Croatia
Katarina Tomičić-Pupek, University of Zagreb, Croatia
Jan Tretmans, Radboud University Nijmegen, Netherlands
Wim Van Petegem, KU Lueven, Belgium
Jaume Jordán Prunera, Universitat Politècnica de València
Violeta Vidaček-Hainš, University of Zagreb, Croatia
Sirje Virkus, Tallinn University, Estonia
Mladen Vouk, North Carolina State University, USA
Marko Tadić, University of Zagreb, Croatia
Lode Lauwaert, KU Lueven, Belgium
Violeta Cvetkoska, University in Skopje, Macedonia
Krzysztof Podlaski, University of Lodz, Poland

Publishing Board
Marko Matus, University of Zagreb Faculty of Organization and Informatics, Croatia
Dijana Peras, University of Zagreb Faculty of Organization and Informatics, Croatia
Bernarda Kos, University of Zagreb Faculty of Organization and Informatics, Croatia
Lea Masnec, University of Zagreb Faculty of Organization and Informatics, Croatia

All rights reserved. No parts of this publication may be reproduced, stored in a retrieval system or transmitted in any form by any means, electronic, mechanical photocopying, recording or otherwise without the prior permission of the copyright owner.
CECIIS Proceedings have been indexed in the following databases: ProQuest (SciTech Premium Collection, ProQuest Technology Collection and ProQuest Advanced Technologies & Aerospace database) and Web of Science.

ISSN 1847-2001 (Print)
ISSN 1848-2295 (Online)

Creative Commons Licence
Central European Conference on Information and Intelligent Systems Proceedings is licensed under a Creative Commons Attribution-NonCommercialNoDerivatives 4.0 International License.
http://creativecommons.org/licenses/by-nc-nd/4.0/

Contents

Index of Authors	VII
List of Reviewers	IX
Preface	XI

Computer Games

Tomislav Peharda, Predrag Peharda **Combining Real-World Maps with Generative AI for Open-World Games**	3
Bogdan Okreša Đurić, Marko Kir **Towards An Ontology-Driven Framework for Generated Game Assets**	9
Bogdan Okreša Đurić, Lovro Đurđević **Where Benchmark Intrusion Detection Meets Cloud Game Streaming: Two Generalisation Gaps in a Deep-Learning Detector**	17
Predrag Peharda, David Slavik **A Framework for Genre-Aware Player Adaptation in LLM-Driven Game AI**	25
Markus Schatten, Marinela Schatten, Elvira Gotal **Conversing with Judita and Werther: Gamified Cognitive Agents for Croatian-Language Literature Learning**	31
Markus Schatten **EduGameGen: A Framework for Educational Game Generation with Large Language Models**	39

Cyber-Physical and Embedded Systems

Maroš Valášek, Tomáš Ranuša, Gabriel Gašpar, Roman Budjač **Performance Evaluation of Python Web Frameworks for Resource-Constrained Embedded Systems**	49
Gabriela Krizanova, Michal Kopcek, Roman Pikna **Behavioural model of a bulk material hopper for virtual commissioning**	57
Zvonimir Šimić, Danijel Topić, Marina Dubravac, Lucija Gujić **Flexibility in the low voltage PV-rich distribution network with demand response and BESS**	65
Tadej Šeruga, Igor Bernik **Cyber Security and Resilience of Artificial Intelligence Systems in the Nuclear Sector: A Literature Review**	73
Kolok Peter **A Pulse-Based Semiconductor Reverse Breakdown Voltage Meter Using Inductive Flyback Energy Transfer**	83
Marek Tebeľák, Lukáš Čechovič, Ondrej Karpiš, Ján Kapitulík **Predictive Control Algorithm for Pneumatic Positioning Cyber-Physical Systems Using Standard ON/OFF Valves**	91
Thi Kim Anh Vo **Memory-Optimized Persistent Sheaf Laplacians for Real-Time Biomedical Signal Anomaly Detection on Resource-Scarce Platforms**	99
Andrej Pavešić, Renata Mekovec **Mobile application development: securing user data**	113

Digital Transformation and Corporate Culture

Katarina Kostelić, Cristina Fleșeriu, Vlad Bocăneț
From Artificial Intelligence as Automation to Generative AI as a Work Tool: A Text Network Analysis of Public Discourse Before and After ChatGPT 123

Antun Biloš, Berislav Andrlić, Vladimir Ristanović
Booking Platform, Generative AI, or Open Web? An Experimental Comparison of User Experience in Tourism Accommodation Search 131

Anamarija Mladinic
Designing Olivia AI: User Expectations for a Trustworthy GDPR Compliance Assistant 141

Antonija Srok, Rikardo Protrka
Cybersecurity and Patient Data Protection in Healthcare: Research Trends and Organisational Challenges in the Context of NIS2 149

Matija Šajn
Generative Artificial Intelligence in Business Process Modeling: A Systematic Literature Review of Applications, Opportunities, and Limitations 157

Innovation, Entrepreneurship and Economic Growth

Nikolina Dreven, Lorena Pikl
Generative AI as a Job Resource and Demand: A JD–R Perspective on Work Engagement and Extra-role Behavior 167

Antonia Kurtović, Nora Miškulin, Vitomir Pavičić
The Invisible Workflow: A Cross-Stakeholder Study of Early-Stage Startup Tracking 177

Stjepan Mihalic, Tomislav Bakovic, Josip Mihalic
The Critical Steps for Using AI: A Case Study from a Manufacturing Company 189

Intelligent Systems

Ayşenur Elif Abanoz, Sevgi Koyuncu Tunç, Ece Çağlayan, Şemsettin Ekin Işık, Lütfiye Dila Çalık, Ahmet Burak Albayrak, Murat Tut, Utku Eyerci, Ayberk Özel, İlke Durmaz, İbrahim Sefa Sağlam
A Systematic Mapping of Paradigms, Devices, and Reproducibility Practices 201

Luka Lukač, Damjan Strnad
The Impact of Quantisation on Performance of Generative Adversarial Networks 211

Marko Horvat, Sven Leo Kupres, Marko Delimar
Assessing the importance of knowledge graphs for improving LLM answer accuracy using data augmentation 217

Lazerta Rista
Language-Aware Evaluation of AI-Driven Search: A Benchmark Study of Asymmetry in a Low-Resource Language 227

Dino Vlahek, Livija Ružić
Machine Learning for Subject Classification of Books Using Open Bibliographic Metadata 235

Antonia Šarčević, Alan Jović, Jakša Vukojević, Nikolina Frid
Cross-Method Analysis of Explainability in EEG-Based Major Depressive Disorder Classification 245

Darko Andročec
Vibe Coding and Web Application Security: A Twin-Prompt Study 253

Oliver Lohaj, Vladyslav Stakhov
Development of a Browser Extension for Toxic Content Detection Using an NLP Model 261

Kolok Peter
Correlation-Based Reduction of Environmental Input Variables for Efficient GRU-Based Prediction in IoT Systems 269

Hannah Hansa, Hanlie Smuts
Responsible AI Usage to Improve Knowledge Management For Sustainability Outcomes 279

Hanlie Smuts, Blossom Alloziem, Catherine Dyason, Keiara Govender
Preserving Human Agency in AI-Driven Interactions: The Role of User Experience Design 289

Dominik Čambál, Monika Herchlová, Juraj Ďuďák
A Comparative Study of SNN Learning Paradigms in Hybrid Graph-Spiking Neural Networks for Biomedical Data Analysis 297

Georgi Trajkov, Marko Grobelnik, Dunja Mladenić
From News Text to Causal Infographics via SemanticTriplets: Making an Opaque Image Pipeline Inspectable 305

Marko Horvat, Sven Leo Kupres
Evaluating AI-Generated Code: A Taxonomy and Comparative Catalog of Vibe-Coding Benchmarks 313

Goran Bubaš, Andreja Kovačić
Using Mind Maps to Investigate the Typology of Students' Use of Large Language Models: A Preliminary Study 325

David Slavik, Tim Juić
A Hybrid Document Reading, Few-Shot Classification, and LLM Extraction Pipeline for Accounting Document Processing 337

Learning Analytics, AI and Technologies in Education

Sam Sabah, Alexander Tillmann, David Weiß, Hendrik Drachsler
When the Map Becomes a Place: Immersive Virtual RealityTerrain Exploration with GeoVR 347

Sirje Virkus, Muhammad Shaheer Alam
Knowing but Not Doing: The Awareness-Practice Gap in Responsible and Green AI Use 355

Jasenka Begić, Tihana Banko, Ana Lokas Ćošković
Perceptual Indicators for Teaching Quality: Transforming SET Data into Learning Analytics Tools for Adult Online Instructors 365

Branislav Fecko, Jozef Dziak, Tibor Vince, Ján Molnár
Conversational AI for Electrical Measurement Education: An Open-Source LLM Framework 375

Nikola Kadoić, Blaženka Divjak, Barbi Svetec, Thomas Penniston, Mirza Žižak, Dino Vlahek
Review of Learning Outcomes Prioritisation Approaches 385

Sabina Rako
Students' Perceived Usefulness of Prompts in Blended Learning: Two Pilot Studies 395

Goran Hajdin
Pedagogical Friction in ICT-Integrated K-12 Classrooms: Cognitive Offloading, Technostress, and Shallow 401

Neven Pintarić, Ivan Peraić
Evaluation of Student Assignments Using Local Large Language Models and Graph-Based Evidence 411

Vjeran Bušelić, Dinko Horvat
Schema Before Output: Metacognitive Schema-First Scaffolding in Computing Higher Education in the GenAI Era 423

Vjeran Bušelić, Dinko Horvat
Skills Never Formed: AI Adoption and Career Formation Risk in IT Higher Education 435

Michael Eichhorn, Julia Schmitt, Alexander Tillmann, Josmario Albuquerque, Paul Astles, Catriona Matthews, Bart Rienties, Blazenka Divjak, Tanja Šestanj-Perić, Egle Gedrimiene, Hanni Muukkonen, Mirza Žižak
Exploring Authentic Learning Scenarios in Online and Blended Environments Across European Universities 445

Hanni Muukkonen, Egle Gedrimiene, Blaženka Divjak, Barbi Svetec, Julia Schmitt, Bart Rienties, Michael Eichhorn, Alexander Tillmann
Synergies of learning design and analytics: Design patterns of higher education courses 453

Strategic Planning and Decision Making

Matea Matic Sosic
Social Influence and e-WOM in Digital Environments: Examining Purchase and Recommendation Intentions through Social Media 465

Bisera Andrić Gušavac, Mihailo Vučeljić, Milena Popović, Biljana Panić, Dragan Pamučar
Application of Business Analytics in Demand Planning: An Example from the Fast Food Industry 475

Aleksandar-Jon Stefan, Artur Bjelica
Harnessing friction topology for stage-sensitive checkoutdiagnostics in supplement e-commerce 483

Klara Šilec, Violeta Vidaček-Hainš, Zrinka Lacković Vincek
Virtual Collaboration in Higher Education: Economic and Competence Implications 491

Milanka Dimovska, Violeta Cvetkoska
Artificial Intelligence in the Workplace: A Threat or an Opportunity for Better Performance? Insights from The Republic of North Macedonia 501

Predrag Trpeski, Violeta Cvetkoska, Filip Peovski, Marijana Cvetanoska, Gunter Merdzan, Kristijan Kozheski
Firms' Workforce Planning Orientations Under Structural Skill Shortages in North Macedonia 507

Marija Gombar, Marija Boban
Not All Technostress Is Equal: Techno-Complexity, Perceived Ease of Use, and Developers' Acceptance of Emerging Development Technologies 515

Peter Marton
GIS Support for Managerial Decision-Making in Enhancing Cycling Tourism Routes 523

Krunoslav Matešić, Mirko Palić, Sabina Sačer
Age Differences In The Effectiveness Of Personality-tailored Ads In Social Networks: Evidence from Croatia 531

Violeta Cvetkoska, Igor Ivanovski, Filip Peovski, Bojan Kitanovikj, Ivona Serafimovska
Circular Migration without Return Infrastructure: Policy and Social-Partner Sensemaking of Skill Shortages, Brain Drain and Labour-Market Governance in a Developing Context 539

Katarina Pažur Aničić, Vjeran Bušelić
Career Paths of IT Project Managers: Exploring Generic Competencies and Progression Patterns 549

Petra Vondra, Nina Begičević Ređep, Tomislav Peharda
A Conceptual Framework for AI-Supported Decision Making in Pre-Incubation Programmes 559

Strategic Planning and Decision Making

Christina Nilsson
When AI Targets Core Tasks: Human Actors and Role Repositioning in Intelligent Documentation Systems 571

Zahra Bibi, Nida Fatima, Rubina Khadim, Iftakhar Rasheed
Privacy Preserving Federated Learning withDifferential Privacy for DDoS Intrusion Detectionin Smart City IoT Networks 581

Lovro Posarić, Tara Nikolić
Towards Autonomous Resilience: Architecting Agentic Middleware for the Cloud-Edge Continuum 589

Štefan Horvat
Evaluating event influence on TGN performance in bike sharing station occupancy forecasting 597

Dino Vlahek, Sandra Lovrenčić, Dijana Oreški, Nina Begičević Ređep, Tena Jagačić, Renato Punčec, Hrvoje Aničić, Marko Mrkonjić
From industrial data to business decisions: Industrial case studies as teaching instruments in interdisciplinary analytics education 603

Hana Ivančić, Tomislav Jagušt
An AI Companion for Self-Regulated Learning: Integrating Conversational AI, Predictive Analytics, and Explainability 615

Claudia-Maria Miu, Delia-Elena Vranceanu, Andreea-Cristina Stoica
Artificial Intelligence Disclosure and Magical Perception:Effects on Consumer Trust in Chatbot Interactions 623

Trends in Software Engineering

Melita Draganić, Robertina Zdjelar, Velibor Božić
The Role of Artificial Intelligence in Objectifying the CMM Model for Assessing the Maturity of Information Systems 633

Maja Benkus, Neven Vrček
The Adoption of Vibe Coding and the Transformation of Programming Practices Induced by Generative Models: A SonarQube-Based Analysis 643

Sandra Melovic, Dusanka Dakic, Darko Stefanovic, Teodora Vuckovic
SwiftUI Adoption in Enterprise Mobile Development: Evaluating the Impact of Incremental UIKit to SwiftUI Migration on Presentation-Layer Complexity 651

Matúš Nečas, Juraj Ďuďák, Roman Budjač
A State-Indexed Execution Model for Embedded Systems 659

Tara Nikolić, Dario Ljubas, Goran Nikšić
Detection without disposition: Security technical debt, scanner-context artefacts, and the triage gap in SME GitHub repositories 667

Publication Ethics And Malpractice Statement 678

Author Guidelines 682

CECIIS 2026 Sponsors 696

CECIIS 2026 Program and Organizing Committees' Members, Research, Professional Tracks' Chairs and Student Poster and Presentation Section Chair 704

Index of Authors

Abanoz, Ayşenur Elif 201
Alam, Muhammad Shaheer 355
Albayrak, Ahmet Burak 201
Albuquerque, Josmario 445
Alloziem, Blossom 289
Andrić Gušavac, Bisera 475
Andrlić, Berislav 131
Andročec, Darko 253
Aničić, Hrvoje 603
Astles, Paul 445
Bakovic, Tomislav 189
Banko, Tihana 365
Begić, Jasenka 365
Ređep, Nina Begičević 559, 603
Benkus, Maja 643
Bernik, Igor 73
Bibi, Zahra 581
Biloš, Antun 131
Bjelica, Artur 483
Boban, Marija 515
Bocăneț, Vlad 123
Božić, Velibor 633
Bubaš, Goran 325
Budjač, Roman 49, 659
Bušelić, Vjeran 423, 435, 549
Çağlayan, Ece 201
Çalık, Lütfiye Dila 201
Čambál, Dominik 297
Čechovič, Lukáš 91
Cvetanoska, Marijana 507
Cvetkoska, Violeta 501, 507, 539
Dakic, Dusanka 651
Delimar, Marko 217
Dimovska, Milanka 501
Divjak, Blaženka 385, 453
Divjak, Blazenka 445
Drachsler, Hendrik 347
Draganić, Melita 633
Dreven, Nikolina 167
Dubravac, Marina 65
Ďuďák, Juraj 297, 659
Durmaz, İlke 201
Dyason, Catherine 289
Dziak, Jozef 375
Eichhorn, Michael 445, 453
Eyerci, Utku 201
Fatima, Nida 581
Fecko, Branislav 375
Fleșeriu, Cristina 123
Frid, Nikolina 245
Gašpar, Gabriel 49
Gedrimiene, Egle 445, 453
Gombar, Marija 515
Gotal, Elvira 31
Govender, Keiara 289
Grobelnik, Marko 305
Gujić, Lucija 65
Hajdin, Goran 401
Hansa, Hannah 279
Herchlová, Monika 297
Horvat, Dinko 423, 435
Horvat, Marko 217, 313
Horvat, Štefan 597
Işık, Şemsettin Ekin 201
Ivančić, Hana 615
Ivanovski, Igor 539
Jagačić, Tena 603
Jagušt, Tomislav 615
Jović, Alan 245
Juić, Tim 337
Kadoić, Nikola 385
Kapitulík, Ján 91
Karpiš, Ondrej 91
Khadim, Rubina 581
Kir, Marko 9
Kitanovikj, Bojan 539
Kopcek, Michal 57
Kostelić, Katarina 123
Kovačić, Andreja 325
Koyuncu Tunç, Sevgi 201
Kozheski, Kristijan 507
Krizanova, Gabriela 57
Kupres, Sven Leo 217, 313
Kurtović, Antonia 177
Lacković Vincek, Zrinka 491
Ljubas, Dario 667
Lohaj, Oliver 261
Lokas Ćošković, Ana 365
Lovrenčić, Sandra 603
Lukač, Luka 211
Marton, Peter 523
Matešić, Krunoslav 531
Matic Sosic, Matea 465
Matthews, Catriona 445
Mekovec, Renata 113
Melovic, Sandra 651
Merdzan, Gunter 507
Mihalic, Josip 189
Mihalic, Stjepan 189
Miškulin, Nora 177
Miu, Claudia-Maria 623
Mladenić, Dunja 305
Mladinic, Anamarija 141
Molnár, Ján 375
Mrkonjić, Marko 603
Muukkonen, Hanni 445, 453
Nečas, Matúš 659
Nikolić, Tara 589, 667
Nikšić, Goran 667
Nilsson, Christina 571
Okreša Đurić, Bogdan 9, 17
Oreški, Dijana 603
Özel, Ayberk 201
Palić, Mirko 531
Pamučar, Dragan 475
Panić, Biljana 475
Pavešić, Andrej 113
Pavičić, Vitomir 177
Pažur Aničić, Katarina 549
Peharda, Predrag 3, 25
Peharda, Tomislav 3, 559
Penniston, Thomas 385
Peovski, Filip 507, 539
Peraić, Ivan 411
Peter, Kolok 83, 269
Pikl, Lorena 167
Pikna, Roman 57
Pintarić, Neven 411
Popović, Milena 475
Posarić, Lovro 589
Protrka, Rikardo 149
Punčec, Renato 603
Rako, Sabina 395
Ranuša, Tomáš 49
Rasheed, Iftakhar 581
Rienties, Bart 445, 453
Rista, Lazerta 227
Ristanović, Vladimir 131
Ružić, Livija 235
Sabah, Sam 347
Sačer, Sabina 531

Sağlam, İbrahim Sefa 201
Šajn, Matija 157
Šarčević, Antonia 245
Schatten, Marinela 31
Schatten, Markus 31, 39
Schmitt, Julia 445, 453
Serafimovska, Ivona 539
Šeruga, Tadej 73
Šestanj-Perić, Tanja 445
Šilec, Klara 491
Šimić, Zvonimir 65
Slavik, David 25, 337
Smuts, Hanlie 279, 289
Srok, Antonija 149
Stakhov, Vladyslav 261
Stefan, Aleksandar-Jon 483
Stefanovic, Darko 651
Stoica, Andreea-Cristina 623
Strnad, Damjan 211
Svetec, Barbi 385, 453
Tebeľák, Marek 91
Tillmann, Alexander 347, 445, 453
Topić, Danijel 65
Trajkov, Georgi 305
Trpeski, Predrag 507
Tut, Murat 201
Valášek, Maroš 49
Vidaček-Hainš, Violeta 491
Vince, Tibor 375
Virkus, Sirje 355
Vlahek, Dino 235, 385, 603
Vo, Thi Kim Anh 99
Vondra, Petra 559
Vranceanu, Delia-Elena 623
Vrček, Neven 643
Vučeljić, Mihailo 475
Vuckovic, Teodora 651
Vukojević, Jakša 245
Weiß, David 347
Zdjelar, Robertina 633
Žižak, Mirza 385, 445
Đurđević, Lovro 17

List of Reviewers

Andročec, Darko, Croatia
Appatova, Victoria, USA
Babojelić, Karlo, Croatia
Baggia, Alenka, Slovenia
Banek Zorica, Mihaela, Croatia
Barać, Dušan, Serbia
Begičević Ređep, Nina, Croatia
Beliga, Slobodan, Croatia
Benšić, Tin, Croatia
Bernik, Andrija, Croatia
Bohanec, Marko, Slovenia
Brank, Janez, Slovenia
Bubaš, Goran, Croatia
Bujan Katanec, Ivana, Croatia
Calcina, Erik, Slovenia
Cerjan, Maja, Croatia
Chandra, Aruna, USA
Cukusic, Maja, Croatia
Cvetkoska, Violeta, Macedonia
de-Marcos, Luis, Spain
Delibašić, Boris, Serbia
Detelj, Kristina, Croatia
Dolenc, Kosta, Slovenia
Domínguez-Díaz, Adrián, Spain
Dreven, Nikolina, Croatia
Dubravac, Marina, Croatia
Dudáš, Adam, Slovakia
Dundek Kokotec, Ivana, Croatia
Dunđer, Ivan, Croatia
Dvorski Lacković, Ivana, Croatia
Eichhorn, Michael, Germany
Faganel, Armand, Slovenia
Ferrer Oliva, Mikel, Spain
Flogie, Andrej, Slovenia
Gligora Marković, Maja, Croatia
Gnip, Peter, Slovakia
Grd, Petra, Croatia
Gregurec, Iva, Croatia
Guardia Ortiz, Lourdes, Spain
Gucek, Alenka, Slovenia
Hajdin, Goran, Croatia
Horvat, Marko, Croatia
Hrustek, Larisa, Croatia
Hurtuk, Ján, Slovakia
Ivković, Nikola, Croatia
Jagačić, Tena, Croatia
Janev, Valentina, Serbia
Jelenković, Leonardo, Croatia
Jeriha, Jan, Slovenia
Kadoić, Nikola, Croatia
Kirinić, Valentina, Croatia
Kitanovikj, Bojan, Macedonia
Kokot, Karolina, Croatia
Kolberg, Dennis, Germany
Koltovska-Nechoska, Daniela, Macedonia
Krašna, Marjan, Slovenia
Križanić, Snježana, Croatia
Kutnjak, Ana, Croatia
Kvet, Marek, Slovakia
Kvet, Michal, Slovakia
Lacković Vincek, Zrinka, Croatia
Lakić, Edin, Slovenia
Lohaj, Oliver, Slovakia
Lukač, Luka, Slovenia
Luković, Ivan, Serbia
Magdalenić, Ivan, Croatia
Mahlasela, Oyena, South Africa
Maina, Marcelo Fabián, Spain
Marić, Matija, Croatia
Maricic, Milica, Serbia
Markotić, Tomislav, Croatia
Marković, Aleksandar, Serbia
Martí, Pasqual, Spain
Mekovec, Renata, Croatia
Meštrović, Ana, Croatia
Mijač, Marko, Croatia
Milić, Luka, Croatia
Mitrović Dankulov, Marija, Serbia
Motamedi, Elham, Slovenia
Načinović Prskalo, Lucia, Croatia
Nikolić, Tara, Croatia
Novak, Matija, Croatia
Oreški, Dijana, Croatia
Paralič, Jan, Slovakia
Pavlina, Krešimir, Croatia
Pecina, Ena, Croatia
Peharda, Tomislav, Croatia
Pesek, Igor, Slovenia
Plantak Vukovac, Dijana, Croatia
Posarić, Lovro, Croatia
Pušnik, Maja, Slovenia
Radošević, Danijel, Croatia
Radovanović, Sandro, Serbia
Rodríguez García, Daniel, Spain
Rožanec, Jože, Slovenia
Sabol, Andrija, Croatia
Šajn, Matija, Croatia
Schatten, Markus, Croatia
Sever, Drago, Slovenia
Šimić, Zvonimir, Croatia
Sittar, Abdul, Slovenia
Strahonja, Vjeran, Croatia
Sturm, Jan, Slovenia
Sulír, Matúš, Slovakia
Szymoniak, Sabina, Poland
Taverner, Joaquin, Spain
Tomaš, Boris, Ireland
Tomičić, Igor, Croatia
Tomičić-Pupek, Katarina, Croatia
Topal, Oleksandra, Slovenia
Umek, Lan, Slovenia
Urh, Benjamin, Slovenia
Van Petegem, Wim, Belgium
Veigend, Petr, Czechia
Vidaček Hainš, Violeta, Croatia
Virkus, Sirje, Estonia
Vlahek, Dino, Croatia
Vrcek, Neven, Croatia
Žajdela Hrustek, Nikolina, Croatia
Žugec, Bojan, Croatia

Preface

It is our pleasure to present the Proceedings of the 2026 Central European Conference on Information and Intelligent Systems. We trust that this collection will prove both valuable and inspiring to its readers.

These Proceedings contain the full-text papers presented at this year's edition of the Central European Conference on Information and Intelligent Systems (CECIIS), an annual international event organised by the **University of Zagreb Faculty of Organization and Informatics, Varaždin, Croatia.** The conference was held from **16 to 18 September 2026** in Varaždin, Croatia. For more than three decades, CECIIS has served as a bridge between academics and practitioners in the field of information and intelligent systems and related disciplines.

The 37th edition of CECIIS was devoted to the theme "The Next Wave of Intelligent Systems", which invited an exploration of the transformative power of intelligent systems at a moment when technological progress is no longer incremental, but systemic and deeply intertwined with how we understand, model and act in the world. At the centre of attention were disruptive forces such as generative artificial intelligence, smart industry, digital twins, quantum computing, biotechnology and renewable energy technologies. Together, these domains are radically reshaping the paradigms of intelligent and information systems, opening new avenues for innovation and for the creation of economic, social and sustainable value. The theme also examined the synergy between people and digital technology, with particular attention to artificial intelligence as a cognitive collaborator rather than a mere instrument of automation.

Although no single event can fully capture the challenge of harnessing the coming developments in intelligent systems, CECIIS 2026 reaffirmed our shared commitment to leading change rather than merely responding to it. Through interdisciplinary collaboration, critical enquiry and bold experimentation, we are jointly shaping a future that is not to be feared, but designed, guided and put to use as a force for good.

It is also worth emphasising that the composition of the Programme and Organising Committees, bringing together academics and experts from a wide range of fields, gives the conference a genuinely cross-disciplinary character. This format steers discussion towards topics likely to gain prominence in the near future, while fostering valuable networking among individuals whose diverse expertise and collective insight form the very foundation of CECIIS.

The Proceedings of CECIIS 2026 comprise 75 rigorously reviewed scientific papers. Each paper was assessed by at least two independent reviewers and subsequently approved by the international Programme Committee. The papers are grouped into the following tracks: Computer Games; Cyber-Physical Systems and Intelligent Energy; Data and Knowledge Bases; Digital Transformation and Corporate Culture; Innovation, Entrepreneurship and Economic Growth; Intelligent Systems; Learning Analytics, AI and Technologies in Education; Strategic Planning and Decision Making; Trends in Software Engineering; and The Next Wave of Intelligent Systems. Papers were selected on the basis of their relevance, originality, applicability to the domain of information and intelligent systems, and overall scholarly merit. The official language of the conference is English.

In total, 91 papers were submitted for review. The review process was carried out by 112 reviewers, roughly half from Croatia and half from abroad, who between them completed 199 reviews, resulting in the acceptance of 74 papers. These papers, written by 180 authors from 14 countries, were presented at the conference. Following the event, a selection of papers will undergo further review with a view to extended publication in the Journal of Information and Organizational Sciences (JIOS), the international scientific journal published by the University of Zagreb Faculty of Organization and Informatics. Previous editions of the Proceedings have been indexed in the following bibliographic databases: Web of Science, ProQuest, DOAJ and EBSCO.

A defining feature of this year's conference was its distinguished invited **keynote speakers**, scholars whose contributions have shaped their respective fields of scientific and professional enquiry. CECIIS 2026 was honoured to welcome the following keynote speakers:

- Prof. Dr.-Ing. Dr. h.c. Detlef Zühlke holds an MSc in electrical engineering and computer science and a PhD in robotics, both from RWTH Aachen, Germany. Until his retirement in 2018 he was Professor of Industrial Automation at the Technical University of Kaiserslautern and Director of the Innovative Factory Systems (IFS) research department at the German Research Center for Artificial Intelligence (DFKI). From 2005 to 2019 he also served as Executive Chairman of SmartFactory KL, a non-profit research association and a worldwide pioneer in smart manufacturing. His research interests lie in industrial control architectures, factory planning and operation, and human–machine systems for industrial applications. He is widely regarded as a pioneer of the Industry 4.0 paradigm, having paved the way into this new era through numerous groundbreaking contributions.

- Jordi Linares-Pellicer is Professor of Computer Science at the Universitat Politècnica de València (UPV), with more than forty years of experience in computing and a career spanning artificial intelligence, human–computer interaction, extended reality and generative models. He leads the VertexLit research group at VRAIN (Valencian Research Institute for Artificial Intelligence) and is Academic Director of ValgrAI, the Valencian Graduate School and Research Network of Artificial Intelligence. He is recognised for bridging frontier AI research with real-world impact in healthcare, education and accessibility, and for his extensive activity as a speaker, educator and public communicator on the latest advances in AI and their implications for society, industry and education.

A number of IT companies and agencies, as partners of the Faculty, took an active part in both the conference programme and the post-conference events by organising sponsored sessions. This collaboration attests to the conference's strong ties with the IT sector and promotes an essential exchange of knowledge and experience. Such synergy between the business sector and the scientific community is central to adapting to change and meeting the challenges facing IT and beyond.

Alongside the scholarly sessions, the conference featured a section dedicated to student posters and presentations, comprising **13 student abstracts by 22 authors from 4 countries.** These contributions added depth and diversity to the intellectual landscape of the conference. The student section was further enriched by an invited lecture: Professor Hanlie Smuts of the University of Pretoria spoke on "Who Owns You Online: What Happens When You Log Out for Good".

The **Cracker Barrel Session**, held as part of the conference, provided an informal and collaborative forum for enhancing decision-making agility in the digital transformation of education within our institution, with a particular focus on the integration of AI.

CECIIS 2026 also introduced a novelty: for the first time in the conference's history, the main programme was preceded by a dedicated pre-conference day featuring two workshops. "Using Learning Analytics to Explore Learning Patterns and Paths", organised by the Laboratory for Learning Analytics at the Faculty of Organization and Informatics together with the TRUELA and SAILeR project teams, offered researchers, including doctoral students, a forum on learning patterns, learning paths and the AI-supported analysis of educational data. "Agility in Action: Decision-Making for Today's Higher Education" drew on interactive activities and real-world cases, including AI adoption, to explore agile decision-making in higher education. The Warm-up Talks provided an informal opening to the conference, a relaxed opportunity to exchange ideas before the formal sessions began, and an evening social event rounded off the day.

We extend our sincere thanks to all who contributed to the conference programme and to these Proceedings: our esteemed keynote speakers, authors, reviewers and session chairs, as well as the members of the Programme and Organising Committees, whose dedication was indispensable. The steadfast support of the University of Zagreb Faculty of Organization and

Informatics underpinned the success of the conference.

Finally, we express our profound gratitude to our respected academic and industrial partners, sponsors and donors, whose unwavering support made CECIIS 2026 such a resounding success. The financial support of our sponsors and donors, including Jadranski naftovod d.d., Oracle Hrvatska d.o.o., InfoDom d.o.o., HBOR d.d., Grad Varaždin, Varaždinska županija, SICK Mobilisis d.o.o., Financijska agencija d.d., Grad Novi Marof, Autocesta Zagreb–Macelj, Grad Čakovec, Međimurska županija, Jedinstvo Krapina d.o.o., Pilana Ratković, Općina Gornji Kneginec, Sirana Krnjak, Vinarija Furjan and Vindija, was invaluable.

In closing, it is our sincere hope that CECIIS 2026 offered its participants an enriching and productive experience, and that readers will find these Proceedings a source of relevant and engaging work. The lasting connections forged with participants over the years strengthen our commitment to continuous improvement. We already look forward to the next Central European Conference on Information and Intelligent Systems, to be held in autumn 2027.

September 2026

On behalf of the Programme and Organizing Committees

Full Prof. Nina Begičević Ređep, PhD and Asst Prof. Bogdan Okreša Đurić, PhD

Editors

Computer Games

Tomislav Peharda, Predrag Peharda
Combining Real-World Maps with Generative AI for Open-World Games

Bogdan Okreša Đurić, Marko Kir
Towards An Ontology-Driven Framework for Generated Game Assets

Bogdan Okreša Đurić, Lovro Đurđević
Where Benchmark Intrusion Detection Meets Cloud Game Streaming: Two Generalisation Gaps in a Deep-Learning Detector

Predrag Peharda, David Slavik
A Framework for Genre-Aware Player Adaptation in LLM-Driven Game AI

Markus Schatten, Marinela Schatten, Elvira Gotal
Conversing with Judita and Werther: Gamified Cognitive Agents for Croatian-Language Literature Learning

Markus Schatten
EduGameGen: A Framework for Educational Game Generation with Large Language Models

Combining Real-World Maps with Generative AI for Open-World Games

Tomislav Peharda, Predrag Peharda
University of Zagreb Faculty of Organization and Informatics
Artificial Intelligence Laboratory
Pavlinska 2, 42000 Varaždin, Croatia
tpeharda@foi.hr, pehardap@gmail.com

Abstract. *Large-scale virtual worlds are typically either geographically accurate but empty, or richly generated but ungrounded, transient, and difficult to edit. This paper proposes a four-layer reference architecture combining both approaches. Geographic Information System (GIS) data provides the geographic skeleton, a semantic layer converts it into a typed graph of places, and generative Artificial Intelligence (AI) enriches it with interiors, decoration, Non-Player Characters (NPCs), and location-grounded narratives. A deployment layer targets Unreal Engine, Unity, Roblox, and Unreal Editor for Fortnite (UEFN). The architecture is positioned against concurrent work on urban digital twins, autonomous-driving simulators, and gameplay world models.*

Keywords. generative AI, GIS, OpenStreetMap, open-world games, sandbox games, world generation, large language models

1 Introduction

Building large-scale virtual worlds is one of the most labour-intensive parts of video-game development. Two approaches have developed to reduce this effort. The first uses real-world data as input: OpenStreetMap (OSM) provides building footprints and road networks (OpenStreetMap Contributors, 2004), Digital Elevation Models (DEMs) provide terrain, and orbital imagery provides textures and land cover. Tools such as Arnis (E. & contributors, 2022) convert this data into Minecraft worlds at city scale. Cesium for Unreal and Unity (Cesium GS, 2025) stream globe-scale 3D content into game engines. Microsoft Flight Simulator 2024 (Microsoft & Asobo Studio, 2024) reconstructs the entire planet from Bing Maps imagery. The second approach generates the world from scratch using generative models. Genie 3 (Google DeepMind, 2025) generates navigable environments frame-by-frame from text. World Labs Marble (World Labs, 2025) generates persistent, editable 3D scenes from text, image, video, or sketch. Meta's WorldGen (D. Wang et al., 2026) targets game engines but does not ground its output in real-world geography.

The two approaches are usually framed as alternatives. This paper treats them as complementary. The GIS approach produces accurate but empty worlds. Street layouts are correct, but buildings are empty shells with no inhabitants, quests, or gameplay logic. The generative approach produces detailed worlds that are not tied to any real location. A game set in a specific real place, with generated content and characters appropriate to that place, requires both.

This paper proposes a four-layer reference architecture for the hybrid. The narrative layer (NPCs, quests, and back-stories grounded in real geography) is the main focus. This is the area least addressed by existing work and where the player-experience impact is largest.

The remainder of the paper is organised as follows. Sec. 2 reviews related work in GIS-to-game tools, generative world models, and hybrid systems. Sec. 3 presents the four-layer architecture. Sec. 4 walks through a concrete example of generating a game world for the city of Split. Sec. 5 discusses applications and open challenges. Sec. 6 concludes.

2 Related work

The GIS-to-game pipeline is mature. Arnis (E. & contributors, 2022) is written in Rust, licensed Apache 2.0, and converts OSM and elevation data into Minecraft worlds. It has over sixteen thousand GitHub stars and shipped v2.8 in May 2026. Cesium for Unreal and Unity (Cesium GS, 2025) streams 3D Tiles, photogrammetry, terrain, and imagery into game engines, and exposes a full-scale WGS84 globe to the engine runtime. Microsoft Flight Simulator 2024 (Microsoft & Asobo Studio, 2024) uses Bing Maps and cloud-side machine learning to reconstruct the planet at consumer-game fidelity. None of these systems generate content. Cesium for Unreal has no inhabitants. Arnis has no quests. Microsoft Flight Simulator has no gameplay logic. The output is a skeleton.

The generative approach combines Large Language Models (LLMs), diffusion-based image and video

generation, and world foundation models. Genie 3 (Google DeepMind, 2025) generates 720p navigable environments frame-by-frame from a text prompt. World Labs Marble (World Labs, 2025) generates persistent, editable 3D scenes from text, image, video, or rough sketches. Sora 2 (OpenAI, 2025) was a consumer-facing version of the same approach and was discontinued in April 2026. Hunyuan3D 2.0 (Zhao et al., 2025) generates game-ready textured 3D assets at scale. None of these systems takes a real-world location as input. The output is not tied to any actual place.

A body of academic work has begun to combine the two approaches, almost entirely outside the games context. The closest existing work is UrbanWorld (Shang et al., 2024), which converts an OSM-derived urban layout into a 3D scene through a four-stage LLM and diffusion pipeline, and RAISECity (S. Wang et al., 2025), a multimodal agent framework for reality-aligned 3D world generation at city scale. Both target embodied AI and urban science rather than gameplay. In the autonomous-driving simulator world, CARLA's 0.9.16 release (CARLA Team, 2025) ships a Digital Twin Tool that auto-builds environments from OSM and integrates NVIDIA Cosmos Transfer (NVIDIA, 2025) for generative scene-style variations. This is the most direct existing example of the hybrid architecture, but serves vehicle training rather than gameplay. Xu and colleagues' scoping review (Xu et al., 2024) surveys the wider Generative AI (Gen AI)-for-urban-digital-twins space. On the gameplay side, Meta's WorldGen (D. Wang et al., 2026) uses an LLM to plan scene layout but takes text prompts as input rather than geographic data. Roblox Cube (Roblox, 2025) generates functional 3D assets, also without a GIS input. Epic's State of Unreal 2025 (Epic Games, 2025) introduced the Persona Device for player-talkable AI NPCs in UEFN, but it is not connected to any GIS input. Individual components exist on both sides, but no system targets sandbox games with grounded NPCs, quests, and narrative. Table 1 summarises the systems discussed above along five dimensions: whether the system takes real-world geographic data as input, whether it targets game engines, whether it generates NPCs, whether it generates narrative content, and whether its output is a persistent artefact that can be edited or shipped with a game build.

Another set of systems targets the NPC and narrative side of games directly. Park et al. (Park et al., 2023) demonstrated that an LLM can drive a population of agents with persistent memory, daily routines, and emergent social behaviour in a small fictional town. Commercial offerings such as Inworld AI and Convai bring similar techniques to Unreal Engine and Unity as middleware for NPC dialogue. AI Dungeon and similar systems have shown that LLMs can drive interactive narrative at run time. None of these systems condition the generated NPCs or narrative on the geographic place in which they are staged. The NPCs are either generic or tied to a hand-authored fictional setting. The architecture proposed in this paper combines these dialogue and narrative techniques with the GIS grounding from the first group of related systems.

3 Proposed architecture

This paper outlines a four-layer reference architecture, depicted in Fig. 1.

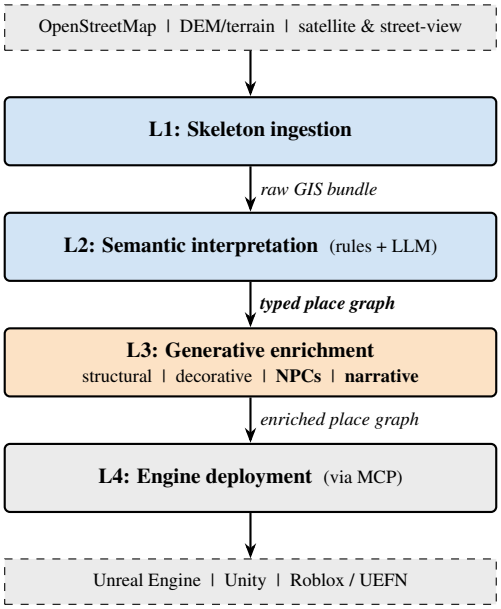


Figure 1. The four-layer reference architecture. Blue layers (L1, L2) are the deterministic GIS-grounded part; orange (L3) is the generative part; grey (L4 and I/O bands) is engine-side deployment. The L2 typed place graph is the interface between the deterministic and generative halves. Everything above it is geographic ground truth, everything below is generated.

3.1 Layer 1: Skeleton ingestion

Layer 1 ingests OSM via the Overpass API, DEMs from public sources (USGS 3DEP, Copernicus, AWS Terrain Tiles), orbital imagery (Sentinel-2, Maxar, Bing), and optionally street-view imagery for facade textures. The output is a uniformly projected and tiled raster-plus-vector bundle for the bounding region of interest. Existing libraries cover this step: Overpass-turbo for OSM queries, GDAL and rasterio for raster handling, and projects such as Arnis demonstrate that this entire stack is mature production tooling. This paper treats Layer 1 as a solved problem and as the boundary between existing tools and the proposed architecture. The architecture assumes that the input bundle contains, at minimum, a building footprint vector layer with OSM tags, a road network with classifications, a terrain raster, and a land-use raster.

Table 1. Comparison of related systems across five dimensions. ✓ indicates the system supports the feature, – indicates it does not.

System	GIS input	Targets games	Generates NPCs	Generates narrative	Persistent output
Arnis	✓	✓	–	–	✓
Cesium for Unreal	✓	✓	–	–	✓
MS Flight Sim 2024	✓	✓	–	–	✓
Genie 3	–	–	–	–	–
World Labs Marble	–	–	–	–	✓
Meta WorldGen	–	✓	–	–	✓
Roblox Cube	–	✓	–	–	✓
UEFN Persona Device	–	✓	✓	–	✓
UrbanWorld	✓	–	–	–	✓
RAISECity	✓	–	–	–	✓
CARLA + Cosmos	✓	–	–	–	✓
This paper	✓	✓	✓	✓	✓

3.2 Layer 2: Semantic interpretation

Raw GIS data is geographically accurate but has limited semantic content. An OSM `building=yes` tag indicates that a polygon is a building, but not the type of building, the era it was built in, or the activities that take place inside it. Layer 2 converts the raw data into a *typed place graph*. Nodes are typed entities such as a primary school, a residential road, or an Italian restaurant. Edges describe spatial and functional relations such as adjacency, access, or proximity. The interpretation uses rules (OSM tag schemas, land-use polygons) and an LLM that fills in gaps and resolves ambiguity.

The rule layer handles the cases where OSM tags are unambiguous: a polygon tagged `amenity=school` becomes a school node, a way tagged `highway=residential` becomes a residential road edge. The LLM layer handles the more common case of missing or ambiguous tags. For example, an unlabelled rectangular polygon next to a labelled school in a residential area is more likely a sports field than a parking lot. The LLM is prompted with the candidate polygon and its neighbourhood context, and produces a typed label with a confidence score. Labels below a confidence threshold are flagged for designer review.

The typed place graph is the interface between L2 and L3. Everything above this line is geographic ground truth, everything below is generated.

3.3 Layer 3: Generative enrichment

Layer 3 contains the generative parts of the architecture. Its outputs fall into four categories: structural, decorative, populated, and narrative.

Structural enrichment. Building interiors are generated by an LLM conditioned on the building type, era, and the player-supplied style prompt, with the constraint that the generated floor plan fit within the footprint provided by L2. The number of floors is taken from OSM `building:levels` where available and inferred from height tags or street-view imagery otherwise. Interior assets are drawn from a retrieval-augmented library (Lewis et al., 2020) of pre-authored components and supplemented by text-to-3D generation (Zhao et al., 2025) when no library entry fits. The output of this step is a per-building floor plan and asset list, attached to the corresponding place graph node.

Decorative enrichment. Vegetation is sampled from species distributions appropriate to the geographic latitude and biome of the target region, which can be derived from the terrain raster and global biome datasets. Street furniture (benches, lampposts, hydrants) is placed deterministically where OSM provides the data and probabilistically where it does not, using density priors per land-use class. Facade textures, when not derivable from street-view imagery, are generated by a diffusion model conditioned on a regional style prompt and the era tag of the building.

Populated enrichment. Each typed place node receives a population pass that instantiates one or more NPCs appropriate to its type. A school receives teachers and students. A restaurant receives staff and patrons. A residential building receives household members. The number and type of NPCs per node follows configurable density priors that can be adjusted per region or per game type. NPCs are generated with persistent identities, life-histories, and behavioural patterns, following the generative-agents approach (Park et al., 2023). The difference is that the identities are conditioned on the real place: name, history, and dialect appropriate to the region. The LLM prompt for an NPC is given both the local node context (a bakery owner in a coastal Mediterranean city) and the regional context (Adriatic coast, local dialect area), allowing consistent voice and background across all NPCs in a region.

Narrative enrichment. Quests, fetch tasks, conversations, and lore are generated against the place graph. Two types of grounding are supported. Historical grounding draws on the documented history of the area, for example a quest set in Diocletian's Palace in Split with real Roman-era context. Synthetic ground-

ing generates fiction consistent with the affordances of the place, for example a heist in a real bank or a quest taking place in a real plaza. The geographic grounding is, to our knowledge, the main novelty relative to prior LLM-driven NPC and narrative systems.

3.4 Layer 4: Engine deployment

Layer 4 targets one of four commercial engines and platforms: Unreal Engine, Unity, Roblox, and UEFN. For each target, the enriched graph is compiled into the native asset bundle and NPC runtime. The deployment is engine-specific but architecturally uniform. The interface at the L3-L4 boundary is the enriched graph itself, which is engine-agnostic. Compilation on the Unreal side uses Cesium for Unreal for the streamed globe and converts the enriched-graph assets into native Actors. On the Roblox side, the same graph compiles to Roblox Studio scripts and Cube-generated 3D assets. The UEFN Persona Device (Epic Games, 2025) and Roblox Cube (Roblox, 2025) provide engine-side runtimes for generative NPCs and assets respectively. Custom socket protocols used in earlier systems are replaced by the Model Context Protocol (MCP) (Anthropic, 2024) as the interface between models and engines. Runtime model calls, for example during a player conversation with an NPC, are routed through the same MCP interface, which allows the model provider to be swapped without engine-side changes.

4 Example scenario

This section walks through an end-to-end example: generating a small open-world game set in the historical centre of Split, Croatia. The bounding region is roughly 1 km by 1 km around Diocletian's Palace, the Roman-era core of the city. The target engine is Unreal Engine.

Layer 1. An Overpass API query retrieves all OSM features within the bounding box: roughly 500 building footprints, the road network, parks, the waterfront promenade, and tagged points of interest. DEMs are pulled from Copernicus at 10 m resolution. Sentinel-2 imagery provides the satellite raster and ESA WorldCover provides land-use classification. Optionally, Mapillary street-view imagery is queried along the road network for facade reference.

Layer 2. The interpretation step produces a typed place graph with about 500 building nodes, 100 road segment edges, and around 80 Point of Interest (POI) nodes. OSM tags carry the obvious cases directly: Diocletian's Palace is tagged `historic=palace`, the Cathedral of Saint Domnius is tagged `building=cathedral`, the Peristyle is tagged as a public square. The LLM resolves the remaining cases. Untagged polygons within the palace walls are labelled as residential or commercial buildings of the period based on their footprint and neighbourhood. Untagged polygons along the waterfront are labelled as cafes, restaurants, or shops based on the surrounding land-use class. Fig. 2 shows a small subgraph produced by this step.

Layer 3. For structural enrichment, Diocletian's Palace gets a generated layout consistent with late-Roman imperial architecture, the cathedral gets a nave-and-side-chapel layout adapted from the original mausoleum, and residential buildings inside the old town get tight apartment floor plans appropriate to Mediterranean urban housing. For decorative enrichment, the parks and palace courtyards are populated with cypress, olive, and pine trees characteristic of the Mediterranean biome, street furniture is placed along the waterfront and pedestrian zones, and facade textures are generated in the Roman-to-Venetian-revival range typical of the city centre. For populated enrichment, the cafes along the Riva receive baristas and patrons, the palace receives guides and visitors, the cathedral receives clergy and tourists, and residential buildings receive household-scale populations. NPC identities use Croatian names and references to the Chakavian dialect of the Adriatic coast. For narrative enrichment, the system generates quests that reference the documented history of the city. Diocletian's Palace was built around 305 AD as a retirement residence for the Roman emperor Diocletian. The city of Split grew within and around its walls after the Avars and Slavs destroyed nearby Salona in the seventh century. Quests can be staged around any of these periods. Smaller quests are generated for individual locations, for example a fetch task involving a fish vendor on the Riva or a conversation with a guide at the Peristyle.

Layer 4. The enriched graph is compiled into an Unreal Engine project. The globe is streamed via Cesium for Unreal, the generated assets are imported as Actors with their place-graph identity preserved as Actor metadata, and the NPC runtime is wired through the MCP interface so that conversations are served by an LLM at play time. The resulting project is a playable Unreal scene in which the player can walk through a recognisable Split and interact with NPCs whose identities and dialogue are grounded in the actual city.

Sample player experience. A player spawns on the Peristyle. The surrounding cafes are populated with NPCs drinking coffee. The player approaches one of the fish vendors on the Riva, who uses a Croatian name and refers to her family's history in the fishing trade. She offers a fetch quest involving a delivery to the cathedral. As the player walks toward the cathedral through the narrow palace streets, the geography matches the real layout, the buildings have appropriate facades, and other NPCs along the way react to the player's progress. The quest line, the NPCs the player meets, and the locations they visit are all generated. The spatial structure and the historical and cultural context match the real city. Re-running the generation with the same seed produces the same world, allowing the

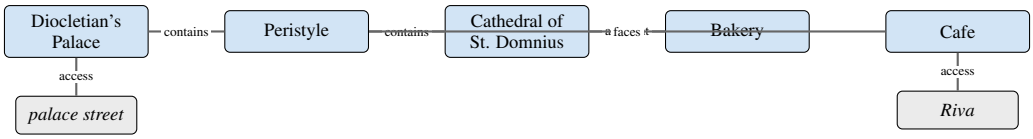


Figure 2. A subgraph of the L2 typed place graph for the Split example. Blue nodes are typed place entities; grey nodes are roads. Edges are typed relations (contains, adjacent, faces, access).

game to be shipped as a deterministic artefact. Fig. 3 summarises the flow of the example scenario across the four layers.

5 Discussion

GIS-only worlds are accurate but static and uninhabited. Purely generative worlds are detailed but not tied to any real location. A hybrid uses GIS for spatial accuracy and generative AI for content. This allows a game to be set in a real place without manual authoring of every detail. A game set in the player's hometown becomes feasible without a dedicated content team per town. A historical reconstruction (ancient Rome, medieval Dubrovnik, eighteenth-century Edinburgh) can be assembled from the same architecture by changing the era prompt and the historical sources fed into the narrative layer. Location-based games can stage encounters that fit the specific place, rather than placing generic content at arbitrary coordinates.

The hybrid keeps the editability and persistence that pixel-space generative world models lack. The enriched place graph is a typed, serialised artefact that ships with the game build. It is editable by designers and inspectable by audit tools. The compute cost is paid once per region, not per frame of play.

Several open challenges remain.

Cultural sensitivity. Generating content set in real places risks producing insulting or inaccurate content about real communities and sacred sites. Religious buildings, sites of historical trauma, and contested borders all require treatment that current LLMs are unreliable at delivering without strong guardrails. A region-specific editorial layer is a prerequisite for any deployment beyond a research prototype. For sensitive cases, involvement of local communities is also required.

Multiplayer consistency. If two players visit the same real place at different times or on different clients, the generated content should be substantially the same. This is straightforward when the generators are deterministic functions of (place, seed, version), as in the Split example above. It is harder when remote model providers introduce non-reproducible sampling, and harder still when generated content evolves over time as players interact with it. The architecture must include a content versioning mechanism that allows two clients to agree on the version of the world they are experiencing.

Scale. Generating per-building interiors at city scale is computationally prohibitive without a tiered strategy. A practical deployment would generate full-fidelity interiors only for buildings the player is near, lower-fidelity placeholders for buildings in the visible distance, and only the typed place graph node for everything further away. Smooth transitions between tiers and pre-generation of high-traffic areas are open problems.

Asset reuse economics. Generating each interior from scratch is wasteful when most buildings of a given type are similar. A more economical strategy is to compose interiors from a curated library of typed components and perturb each instance slightly. This requires a substantial library to exist before the architecture works well at scale.

Legal status. A game set in a real city raises questions about likeness rights, trademarks (a real coffee chain's logo on a real corner), and the depiction of real inhabitants by NPCs that resemble them. These are legal rather than technical questions, but they shape what the architecture can be permitted to do in a commercial product.

6 Conclusion

Sandbox and open-world games can benefit from combining the GIS skeleton approach and the generative-detail approach. The intersection has been explored for urban science and autonomous-vehicle training, but not for gameplay. This paper has outlined a four-layer reference architecture that combines GIS ingestion, semantic interpretation, generative enrichment, and engine deployment. Grounded NPCs and narratives are highlighted as the main novelty. Future work will turn the proposal into an implementation. The narrative layer is the priority, as it is the least explored. We will also conduct user studies of designer-facing and player-facing variants of the architecture.

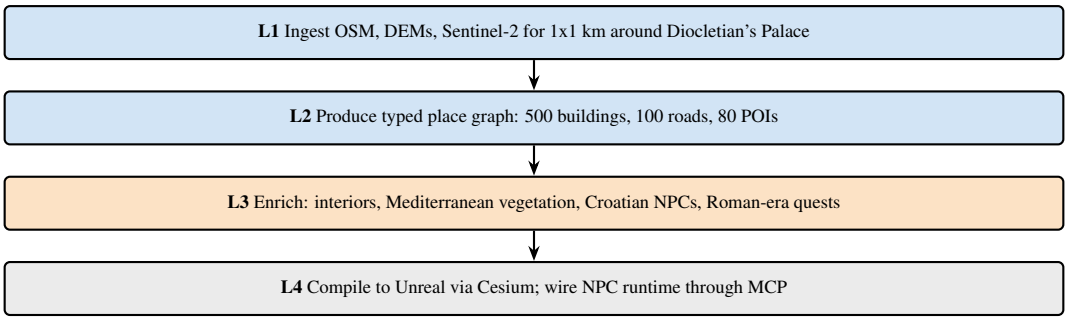


Figure 3. The four-layer pipeline applied to the Split example scenario, from raw GIS ingestion to a playable Unreal Engine build. Colours match Fig. 1.

References

Anthropic. (2024, November). Introducing the model context protocol.

CARLA Team. (2025, September). CARLA 0.9.16: Digital twin tool, NVIDIA Cosmos Transfer and NuRec integration.

Cesium GS. (2025). Cesium for Unreal and Unity: 3d geospatial streaming for game engines.

E., L., & contributors. (2022). Arnis: Generate minecraft worlds from real-world locations [Apache 2.0; uses OpenStreetMap and AWS Terrain Tiles. Accessed: 2026-06-07].

Epic Games. (2025). State of Unreal 2025: Persona device, scene graph, developer assistant.

Google DeepMind. (2025, August). Genie 3: A new frontier for world models.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, *33*, 9459–9474.

Microsoft & Asobo Studio. (2024). Microsoft flight simulator 2024.

NVIDIA. (2025). Cosmos: World foundation models for physical AI.

OpenAI. (2025, September). Sora 2 [Discontinued April 2026].

OpenStreetMap Contributors. (2004). OpenStreetMap.

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. *Proceedings of the 36th annual acm symposium on user interface software and technology*, 1–22.

Roblox. (2025, March). Introducing Roblox cube: A foundation model for 3d intelligence.

Shang, Y., Lin, Y., Zheng, Y., Fan, H., Ding, J., Feng, J., Chen, J., Tian, L., & Li, Y. (2024). Urbanworld: An urban world model for 3d city generation. *arXiv preprint arXiv:2407.11965*.

Wang, D., Jung, H., Monnier, T., Sohn, K., Zou, C., Xiang, X., Yeh, Y.-Y., Liu, D., Huang, Z., Nguyen-Phuoc, T., et al. (2026). Worldgen: From text to traversable and interactive 3d worlds. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27124–27135.

Wang, S., Zheng, Z., Shang, Y., He, L., Yu, Y., Hangyu, F., Feng, J., Liao, Q., & Li, Y. (2025). Raisecity: A multimodal agent framework for reality-aligned 3d world generation at city-scale. *arXiv preprint arXiv:2511.18005*.

World Labs. (2025, November). Marble: Generate, edit, and inhabit persistent 3d worlds.

Xu, H., Omitaomu, F., Sabri, S., Zlatanova, S., Li, X., & Song, Y. (2024). Leveraging generative ai for urban digital twins: A scoping review on the autonomous generation of urban data, scenarios, designs, and 3d city models for smart city advancement. *Urban Informatics*, *3*(1), 29.

Zhao, Z., Lai, Z., Lin, Q., Zhao, Y., Liu, H., Yang, S., Feng, Y., Yang, M., Zhang, S., Yang, X., et al. (2025). Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation. *arXiv preprint arXiv:2501.12202*.

Towards An Ontology-Driven Framework for Generated Game Assets

Bogdan Okreša Đurić, Marko Kir
University of Zagreb Faculty of Organization and Informatics
Artificial Intelligence Laboratory
Pavlinska 2, 42000 Varaždin, Croatia
dokresa@foi.unizg.hr, mkir21@student.foi.hr

Abstract. *Procedural content generation has long used ontologies as an index over a fixed library of hand-made assets. Generative models can now synthesise assets on demand, but their output is difficult to control and keep coherent. We propose a conceptual framework that treats asset realisation as a modular step: an ontology specifies what an asset must be and a realiser produces it, either by retrieving from a library or by generating it and validating the result with a vision-language model. We describe the framework, a local-first instantiation in a procedurally generated detective game, and open questions for empirical study.*

Keywords. ontology, procedural content generation, generative AI, vision-language models, neuro-symbolic, video games

1 Introduction

Building the contents of a game world by hand is a slow and expensive process, and it is often the part of game development that creates the most problems with growing scale. This is the gap that procedural content generation (PCG) has tried to close for the past two decades by writing scripts and software that produce levels, maps, items, quests, and whole worlds rather than authoring them one by one (Togelius et al., 2011). A recurring idea within that effort is to give the generator a model of the domain it is generating for, so that what it produces are not simply randomised variations, but meaningful combinations, e.g. a chest belongs in a room, a key opens a door, a witness has something to hide. Ontologies, in their knowledge-representation context, are a natural way to write such a model down, and they have been used to describe game worlds and to drive their generation (Sacco et al., 2016; Tutenel et al., 2009).

Until recently, however, the ontology in such systems did little more than point at things. A slot in the world, e.g. "a medieval cooking implement on the kitchen table", was resolved by looking the description up against the descriptions of a curated library of hand-made assets and binding the slot to whichever model matched best. The ontology was, in effect, a semantic index over a fixed catalogue. The recent rapid development of capable generative models, which can synthesise a three-dimensional object or a texture from a textual description, changes what is possible in the described scenario: the wanted model no longer needs to be sought for and hopefully found, it can be created. The difficulty is that generative models are challenging to steer. Left to themselves, they produce output that is inconsistent in style, ignores constraints of scale or function, and is rarely ready to drop into a playable level without further manual work (H. Lee et al., 2024).

This paper argues that such processes, depending on generative models, can be guided by an ontology, and proposes a conceptual framework in which asset realisation is a modular operation. The ontology specifies what an asset must be, while a realiser turns that specification into a concrete, placeable asset. A realiser may retrieve a pre-made asset from a library, which implements the classical approach as a special case, or it may generate one and then check the result against the specification using a vision-language model (VLM) before accepting it. The retrieval and generative realisers follow the same process, which is what lets one be substituted for the other and what lets the earlier approach be read as the baseline which the framework described in this paper departs from.

The conceptual framework presented here treats asset realisation as a modular step: an ontology specifies what an asset must be, and a realiser produces it, either by retrieving from a library or by generating it and validating the result against the specification with a VLM. The baseline retrieval realiser is the system built for the second author's master's thesis (Kir, 2026); the generative realiser is the extension this paper sets out. We deliberately keep the framework independent of any particular generative model, and describe a local-first instantiation, running self-hosted generation and validation, as one instance of it rather than as the contribution itself.

The rest of the paper is organised as follows. Section 2 reviews the related work across PCG, semantic game modelling, generative 3D content, vision-

language validation and neuro-symbolic control, and positions the framework against the closest systems. The framework is presented in Section 3. Section 4 describes use-case examples of retrieval and generation. Section 5 discusses what the framework covers, what remains difficult, and how it should be evaluated. The conclusions are laid out in Section 6.

2 Background and Related Work

The vocabulary this paper relies on comes from the search-based PCG taxonomy of Togelius et al. (2011), which separates what is generated, how it is represented, and how its quality is judged. Two broad families of methods can be recognised therein: constructive and search-based approaches, which build or optimise content directly, and approaches that learn to generate from examples, comprised under the name PCG via machine learning by Summerville et al. (2018). A thorough introduction to the topic of PCG is given by Shaker et al. (2016). Recent survey papers observe the rapid entry of large language models into the generation pipeline (Farrokhi Maleki & Zhao, 2024; Gallotta et al., 2024). Our framework is agnostic about which of the recognised method families plays the role of the realiser. What it provides is a semantic specification around the generator and a validation step after it.

The argument that game worlds should be backed by a semantic model, rather than by ad-hoc data structures, was made earlier by Tutenel et al. (2009), who showed that semantics improve both authoring and runtime consistency. Publications coming from the authors of the same community used Semantic Web ontologies and linked data to drive generation: Sacco et al. (2016) generate game content from ontologies and open knowledge sources, Sacco et al. (2017) give a vocabulary for describing game characters, and Barros et al. (2016) generate adventures from open data. An ontology combining role-playing-game concepts with multi-agent organisation was defined in (Okreša Đurić & Schatten, 2016) and used to drive model-driven gameplay (Schatten et al., 2017). In all of these, however, the ontology informs selection or arrangement of existing content; it does not specify content to be created or generated.

Text-to-3D generation became practical with optimisation-based methods that distil a two-dimensional diffusion prior into a three-dimensional shape, of which DreamFusion is the prime example (Poole et al., 2023). The field has since broadened considerably, as reported in the state-of-the-art report by Fan et al. (2026) and H. Lee et al. (2024). Of particular relevance is recent work explicitly targeting game-ready output, with engine-friendly topology and materials (Lei et al., 2025), since the requirement that a generated asset actually drop into a playable level is the hard part for our purposes. The use of vision-language models to steer 3D generation itself is now established: Bai et al. (2025), for instance, employ a VLM as a differentiable semantic and spatial foundation inside the generation process. Our use of a VLM can be observed to be coming after generation rather than within it, as a check against the brief rather than a training signal. These models supply the generative realiser, yet none of them carries a semantic specification of its own.

The idea of using a capable model as an evaluator was established in the "LLM-as-a-judge" setting (Zheng et al., 2023) and extended to the multimodal case, where a VLM scores visual content against stated criteria (S. Lee et al., 2024). Most directly relevant to the contents presented in this paper is the use of a VLM inside a generation loop: Singh and Zheng (2023) decompose a prompt into separate assertions, check each against the generated image with visual question answering, and refine until the assertions are satisfied. Our validation stage is of exactly this kind, with the assertions supplied by the ontology rather than by a free-text prompt.

Specifying a space of acceptable artefacts symbolically, i.e. through a definition and a list of properties, and then instantiating them, i.e. producing members of it, is an old idea in PCG; the answer-set-programming approach of Smith and Mateas (2011) is the clearest precedent, generating content that satisfies declared constraints. Combining such symbolic specification with sub-symbolic generation is the subject of neuro-symbolic AI, whose integration patterns are catalogued by Kautz (2022) and surveyed by Marra et al. (2024). Our framework is a neuro-symbolic one in the following sense: a symbolic ontology governs a sub-symbolic generator, and a VLM re-expresses the generator's output in symbolic terms so that it can be checked.

The content presented in this paper can be recognised as being near two publications in particular, even though they clearly differ. Holodeck (Yang et al., 2024) generates 3D environments from a language description by prompting a large language model for spatial constraints and then retrieving assets from a library of tens of thousands of annotated models. SceneCraft (Hu et al., 2024) synthesises a scene as engine code and, notably, uses a VLM to inspect the rendered result and refine it. Both of the approaches presented in these papers arrange or place assets rather than generating them, and both rely on the implicit semantics of a language model rather than on a formal ontology. The framework proposed here differs on two perspectives at once: the specification is an explicit ontology, and the realiser may generate the asset, with the VLM validating it against that specification.

A more recent line of work does generate assets from a structured specification rather than retrieving them, which brings it closer to the present proposal. PhysForge (Yang et al., 2026) has a vision-language model plan a hierarchical "blueprint" of material, functional, and kinematic constraints, which a physics-grounded

generator then realises into simulation-ready assets. CubePart (Zhu et al., 2026) takes a user-defined parts schema and generates game-engine-ready meshes that respect that structure without manual post-processing. Closer still to the idea proposed herein, the agentic scene generation of Liu et al. (2025) couples asset generation with a vision-language verification step that rechecks the result against a structured spatial context and regenerates when it is inconsistent.

These systems share with our framework the use of a structured specification to steer generation, and in one case a generate-then-verify loop. They differ from it, however, on the points that define our contribution: the specification is authored implicitly by a language model rather than declared as a formal ontology. Furthermore, the VLM acts as a planner or a reward signal rather than as an acceptance gate that admits or regenerates a candidate. Finally, none of them treat retrieval and generation as one interchangeable realiser, which is what lets the classical, library-bound approach be recovered as a special case of the same framework. We therefore claim novelty of the proposed framework for this specific combination, and note that, were a system to couple a formal ontology with generative assets and a validation loop in the same manner, the claim would narrow to its application within a procedurally generated detective game.

Procedurally generated mystery is itself small literature, confined largely to workshop work that generates logically consistent murderous story plots from constraints (Mohr et al., 2018; Stockdale, 2016). We draw on it for the testbed but make no narrative-generation contribution of our own.

3 The Framework

The framework rests on a single abstraction. The ontology declares slots, each of which must be filled by a concrete asset, and for each slot the system derives a realisation brief, i.e. a short structured statement of what the asset must be. A realiser is any component that takes a brief and returns a bound asset. Because of this modular and clearly defined nature of the described process, i.e. brief as an input, a bound asset as an output, different realisers can be substituted freely, and the classical retrieval approach and the generative approach proposed here become two instantiations of one pattern rather than two unrelated systems. The described structure can be seen organised in layers in Fig 1.

The knowledge layer is the ontology itself, and it is the part of the system that knows nothing about any engine, library, or generator. It holds the asset taxonomy, the properties, and attributes of each type, e.g. a lamp may be a light source and, in a particular game, a possible weapon, the spatial and connectivity rules that make a level playable, the constraints on style and scale, and the schema of the mystery. This independence is deliberate, as it lets the layers beneath it change without disturbing the model of the domain.

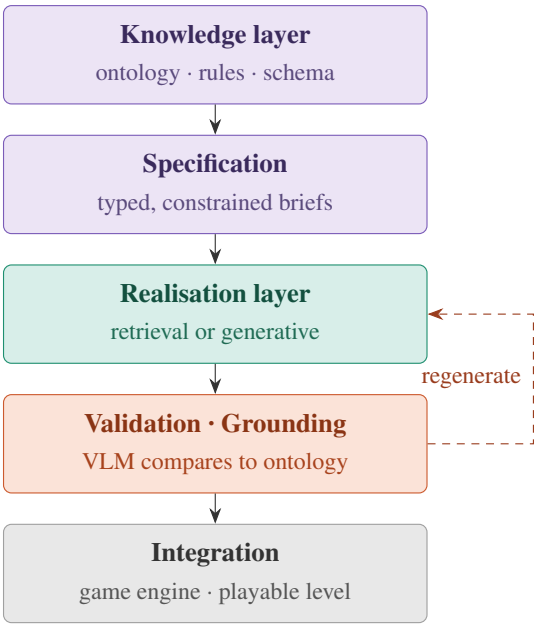


Figure 1. The layered structure of the proposed framework, where a symbolic knowledge layer derives constrained briefs, a swappable realisation layer produces assets, and a validation stage grounds them against the ontology and returns to regenerate on rejection before integration into the engine.

The specification layer reasons over a specific instance of a world, e.g. the particular mystery and the particular layout that have been generated, and emits the set of slots that need filling, together with a brief for each. The brief assembles the slot's semantic type, the constraints that apply to it, and its relational context, and it states the criteria under which a candidate asset will be accepted. The brief can be considered as a type of contract, since both the realiser and the validation stage are written and evaluated against it.

The realisation layer is the swappable and most modular component and the central point of the contribution presented here. A retrieval realiser resolves a brief by selecting from a curated library, which is the baseline. On the other hand, a generative realiser resolves it by synthesising a candidate and validating it, which is detailed in Fig. 2.

The validation and grounding stage closes the loop. A generated candidate asset is inspected by a VLM, which is asked, in effect, whether the candidate satisfies the brief: is it the right kind of object, in the right style, at the right scale, plausibly placeable where the brief intends? A candidate that satisfies the brief is bound, whereas one that does not is rejected and regenerated. It is worth stating plainly where the control over the generator is located because it is not inside the generator. The generator is controlled in the two stages that surround it, as shown in Fig. 2: the conditioning

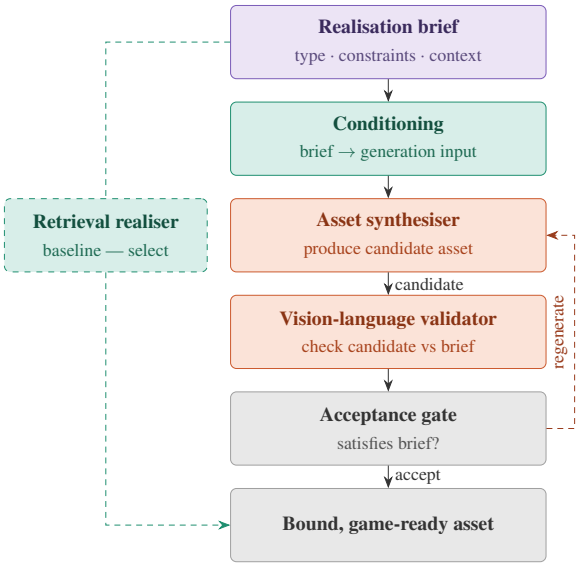


Figure 2. The realisation and grounding loop containing the following steps: 1) the brief is conditioned into a generation input, 2) a synthesiser produces a candidate, 3) a vision-language validator checks it against the brief, 4) the acceptance gate binds the asset or returns to regenerate. The retrieval realiser is the baseline, as it provides the brief-to-asset process by selection.

that turns the brief into the generator's input, and the acceptance gate that enforces the brief on the output, with the VLM serving as the means of applying a symbolic criterion to a synthesised image.

The integration layer assembles bound assets into the playable level and runs the game. It is the only layer that is engine-specific.

The scope of the described process should be discussed briefly. The same validation stage serves both realisers: in the generative case it is the VLM check just described, and in the retrieval case it resolves to a simple library-fit check. Presenting validation as one stage with two implementations, rather than as a framework feature added to the generative process, is what keeps the baseline and the proposed system inside a single architecture.

To make the loop concrete, consider a single slot from a generated case: a cooking tool resting on the kitchen table, which the mystery has marked as a candidate murder weapon. The specification layer derives the brief shown in Fig 3 from the ontology, drawing the slot's type, its attributes, the level's style definition, and the placement constraints from the classes and individuals already present in the world instance.

Conditioning turns this brief into the generator's input, and the synthesiser returns a first candidate: a cauldron. Rendered from several viewpoints and put to the validator, it is checked against the four acceptance criteria in turn. The validator confirms that the object is a single cooking implement, that its period and style

```
slot:       kitchen/implement_1
type:       CookingImplement
style:      medieval, low-poly, muted palette
attributes: graspable, can_be_weapon
scale:      longest dimension 20-60 cm
place:      upright on KitchenTable, near Hearth
role:       candidate murder weapon
accept if:
  - one cooking implement, nothing else
  - period and art style match the level
  - longest dimension within the set
  - rests upright and stable on a surface
```

Figure 3. A realisation brief for one slot, derived from the ontology and passed unchanged to either realiser

match the level, and that it stands upright on a flat surface, but it reads the proportions as roughly knee-height, well above the sixty-centimetre ceiling the brief allows. The candidate is rejected, and the slot is returned for regeneration with the scale constraint tightened. The second candidate, a smaller cast-iron pot, satisfies all four criteria and is bound to the slot. The check that earned its place here is the one a casual glance would have let through: the cauldron looked entirely plausible in isolation, and only the brief's explicit scale band, applied by the validator, caught that it would have towered over the table it was meant to sit on.

The above example also shows the retrieval baseline and the generative realiser meeting at the same interface: the baseline would resolve this identical brief by selecting the nearest medieval pot from the library, where the generative realiser synthesises one and proves it against the brief before accepting it.

4 Use-Case Examples

The example use-case scenario is based on a small-scale detective game: a single location, a handful of characters, one of them a victim and one is a culprit, a scattering of clues, and a player whose task is to use the clues to deduce who is the culprit. The world, the cast, and the mystery elements are generated from the ontology at the start of the play, so that no two cases are identical. The game is built in a contemporary engine, e.g. Godot, with the ontology and the realisers running in a separate local service that the engine talks to.

4.1 The Baseline Realiser

In the context of the baseline example, which is the system of the second author's thesis (Kir, 2026), each slot is bound by retrieval from a curated library of assets. The ontology defines the parts of a room as a small grid of floors, walls, objects, and placement points for items and characters, and the generator assembles a playable map by selecting and placing library assets according to the ontology's rules. Dialogue is authored per role

and assigned to characters according to the parts they have been dealt in the mystery. This is the approach that deals only with retrieving the assets, i.e. it should produce a working game, or at least an in-game world, in which the ontology indexes a fixed catalogue.

4.2 The Generative Realiser

The generative realiser replaces selection with synthesis, as shown in Fig. 2. A brief arriving at the generative realiser is first conditioned into the form the synthesiser expects. This is a deterministic step that translates the ontology's constraints, the asset's type, its style description, its rough dimensions, and its relational context, into the generator's input. The synthesiser produces one or more candidate assets. Each candidate is passed to the vision-language validator, which checks it against the brief's acceptance criteria. The acceptance gate then binds a candidate that passes the validator or, if it does not, returns to synthesis. Accepted assets pass through a step that prepares them for the engine before they are bound into the level.

The instantiation is deliberately local-first: generation and validation run on self-hosted models, with nothing leaving the machine. The motivation is partly practical, since it removes per-call cost and works offline, and partly methodological, since a self-contained pipeline is far easier for a reviewer or a successor to reproduce than one that depends on a remote service whose behaviour changes without notice. Furthermore, online services often feature various constraints on what may or may not be generated, while this limitation can be mitigated when locally hosted models are used. We stress that the choice of any particular model is an instantiation detail; the framework is stated over the roles needed by the process, not the particular models that fill them.

5 Discussion

The role of the ontology in the finalised process is threefold:

- It makes the generator controllable, by supplying the constraints that are enforced at the conditioning and acceptance layers.
- It makes the pipeline interpretable because a rejection is attributable to a specific unsatisfied criterion rather than to an opaque score.
- It makes the realiser swappable because retrieval and generation differ only in how the brief is approached, not in the role they play in the whole process.

The first of these is the answer to the persistent issue of generative models being uncontrollable. The control mechanism can be found in the proposed framework, but it may be recognised in the ontology-based briefs and the symbolic context given to the model rather than in the model itself.

The obvious challenge that can be recognised in the proposed framework is generating truly game-ready assets. A synthesised asset that looks correct in a preview may still have unusable topology, the wrong scale, or geometry that the engine cannot collide against, and no amount of semantic validation makes a malformed mesh playable. The acceptance gate mitigates this only to the extent that the VLM can detect it from a rendered view, which is a real limitation and a place where the framework depends on how good a generation model is combined with it. A reasonable fallback, where 3D synthesis proves unreliable, is to generate two-dimensional textures or sprites instead, which current models handle far more dependably. The proposed framework is indifferent to which assets are generated, since both are simply made by different realisers, even though the process itself is unmodified.

The questions worth asking are, for example:

- Does the ontology-conditioned generation produce assets that satisfy their briefs?
- At what cost are those assets produced?
- Does the resulting level remain playable?

These suggest a small set of measures, such as:

- the acceptance rate of generated candidates,
- the number of regenerations needed before acceptance,
- the proportion of brief constraints satisfied,
- a rating of stylistic coherence across a generated level, which a panel of judges or a selected VLM can supply,
- the validity of placements,
- the playability of the finished level.

Several of these measures can be taken with existing instruments rather than devised from scratch. The stylistic and surface quality of generated candidates can be scored with a purpose-built vision-language metric for 3D objects such as Gen3DEval (Maiti et al., 2025), which is designed to align with human judgement without requiring a ground-truth reference, and the scene-level measures can follow the protocol used for comparable systems (Liu et al., 2025), combining standard image–text similarity with human and model ratings of aesthetic quality and functional plausibility.

The natural comparator is the retrieval baseline, against which the generative realiser should be measured for coherence, variety, and authoring effort. A proof-of-concept at modest scale is planned for a future study. Such an example can be used to provide empirical answers to validity and feasibility of the proposed

framework by providing the answers to the above measures, including taking into account the amount of time needed for the whole process to run.

The proposed framework is devised with generalisation in mind. Nothing in the framework is specific to detective games, which are used as examples' genre. The genre is a convenient testbed because its small, well-structured worlds keep the ontology tractable and the playability criteria sharp, but the same brief-and-realiser arrangement applies wherever a game world can be specified semantically and its contents synthesised. Furthermore, the framework is not created with a particular set of models in mind either, i.e. the proposed framework defines the steps of a process and the roles that are needed for its completion, but the specific actors in the process may be interchangeable.

6 Conclusion

We have argued that the ontology in ontology-driven content generation can do more than simply index a catalogue: it can specify assets to be synthesised, and a VLM can check that the synthesis met the specification. Defined as a swappable modular realiser between a symbolic specification and a validation loop, this approach generalises the often-used retrieval method and proposes a way to incorporate generative models into the game assets' development pipeline as a controllable resource. The framework is presented here with the described baseline and generative use-case scenarios, followed by a conceptual study of those specific systems. Further research opportunities are observed in developing an empirical study of a specific instantiation, along with the brief schema that it is based on.

Acknowledgement

The authors acknowledge that Claude Opus 4.8 by Anthropic was used to assist in editing and streamlining sections of this paper.

References

Bai, W., Li, Y., Luo, W., Chen, W., & Sun, H. (2025). Vision-language models as differentiable semantic and spatial rewards for text-to-3d generation [Preprint]. *arXiv preprint arXiv:2509.15772*.

Barros, G. A. B., Liapis, A., & Togelius, J. (2016). Playing with data: Procedural generation of adventures from open data. *Proceedings of the First International Joint Conference of DiGRA and FDG*. https://doi.org/10.26503/dl.v2016i1.801

Fan, H., Chen, H., Huang, Z., Liu, Z., & Sheng, L. (2026). 3D Asset Generation: A Survey of Evolution Towards Autoregressive and Agent-Driven Paradigms. *Frontiers of Computer Science*, *20*(11), 2011710. https://doi.org/10.1007/s11704-025-50381-5

Farrokhi Maleki, M., & Zhao, R. (2024). Procedural Content Generation in Games: A Survey with Insights on Emerging LLM Integration. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, *20*(1), 167–178. https://doi.org/10.1609/aiide.v20i1.31877

Gallotta, R., Todd, G., Zammit, M., Earle, S., Liapis, A., Togelius, J., & Yannakakis, G. N. (2024). Large Language Models and Games: A Survey and Roadmap. *IEEE Transactions on Games*, 1–18. https://doi.org/10.1109/TG.2024.3461510

Hu, Z., Iscen, A., Jain, A., Kipf, T., Yue, Y., Ross, D. A., Schmid, C., & Fathi, A. (2024). SceneCraft: An LLM Agent for Synthesizing 3D Scenes as Blender Code. *Proceedings of the 41st International Conference on Machine Learning*, *235*, 19252–19282.

Kautz, H. A. (2022). The Third AI Summer: AAAI Robert S. Engelmore Memorial Lecture. *AI Magazine*, *43*(1), 105–125. https://doi.org/10.1002/aaai.12036

Kir, M. (2026). *Developing a Detective Video Game Based on Ontologies and Procedural Content Generation* [Master's thesis, University of Zagreb Faculty of Organization and Informatics] [Forthcoming].

Lee, H., Savva, M., & Chang, A. X. (2024). Text-to-3D Shape Generation. *Computer Graphics Forum*, *43*(2), e15061. https://doi.org/10.1111/cgf.15061

Lee, S., Kim, S., Park, S., Kim, G., & Seo, M. (2024). Prometheus-Vision: Vision-Language Model as a Judge for Fine-Grained Evaluation. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 11286–11315). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.672

Lei, B., Li, Y., Liu, X., Yang, S., Xu, L., Huang, J., Tang, R., Weng, H., Liu, J., Xu, J., Zhou, Z., Zhu, Y., Xing, J., Xu, J., Ma, C., Yan, X., Yang, Y., Wang, C., Xu, D., … Guo, C. (2025). Hunyuan3D studio: End-to-end AI pipeline for game-ready 3d asset generation [Technical Report]. https://arxiv.org/abs/2509.12815

Liu, X., Tai, Y.-W., & Tang, C.-K. (2025). Agentic 3D scene generation with spatially contextualized VLMs [arXiv:2505.20129].

Maiti, S., Agapito, L., & Kokkinos, F. (2025). Gen3DEval: Using vLLMs for Automatic Evaluation of Generated 3D Objects. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 18552–18562. https://doi.org/10.1109/CVPR52734.2025.01729

Marra, G., Dumančić, S., Manhaeve, R., & De Raedt, L. (2024). From statistical relational to neurosymbolic artificial intelligence: A survey. *Artificial Intelligence*, *328*, 104062. https://doi.org/10.1016/j.artint.2023.104062

Mohr, H., Eger, M., & Martens, C. (2018). Eliminating the Impossible: A Procedurally Generated Murder Mystery. *Joint Proceedings of the AIIDE 2018 Workshops Co-Located with 14th AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*. Retrieved June 29, 2026, from https://ceur-ws.org/Vol-2282/

Okreša Đurić, B., & Schatten, M. (2016–June 3). Defining Ontology Combining Concepts of Massive Multi-Player Online Role Playing Games and Organization of Large-Scale Multi-Agent Systems. In P. Biljanović (Ed.), *39th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)* (pp. 1330–1335). IEEE. https://doi.org/10.1109/MIPRO.2016.7522346

Poole, B., Jain, A., Barron, J. T., & Mildenhall, B. (2023). DreamFusion: Text-to-3d using 2d diffusion [arXiv:2209.14988]. *The Eleventh International Conference on Learning Representations (ICLR)*.

Sacco, O., Liapis, A., & Yannakakis, G. N. (2016). A Holistic Approach for Semantic-Based Game Generation. *2016 IEEE Conference on Computational Intelligence and Games (CIG)*, 1–8. https://doi.org/10.1109/CIG.2016.7860386

Sacco, O., Liapis, A., & Yannakakis, G. N. (2017). Game Character Ontology (GCO): A Vocabulary for Extracting and Describing Game Character Information from Web Content. *Proceedings of the 13th International Conference on Semantic Systems*, 9–16. https://doi.org/10.1145/3132218.3132233

Schatten, M., Okreša Đurić, B., Tomičić, I., & Ivković, N. (2017–June 23). Agents as Bots – An Initial Attempt Towards Model-Driven MMORPG Gameplay. In Y. Demazeau, P. Davidsson, J. Bajo, & Z. Vale (Eds.), *Advances in Practical Applications of Cyber-Physical Multi-Agent Systems: The PAAMS Collection* (pp. 246–258). Springer International Publishing. https://doi.org/10.1007/978-3-319-59930-4_20

Shaker, N., Togelius, J., & Nelson, M. J. (2016). *Procedural content generation in games*. Springer. https://doi.org/10.1007/978-3-319-42716-4

Singh, J., & Zheng, L. (2023). Divide, Evaluate, and Refine: Evaluating and Improving Text-to-Image Alignment with Iterative VQA Feedback. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in neural information processing systems* (pp. 70799–70811, Vol. 36). Curran Associates, Inc.

Smith, A. M., & Mateas, M. (2011). Answer Set Programming for Procedural Content Generation: A Design Space Approach. *IEEE Transactions on Computational Intelligence and AI in Games*, *3*(3), 187–200. https://doi.org/10.1109/TCIAIG.2011.2158545

Stockdale, A. (2016). ClueGen: An exploration of procedural storytelling in the format of murder mystery games. *Experimental AI in Games (EXAG) Workshop at AIIDE, AAAI Technical Report WS-16-22*.

Summerville, A., Snodgrass, S., Guzdial, M., Holmgård, C., Hoover, A. K., Isaksen, A., Nealen, A., & Togelius, J. (2018). Procedural content generation via machine learning (PCGML). *IEEE Transactions on Games*, *10*(3), 257–270. https://doi.org/10.1109/TG.2018.2846639

Togelius, J., Yannakakis, G. N., Stanley, K. O., & Browne, C. (2011). Search-based procedural content generation: A taxonomy and survey. *IEEE Transactions on Computational Intelligence and AI in Games*, *3*(3), 172–186. https://doi.org/10.1109/TCIAIG.2011.2148116

Tutenel, T., Bidarra, R., Smelik, R. M., & de Kraker, K. J. (2009). Using semantics to improve the design of game worlds. *Proceedings of the Fifth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE)*, 100–105.

Yang, Y., Sun, F.-Y., Weihs, L., VanderBilt, E., Herrasti, A., Han, W., Wu, J., Haber, N., Krishna, R., Liu, L., Callison-Burch, C., Yatskar, M., Kembhavi, A., & Clark, C. (2024). Holodeck: Language guided generation of 3d embodied AI environments. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16227–16237. https://doi.org/10.1109/CVPR52733.2024.01536

Yang, Y., Wang, C., Ye, J., Li, Y., Chen, Z., Huang, Z., Mu, Y., Chen, Z., Guo, C., & Liu, X. (2026). PhysForge: Generating physics-grounded 3d assets for interactive virtual world [arXiv:2605.05163]. *Proceedings of the International Conference on Machine Learning (ICML)*.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-judge with MT-bench and chatbot arena [arXiv:2306.05685]. *Advances in Neural Information Processing Systems 36 (NeurIPS), Datasets and Benchmarks Track*.

Zhu, Y., Deng, K., Fauconnier, J.-P., Navarro, I., Li, D., Pun, A., Zhang, Y., Zhuang, P., Sun, X., Agrawala, M., Bhat, K., & Zhou, T. (2026). CubePart: An open-vocabulary part-controllable 3d generator [arXiv:2605.28763]. *ACM SIGGRAPH 2026 Conference Papers*.

Where Benchmark Intrusion Detection Meets Cloud Game Streaming: Two Generalisation Gaps in a Deep-Learning Detector

Bogdan Okreša Đurić, Lovro Đurđević
University of Zagreb Faculty of Organization and Informatics
Artificial Intelligence Laboratory
Pavlinska 2, 42000 Varaždin, Croatia
dokresa@foi.unizg.hr, ldurdevic22@student.foi.hr

Abstract. *Cloud game-streaming services such as GeForce NOW are high-value, latency-sensitive targets, making learned intrusion detection attractive. This paper asks whether a benchmark-trained detector can serve that purpose, using a single dense neural network as a fixed instrument. The model is evaluated at three removes from its NSL-KDD training distribution: a validation split, the held-out KDDTest+ partition, and game-streaming traffic. Validation accuracy above 0.99 falls to 0.776 on KDDTest+, concentrated in minority attack classes, and the benchmark feature schema does not map onto encrypted streaming traffic. Class imbalance and feature-representation mismatch, not model capacity, prevent transferring benchmark-trained detectors to streaming infrastructure.*

Keywords. cloud game streaming, network intrusion detection, deep learning, NSL-KDD, generalisation, feature extraction

1 Introduction

Cloud game-streaming services, in which a game runs on remote infrastructure and is delivered to the player as an interactive video stream, have grown into a significant class of online services. Platforms such as NVIDIA GeForce NOW, Xbox Cloud Gaming, and Amazon Luna run games on multi-tenant servers and stream the rendered output to thin clients over the Internet. The traffic they generate is continuous, high in bitrate, and acutely sensitive to latency and loss, since even a small disruption is immediately visible to the player as stutter or lag. These same properties make these services attractive targets: the gaming sector is consistently among the most heavily targeted by distributed denial-of-service campaigns (Yoachimik et al., 2026).

It is natural to ask whether the learned intrusion detectors that perform so well on standard benchmarks could be applied to protect such infrastructure. Network intrusion detection has long been treated as a classification problem suitable for machine learning. Among the public benchmarks, the NSL-KDD dataset (Tavallaee et al., 2009) remains one of the most widely used, with a substantial body of work reporting classification accuracy exceeding 99 % on it. Taken at face value, such figures suggest that a capable detector is easily obtained and might simply be pointed at a new domain.

This paper tests that suggestion and finds it wanting, in two distinct ways. Rather than proposing a new architecture or pursuing an improved accuracy figure, it takes a single, deliberately modest deep neural network and treats it as a fixed instrument, observing its behaviour in three different scenarios based on the data on which it was trained. The first is an in-distribution validation split of the training partition. The second is the NSL-KDD test partition, which is constructed to differ from the training partition in attack difficulty and class composition (Tavallaee et al., 2009). The third is deployment against game-streaming traffic, captured both from a self-hosted streaming stack and from a commercial service. The three evaluation points yield two gaps in performance, which we characterise and then diagnose.

The four attack categories of NSL-KDD map directly onto threats to game-streaming infrastructure, which is what makes the benchmark a reasonable starting point for this research. Denial of service corresponds to the attacks of large volume that routinely disrupt latency-sensitive game services; reconnaissance (Probe) corresponds, e.g. to scanning of the streaming edge; remote-to-local (R2L) corresponds to account takeover, which on these platforms exposes payment details and game libraries; and user-to-root (U2R) corresponds, in a multi-tenant cloud setting, to tenant-isolation breakout, the escalation from a guest context to the shared host. The mapping is clean at the level of threat categories, which is precisely why the failure documented below is informative rather than expected: the challenge is not that the threats differ, but that the

traffic does.

The contribution of the paper is therefore methodological rather than architectural. It offers an account of how the apparent effectiveness of one benchmark-trained detector degrades as the evaluation moves towards a realistic target, together with a diagnosis of the two causes responsible. The first gap is attributable to severe class imbalance in the benchmark; the second to a mismatch between the connection-record features the model was trained on and the continuous media traffic that game streaming produces. The second gap is the more significant for the streaming setting, and the more general, since it is not concerned with any particular model.

The remainder of the paper is organised as follows. Section 2 reviews the relevant background and related work, including the traffic characteristics of cloud gaming. Section 3 describes the model, the data, the three evaluation points, and the deployment. Section 4 reports the results and the two gaps. Section 5 discusses what binds the detector's performance and what a more informative evaluation protocol would require. Section 6 concludes and outlines future work.

2 Background and Related Work

2.1 The NSL-KDD Dataset

The NSL-KDD dataset was introduced as a refinement of the KDD Cup '99 dataset, which had served for some time as a standard reference in intrusion-detection research (Tavallaee et al., 2009). The principal defect of KDD Cup '99 was a large number of duplicate records in the training set, which biased models towards frequently occurring patterns and produced optimistic evaluation results. NSL-KDD removes the redundant records and selects samples across a range of classification difficulty, rendering a more balanced and more varying benchmark. Each record describes a single network connection through forty-one features that capture properties such as connection duration, protocol, the number of bytes transferred, and statistical indicators of network behaviour, and is labelled as normal traffic or as one of four attack categories: denial of service (DoS), reconnaissance (Probe), remote-to-local (R2L), and user-to-root (U2R).

For the purposes of this paper, the most important property of NSL-KDD is that its test partition, KDDTest+, is deliberately constructed to differ from the training partition, KDDTrain+, in the difficulty and representation of records (Tavallaee et al., 2009). A model evaluated on KDDTest+ is therefore not being asked simply to reproduce the observations applicable to the data it was trained on, and a decline in performance between an in-distribution split and KDDTest+ is a genuine indicator of limited generalisation rather than an artefact of the evaluation. It should also be noted here that NSL-KDD derives from network traffic captured in the late 1990s and is, by common acknowledgement, no longer representative of modern network conditions. This limits the external validity of any result obtained on it, as is pointed out again in Section 5.

2.2 Saturation of Single-Dataset Results

Deep-learning classifiers reporting accuracy above 99 % on NSL-KDD are numerous, and the four-class taxonomy used here has been revisited many times. A recurring feature of the applicable literature is that this figure is obtained on a random split of the training partition rather than on the separate test partition, so that training and evaluation data share a distribution. Such a figure, therefore, measures the capacity of a model to fit a distribution and conveys little about its behaviour on data drawn from elsewhere. This paper treats the difference between the in-distribution figure and the separated figure as the object of study, rather than as a simple, obvious observation, or a result that should be minimised.

2.3 Cross-Dataset Generalisation

A smaller but established line of work has examined how intrusion-detection models transfer across datasets. The consistent finding is that models scoring near-perfectly within a single dataset degrade sharply, often towards chance, when evaluated on a different one. D'Hooge et al. (2020) train a range of supervised algorithms on one dataset and evaluate them on another, restricting attention to attack classes common to both, and report that the classifiers do not perform adequately on traffic from networks unseen at training time. Verkerken et al. (2022) report a comparable degradation for unsupervised methods. Related work on the statistical properties of synthetic intrusion-detection datasets (Layeghy et al., 2021) and on standardising feature sets across datasets (Sarhan et al., 2022) situates the problem in the lack of correspondence between datasets' feature spaces and traffic characteristics. The work reported here is complementary to this line: rather than transferring a model from one labelled dataset to another, it confronts a benchmark-trained model with a class of real traffic, game streaming, that the benchmark schema was never designed to describe, which makes the feature-space mismatch concrete rather than statistical.

2.4 Cloud Game-Streaming Traffic

The properties of the traffic generated by cloud game-streaming services make it inappropriate for the benchmark schema. Di Domenico et al. (2021) analyse Stadia, GeForce NOW, and PS Now, and find that GeForce NOW and Stadia carry their video content over RTP, with bitrates reaching tens of megabits per second depending on resolution and network conditions. A subsequent measurement study of GeForce NOW (Lyu et

al., 2024) provides a detailed structural overview of its service and gameplay flows and extends the analysis to Xbox Cloud Gaming. Typically, it is described as a sustained, high-rate, latency-sensitive flow of encrypted media, dominated downstream by the video stream and accompanied upstream by a thin control channel. This is quite a different structure from the discrete, short-lived, service-typed connection that the NSL-KDD feature schema presupposes, and Section 4 shows that the difference is what defeats the trained model.

2.5 Class Imbalance in Intrusion Detection

The R2L and U2R categories are severely under-represented in NSL-KDD, and the difficulty of detecting them is well documented. A range of solutions has been proposed, including oversampling of the minority classes (Chawla et al., 2002), cost-sensitive training, and related techniques. Per-class reporting of precision, recall, and the F1 measure is an expected practice where minority detection is at issue (Sokolova & Lapalme, 2009). The minority-class collapse observed in Section 4 is, in this sense, a known occurrence. Therefore, the contribution of this paper is not to rediscover it, but to place it within an account of how benchmark accuracy might not represent reality, i.e. how it may exaggerate real-life results.

3 Method

The study uses a single deep neural network, held fixed throughout, as an instrument for measuring generalisation. This section describes the model and its training, the data and its preparation, the three evaluation points, and the deployment in which the second gap is observed.

3.1 Model and Training

The model is a sequential, fully connected network with three hidden layers of 128, 64, and 32 units. The hidden layers use the rectified linear activation, and dropout with rates of 0.3 and 0.2 is applied after the first and second hidden layers respectively as regularisation (Srivastava et al., 2014). The output layer has five units with a softmax activation, producing a probability distribution over the five classes. The network is trained with the Adam optimiser (Kingma & Ba, 2015) and the categorical cross-entropy loss, with a batch size of 128. Training is permitted for up to fifty epochs but is governed by early stopping on the validation loss, with patience of five epochs and restoration of the best weights. In practice, though, training halts after twenty-two epochs. The network has 26 117 trainable parameters. The architecture is deliberately modest and is not varied: because the purpose of the study is to assess generalisation rather than to maximise accuracy. Therefore, holding the model fixed is a feature of the design rather than a limitation of it.

3.2 Data and Preparation

Training uses the KDDTrain+ partition of NSL-KDD, comprising 125 973 records, and evaluation uses the KDDTest+ partition, comprising 22 544 records. The original fine-grained attack labels are mapped to the five-class scheme described in Section 2. Three columns are removed before training: the original label, the difficulty indicator, and `num_outbound_cmds`, whose values are uniformly zero and therefore carry no information. This leaves 40 input features, of which three (`protocol_type`, `service`, and `flag`) are categorical. The categorical features are one-hot encoded and the numeric features are scaled to the unit interval; one-hot encoding expands the input to 121 dimensions, which constitutes the final input to the network.

The distribution of classes in KDDTrain+ is given in Table 1. The imbalance is pronounced: the two minority classes together account for fewer than one per cent of the training records, against more than 90 % for the normal and DoS classes combined. This imbalance is the cause of the first generalisation gap reported below.

Table 1. Class distribution in the KDDTrain+ partition

Class	Records
Normal	67 343
DoS	45 927
Probe	11 656
R2L	995
U2R	52
Total	125 973

3.3 The Three Evaluation Points

The model is evaluated at three points, defined exactly so that the two gaps between them are clear.

In-distribution validation. During training, twenty per cent of KDDTrain+ is held out as a validation split. Because this split is drawn from the same distribution as the training data, performance on it measures how well the model fits that distribution. This is the figure that reaches above 0.99 and corresponds to what much of the literature reports.

Separate benchmark. The model is evaluated on KDDTest+, which differs from KDDTrain+ in attack difficulty and class composition (Tavallaee et al., 2009). This is the first test of generalisation.

Streaming-traffic deployment. The same trained model is confronted with cloud game-streaming traffic, captured both from a self-hosted streaming stack and from a commercial service. This is the second test of generalisation, and it exposes a different failure from the first.

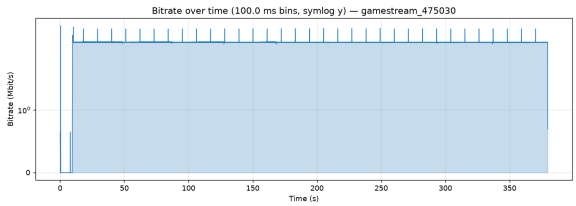


Figure 1. The constant bitrate with an occasional spike, of the self-hosted capture

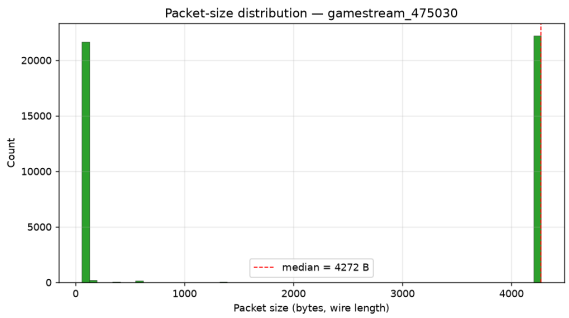


Figure 2. Distribution of packet size of the self-hosted capture

3.4 Deployment and Streaming Traffic

To confront the trained model with realistic streaming traffic, two captures are used:

1. The first is from a self-hosted streaming stack, with Sunshine as the host and Moonlight as the client, communicating over the GameStream protocol. This stack is fully under our control, can be captured at the interface between host and client, and is reproducible by others.

2. The second is a capture of a commercial GeForce NOW session taken at the client.

The self-hosted capture serves as the primary, reproducible exhibit, while the GeForce NOW capture indicates that the result holds on a real service. Both captures record traffic shape only, i.e. packet headers, sizes, timing, direction, and protocol; payloads are encrypted and are neither inspected nor retained.

The traffic of the self-hosted capture (Sunshine as the host and Moonlight as the client) was being captured for a total of 379 seconds, wherein a total of 44 368 packages were captured, carrying a total of 96 674 250 bytes. Out of the captured packets, 21 473 were UDP packets (14 800 RTP packets), and 711 were TCP packets. During the captured period, after the connection was established, the bitrate was constant, with an occasional spike, as seen in Fig. 1. Although there are many small packages, the median size of a package is 4 272 bytes, as shown in Fig. 2.

The captured traffic of a GeForce NOW session shows a similar context, although with slightly different values. The session settings at the time of the capture include the resolution of 1920×1080, with automatic maximum bitrate, balanced streaming quality mode, over an Internet connection providing roughly 300 Mbps download and 150 Mbps upload speeds, running on a Lenovo ThinkBook 15 G2 ITL within the client version 2.0.86.129/ro running Escape Simulator 2 by Pine Studio.

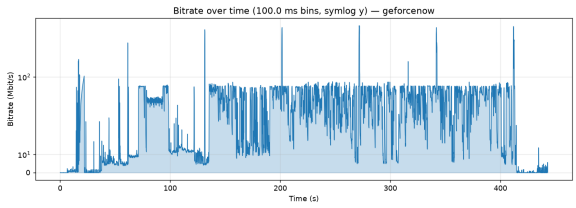


Figure 3. The varying bitrate is almost constant while the player or the scene of the GeForce NOW capture were moving

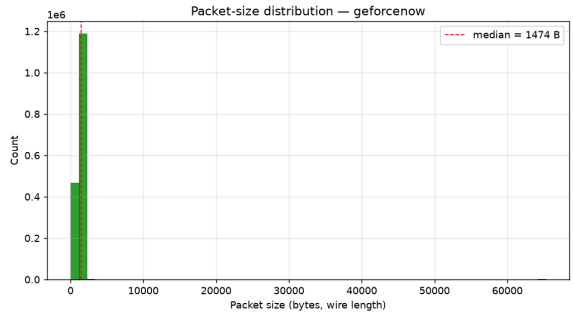


Figure 4. Distribution of packet size of the GeForce NOW capture

The traffic of the GeForce NOW capture was being captured for a total of 442 seconds, wherein a total of 1 662 788 packages were captured, carrying a total of 2 128 208 510 bytes. Out of the captured packets, 816 071 were UDP packets (723 907 RTP packets), 15 028 were TCP packets, and 29 ICMP packets. During the captured period, after the connection was established, the bitrate was varying, but always high while the player or the scene were moving, as seen in Fig. 3. The packages during this capture were much smaller, as seen in Fig. 4, where the median packet size is 1 474 bytes, although their count is vastly greater than the median packets in the self-hosted capture.

The two packet-size distributions are not measured from the same position, and the difference in their medians (4 272 bytes for the self-hosted capture against 1 474 bytes for GeForce NOW) reflects the point of capture rather than a property of the traffic. The GeForce NOW session was captured at the wireless client, where packets appear at their on-wire size, so its median sits just below the Ethernet maximum transmission unit of roughly 1 500 bytes. The self-hosted session was captured at the interface between host and client, where segmentation offload coalesces several outgoing packets into larger units before they are

recorded, wherefore its distribution reflects host-side segments rather than on-wire frames. This caveat applies only to the packet-size comparison. The bitrate and total-volume figures aggregate bytes over time and are independent of how the traffic is segmented, so they remain directly comparable across the two captures. Where on-wire packet size is the quantity of interest, the GeForce NOW capture is the faithful measurement.

3.5 Mapping the Feature Schema onto Streaming Traffic

The model expects an input vector derived from the NSL-KDD connection-record schema, in which a single record describes one discrete, service-typed connection and several features are statistics aggregated over a window of recent connections, for instance counts of connections to the same host or service within a two-second interval. Game-streaming traffic does not present this structure. A streaming session is a small number of long-lived, high-rate media flows rather than a sequence of discrete connections. It is carried over UDP and so has no TCP connection-state flag, and it corresponds to no entry in the benchmark's enumeration of application services. The question for deployment is therefore not how accurately the model classifies streaming traffic, but whether the features it requires can be constructed from such traffic at all.

To answer this, each of the 41 NSL-KDD features is assessed against the captured streaming traffic and classified as derivable, degenerate, or inapplicable. A feature is derivable if it can be computed and carries discriminative information; degenerate if it can be computed but is reduced to a constant for streaming traffic, as the content features describing events within an interactive login session do; and inapplicable if the unit it depends on does not exist in the traffic, as is the case for the service-typed and TCP-flag features and for the windowed counts of repeated discrete connections.

Of the 41 features, four are derivable, fifteen degenerate (including `num_outbound_cmds`, which is constant in NSL-KDD and removed during preprocessing), and twenty-two inapplicable, as set out in Table 2. Close to nine in ten features therefore carry no information or cannot be honestly constructed for a media stream. The four features that do persist are dominated by the two byte-count features, i.e. by data volume, which is the signature that the model associates with the signature the model associates with denial of service.

The effect is more pronounced once the one-hot encoding is considered. The 121-dimensional input comprises the 37 numeric features and the expansion of the three categorical features; with `flag` contributing eleven columns and `protocol_type` three, the `service` feature accounts for the remaining seventy. Both `service` and `flag` are inapplicable to streaming traffic, so together they occupy 81 of the 121 input dimensions. Two-thirds of the vector the model actually receives is thus structurally meaningless for a media stream before the degenerate features are even considered.

4 Results

The three evaluation points are reported in turn, exposing the two gaps identified earlier.

4.1 In-Distribution Performance

During training, validation accuracy rises quickly within the first few epochs and stabilises above 0.99, where it remains for the rest of training; the validation loss declines and then oscillates mildly, while the training loss declines steadily. Figure 5 shows the accuracy and loss across the twenty-two epochs of training. Taken on its own, this figure would suggest a highly capable intrusion detection model, and it is consistent with the results commonly reported for NSL-KDD.

The closeness of the training and validation curves is sometimes read as evidence that a model has not overfitted. That interpretation is correct only in a narrow sense and is worth examining carefully. Because the validation split is drawn from the training distribution, the small gap shows that the model fits that distribution well and has not overfitted to the training sample in particular. On the other hand, it says nothing about how the model will behave on data constructed to differ from the training distribution. The above-described can be observed as a challenge in generalisation, rather than being a sign of an incapable model. Establishing this distinction is what licenses describing the subsequent fall as a generalisation gap rather than a failure of model capacity.

4.2 First Gap: Held-Out Benchmark Performance

On KDDTest+, overall accuracy is 0.776, a fall of more than 20 % from the in-distribution figure. The per-class results are given in Table 3. The fall is not distributed evenly across the classes: it is most visible in the two minority classes. DoS and Probe retain serviceable F1 measures of 0.89 and 0.77, and the normal class is recovered with a recall of 0.97. R2L and U2R, by contrast, attain F1 measures of only 0.12 and 0.09, driven by recall of 0.07 and 0.04 respectively. This can be interpreted as follows: the model misses roughly 19 of every 20 R2L attacks and 24 of every 25 U2R attacks.

The confusion matrix in Table 4 shows where the missed attacks are sent. The overwhelming majority of R2L and U2R records are classified as normal traffic: 2 572 of 2 754 R2L records and 174 of 200 U2R records are predicted as normal. This is the most damaging failure mode available to the trained model because an attack classified as normal produces no

Table 2. Mapping of the 41 NSL-KDD features onto steady-state game-streaming traffic. Group: B basic, C content, T time-based traffic, H host-based traffic. Class: Der. derivable, Deg. degenerate, Inap. inapplicable.

#	Feature	Group	Class	#	Feature	Group	Class
1	duration	B	Der.	22	is_guest_login	C	Deg.
2	protocol_type	B	Der.	23	count	T	Inap.
3	service	B	Inap.	24	srv_count	T	Inap.
4	flag	B	Inap.	25	serror_rate	T	Inap.
5	src_bytes	B	Der.	26	srv_serror_rate	T	Inap.
6	dst_bytes	B	Der.	27	rerror_rate	T	Inap.
7	land	B	Deg.	28	srv_rerror_rate	T	Inap.
8	wrong_fragment	B	Deg.	29	same_srv_rate	T	Inap.
9	urgent	B	Inap.	30	diff_srv_rate	T	Inap.
10	hot	C	Deg.	31	srv_diff_host_rate	T	Inap.
11	num_failed_logins	C	Deg.	32	dst_host_count	H	Inap.
12	logged_in	C	Deg.	33	dst_host_srv_count	H	Inap.
13	num_compromised	C	Deg.	34	dst_host_same_srv_rate	H	Inap.
14	root_shell	C	Deg.	35	dst_host_diff_srv_rate	H	Inap.
15	su_attempted	C	Deg.	36	dst_host_same_src_port_rate	H	Inap.
16	num_root	C	Deg.	37	dst_host_srv_diff_host_rate	H	Inap.
17	num_file_creations	C	Deg.	38	dst_host_serror_rate	H	Inap.
18	num_shells	C	Deg.	39	dst_host_srv_serror_rate	H	Inap.
19	num_access_files	C	Deg.	40	dst_host_rerror_rate	H	Inap.
20	num_outbound_cmds	C	Deg.	41	dst_host_srv_rerror_rate	H	Inap.
21	is_host_login	C	Deg.				

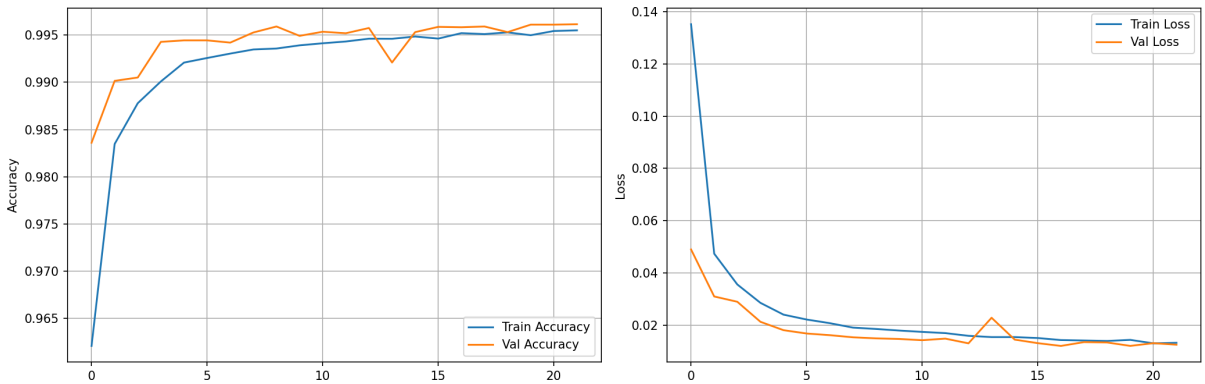


Figure 5. Accuracy and loss over the training epochs

Table 3. Per-class results on KDDTest+

Class	Prec.	Rec.	F1	Samples
DoS	0.96	0.83	0.89	7 458
Probe	0.87	0.70	0.77	2 421
R2L	0.96	0.07	0.12	2 754
U2R	0.90	0.04	0.09	200
Normal	0.68	0.97	0.80	9 711

Table 4. Confusion matrix on KDDTest+; rows are true classes, columns are predicted classes

	DoS	Pr.	R2L	U2R	Norm.
DoS	6 169	22	2	0	1 265
Probe	210	1 695	1	0	515
R2L	0	2	180	0	2 572
U2R	0	14	3	9	174
Normal	54	222	1	1	9 433

alert and passes silently. The pattern of high precision alongside very low recall for these classes is itself informative. When the model does assign a record to R2L or U2R it is usually correct, but it does so very rarely. The above-stated is the signature of a model that has seen too few examples of a class to form a usable decision rule, and it follows directly from the training distribution in Table 1, in which R2L is represented by 995 records and U2R by 52.

4.3 Second Gap: Streaming-Traffic Deployment

The second gap appears when the same model is confronted with game-streaming traffic, and it is of a different kind from the first. The first gap concerns attacks the model was never given enough examples of to learn; the second concerns traffic the model cannot be given at all because the feature schema it requires has

no coherent mapping onto a continuous media stream.

The mapping of Section 3.5 is itself the result. Of the 41 NSL-KDD features, only four can be faithfully derived from a streaming flow, fifteen collapse to constants, and twenty-two depend on a structure: a named service, a TCP connection state, or a population of repeated discrete connections, that a single sustained media flow does not possess. The model cannot be supplied with the inputs it was trained on because those inputs do not exist in this traffic.

The practical consequence follows from which features survive. When only the volume and rate features carry real information, a sustained high-bitrate stream presents the same signature the model associates with flooding, and benign streaming traffic is liable to be classified as a denial-of-service attack. This is the worst available outcome for the streaming setting, where a false positive on legitimate play is itself a denial of service. The failure is not a deficiency of the model, but of the attempt to express streaming traffic in a representation built for late-1990s connection records.

5 Discussion

The two gaps point to a single conclusion: a benchmark-trained detector is bound by class imbalance and by feature representation, not by model capacity, and neither obstacle would be removed by a larger or deeper neural network. At the first gap, the minority classes are represented by too few examples for any model to learn a reliable boundary from them, and additional statistical power applied to the same data would not supply the missing examples. At the second gap, the inputs the streaming setting can offer do not match the representation the model requires; a more expressive model would only fit the benchmark representation more closely and would not reconcile the two. For the streaming target specifically, the second gap is essential: it is not that the model classifies streaming traffic poorly, but that the traffic cannot be faithfully expressed in its required input schema in the first place.

This has a direct effect on the appeal of reusing benchmark-trained detectors to protect game-streaming infrastructure. The threat categories map cleanly onto that setting, which is what makes the idea tempting, but the mapping of threats is not a mapping of traffic. A detector trained on connection-record features cannot be pointed at a continuous encrypted media stream and expected to function, and the volume-and-rate features that do survive the transition are precisely the ones that cause benign streaming traffic to resemble a flood. Protecting streaming infrastructure with learned detection would require features defined natively over media flows, such as bitrate stability, packet-size and inter-arrival distributions, and flow-level rather than connection-level statistics, and training data drawn from such traffic rather than from a connection-oriented benchmark.

The wider methodological point is that results reporting only in-distribution accuracy, as found in much of the NSL-KDD literature, systematically overstate operational capability because they reflect the one quantity least likely to fail. More informative evaluation reports featuring separated performance with per-class metrics would render a minority-class collapse unable to hide inside a favourable average. What is more, where deployment against a particular kind of traffic is considered, an explicit account should be given of whether the feature schema can represent the applicable traffic at all. The feature-extraction path, which is rarely documented, turns out to be decisive and deserves to be reported as a first-class part of the model training process.

Several limitations bound the above proposals. The study uses a single model and a single benchmark; NSL-KDD is an ageing dataset whose traffic is not representative of present-day networks, so the benchmark figures should be read as illustrative. The GeForce NOW capture is a single-user convenience sample that illustrates the representation mismatch rather than characterising the service, and the self-hosted stack stands in for commercial platforms whose proprietary protocols and infrastructure differ. These limitations constrain the generality of the figures but not the logic of the result, which concerns the structure of the evaluation and the accurate representation of the traffic rather than the score of any particular model. The two captures are also taken from different vantage points, so their packet-size distributions are not strictly comparable, as noted in Section 3; their volume and bitrate characteristics are unaffected.

6 Conclusion and Future Work

This paper used a single, fixed deep neural network as an instrument to ask whether a benchmark-trained intrusion detector can be applied to cloud game-streaming traffic. An in-distribution validation accuracy above 0.99 falls to 0.776 on the held-out KDDTest+ partition, with the fall concentrated in the minority R2L and U2R classes, and the NSL-KDD feature schema is shown to have no coherent mapping onto continuous encrypted streaming traffic, with only the volume-and-rate features surviving the move and those causing benign streaming traffic to resemble a denial-of-service attack. The first gap is attributable to class imbalance and the second to the mismatch between the connection-record format the model requires and the media-flow traffic streaming produces. The binding constraints are the data and its representation, not the capacity of the model, and the clean mapping of threat categories onto the streaming context does not carry a mapping of traffic with it.

Several directions are left for future work, named

here so that their absence is understood as a matter of scope. The first is the construction of a model over features defined natively for media flows, trained on game-streaming traffic rather than on a connection-oriented benchmark. The captures and feature analysis reported here are a step towards specifying such features. The second is the mitigation of the minority-class collapse through oversampling, class weighting, or cost-sensitive training (Chawla et al., 2002), evaluated honestly on the selected partition rather than on a resampled test set. The third is the use of modern, flow-based datasets such as CICIDS2017 (Sharafaldin et al., 2018) or UNSW-NB15 (Moustafa & Slay, 2015), whose flow-level features are closer to what a streaming-aware detector would require and whose reconciliation with older schemas remains an open problem (Sarhan et al., 2022).

References

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

D'Hooge, L., Wauters, T., Volckaert, B., & De Turck, F. (2020). Inter-dataset generalization strength of supervised machine learning methods for intrusion detection. *Journal of Information Security and Applications*, *54*, 102564. https://doi.org/10.1016/j.jisa.2020.102564

Di Domenico, A., Perna, G., Trevisan, M., Vassio, L., & Giordano, D. (2021). A Network Analysis on Cloud Gaming: Stadia, GeForce Now and PSNow. *Network*, *1*(3), 247–260. https://doi.org/10.3390/network1030015

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization [arXiv:1412.6980]. *3rd International Conference on Learning Representations (ICLR)*.

Layeghy, S., Gallagher, M., & Portmann, M. (2021). Benchmarking the benchmark – analysis of synthetic NIDS datasets [arXiv preprint arXiv:2104.09029].

Lyu, M., Madanapalli, S. C., Vishwanath, A., & Sivaraman, V. (2024). Network Anatomy and Real-Time Measurement of Nvidia GeForce NOW Cloud Gaming. In P. Richter, V. Bajpai, & E. Carisimo (Eds.), *Passive and Active Measurement: 25th International Conference, PAM 2024, Virtual Event, March 11–13, 2024, Proceedings* (Vol. 1). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-56249-5

Moustafa, N., & Slay, J. (2015). UNSW-NB15: A comprehensive data set for network intrusion detection systems. *2015 Military Communications and Information Systems Conference (MilCIS)*, 1–6. https://doi.org/10.1109/MilCIS.2015.7348942

Sarhan, M., Layeghy, S., Moustafa, N., & Portmann, M. (2022). Towards a standard feature set for network intrusion detection system datasets. *Mobile Networks and Applications*, *27*(1), 357–370. https://doi.org/10.1007/s11036-021-01843-0

Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. A. (2018). Toward generating a new intrusion detection dataset and intrusion traffic characterization. *Proceedings of the 4th International Conference on Information Systems Security and Privacy (ICISSP)*, 108–116. https://doi.org/10.5220/0006639801080116

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*(1), 1929–1958.

Tavallaee, M., Bagheri, E., Lu, W., & Ghorbani, A. A. (2009). A detailed analysis of the KDD CUP 99 data set. *2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA)*, 1–6. https://doi.org/10.1109/CISDA.2009.5356528

Verkerken, M., D'hooge, L., Wauters, T., Volckaert, B., & De Turck, F. (2022). Towards model generalization for intrusion detection: Unsupervised machine learning techniques. *Journal of Network and Systems Management*, *30*(1), 12. https://doi.org/10.1007/s10922-021-09615-7

Yoachimik, O., Pacheco, J., & Cloudforce One. (2026, February). *2025 Q4 DDoS Threat Report: A Record-Setting 31.4 Tbps Attack Caps a Year of Massive DDoS Assaults*. Retrieved September 9, 2026, from https://blog.cloudflare.com/ddos-threat-report-2025-q4/

A Framework for Genre-Aware Player Adaptation in LLM-Driven Game AI

Predrag Peharda, David Slavik
University of Zagreb Faculty of Organization and Informatics
Artificial Intelligence Laboratory
Pavlinska 2, 42000 Varaždin, Croatia
pehardap@gmail.com, dslavik@foi.hr

Abstract. *Large language models make it inexpensive to add a "strategic" adaptation layer to game AI: a component that reads how a player behaves and re-shapes an agent's strategy between encounters. This paper argues that the value of such adaptation is genre-contingent. It undermines genres that reward mastering a stable, learnable system, and benefits genres that reward a responsive or living world. The main problem is therefore calibration, not capability: more adaptive is not always better. In this paper we propose a conceptual framework for deciding when such adaptation is appropriate.*

Keywords. game AI, large language models, player modeling, player adaptation, game design, framework

1 Introduction

AI and games share a long history, beginning with programs built to play games as a test of whether a machine could perform tasks that appeared to require intelligence, and progressing through Checkers, Chess, and Go to modern video games (Yannakakis & Togelius, 2025, pp. 4–16).

Over time, the role of AI in games expanded beyond playing them, to generating content and to adapting the game experience to the individual playing it. Early online games collected player data during play and tuned themselves from it, later titles responded to a player's emotional state, and recent systems delegate the memories and behavior of non-player characters to LLMs (Yannakakis & Togelius, 2025, pp. 4–16). Adapting a game to the player has become a common practice rather than an exceptional one, and much of this work proceeds on the assumption that tuning a game more closely to its player produces a better experience.

LLMs change the cost of building such adaptation. An LLM can take an aggregated description of how a player has behaved and re-shape an agent's strategy through a prompt alone, without a trained model or a manually built difficulty controller. This makes it inexpensive to add a strategic adaptation layer to an agent that already exists. Once a capability becomes inexpensive to build, the relevant question is no longer whether it can be built, but whether it should be applied, and in which contexts.

In this paper, we argue that the value of adapting to the player is genre-contingent. Some genres are built around mastering a fixed system: a competitive opponent is satisfying precisely because the player can study it, learn its patterns, and eventually overcome it, and an agent that quietly reshapes itself to match the player removes that source of satisfaction (Cowley & Charles, 2016). Other genres are built around a world that feels alive and responds to the player, such as a narrative RPG or a companion character, and in these the same adaptation is what makes the experience compelling. A single mechanism therefore produces opposite results depending on what the game is designed to offer.

Our contribution is a conceptual framework for making that determination. A three-tier architecture confines the LLM to a strategic role and keeps it out of the real-time control loop. The framework is conceptual, and in this paper it is not empirically validated.

Section 2 reviews player modeling and DDA, LLMs in games, and the planning limits that motivate keeping the model off the real-time loop. Section 3 presents the framework, Section 4 an illustrative example in which a boss enemy and a companion run on the same mechanism to opposite effect, Section 5 discusses what the framework does and does not claim, and Section 6 concludes and outlines future work.

2 Background and Related Work

2.1 Player Modeling and Dynamic Difficulty Adjustment

Player modeling is one of the established uses of AI in games, alongside playing games and generating content. It refers to the detection, prediction, and expression of the cognitive, affective, and behavioral characteristics that a person exhibits while playing (Yannakakis & Togelius, 2025, pp. 315–316).

A player model, in this sense, is a representation of how an individual interacts with a game and responds

to it, learned from observations of their play. Such a model can be put to use at different stages. It may be applied during the analysis of play, during the design process, or during play itself, and it is this last case that the present paper concerns. Yannakakis and Togelius term it adaptation: the game reads a model of the player and modifies itself in response while the player is still playing (Yannakakis & Togelius, 2025, pp. 320–321). The use cases they group under adaptation include dynamic game balancing, matchmaking, and personalized content, as well as experiences tailored through non-player characters (Yannakakis & Togelius, 2025, p. 322).

The most extensively studied form of this kind of adaptation is dynamic difficulty adjustment (DDA), which modifies a game's features and behavior in real time according to the player's skill in order to keep the player engaged. Approaches range from heuristic adjustments based on game state to methods that respond to the player's measured affective state. Reviews of this work are broadly positive, reporting improvements in engagement and related measures, while also noting that player preferences remain incompletely understood and that further study is needed (Zohaib, 2018). The assumption underlying much of this work is that a closer fit between the game and the player produces a better experience.

A common thread across this work is that effective adaptation depends on a model of the player. The player model is what supplies the constraints the adaptive system acts on, rather than the system adjusting in an unprincipled way. This view frames adaptation as the application of a player model rather than as a standalone mechanism (Cowley & Charles, 2016).

2.2 LLMs in Games

The use of LLMs in games has grown quickly, and recent surveys have begun to map the roles such models can play. Gallotta et al. (2024) catalogue the applications of LLMs in and for games and identify the distinct roles a model can take, principally as a player that makes decisions, as a non-player character that controls dialogue and behavior, and as a generator of content such as levels or whole games. Much of the surveyed work concentrates on these roles. LLMs have proven well suited to generating natural-language content, to driving the speech and behavior of characters, and to playing games whose states and actions can be expressed as text (Gallotta et al., 2024).

A parallel line of work studies LLMs specifically as agents that play games. Hu et al. (2024) survey this area and describe such agents in terms of three components: memory, reasoning, and input and output, organizing the literature across six game genres, including adventure, communication, competition, cooperation, simulation, and crafting and exploration. The emphasis throughout is on the model as a participant that perceives a game and acts within it, rather than as a system that reshapes the game around the person playing it (Hu et al., 2024).

What receives far less attention in this literature is the use of an LLM to model and respond to the player. Gallotta et al. (2024) note this directly, identifying player modeling as one of the most promising but least explored directions for future work on LLMs in games. The present paper takes up that direction. Rather than using an LLM to play a game or to author its content, we consider an LLM that reads a model of the player and adjusts the behavior of a game agent in response, and we ask under what conditions doing so is appropriate.

2.3 Planning Limitations of LLMs

A growing body of work questions how far LLMs can be trusted to reason and plan on their own. Kambhampati (2024) argues that, despite their fluency, LLMs do not perform reliable planning in the way a dedicated planner does; what looks like reasoning is better understood as approximate retrieval over patterns seen in training, and the models are poor at verifying or correcting their own plans. The practical consequence is that an LLM cannot be assumed to produce a sound plan, and it cannot be relied upon to notice when its own plan is unsound (Kambhampati, 2024).

This does not mean LLMs have no place in tasks that involve planning. Kambhampati et al. (2024) propose what they call LLM-Modulo frameworks, in which the model generates candidate solutions but an external component checks them against explicit constraints, accepting only those that pass. In this arrangement the LLM contributes ideas while a separate, verifiable system holds responsibility for correctness. The division of labor is deliberate: the model proposes, and the surrounding system disposes (Kambhampati et al., 2024).

These limitations bear directly on where an LLM can sit within a game. A game's moment-to-moment control loop runs on a strict timing budget and demands consistent, correct behavior many times per second, which is precisely the regime in which an unverified and slow generator is least appropriate. The framework proposed in the next section takes this constraint as a starting point. It keeps the LLM out of the real-time loop and confines it to a slower role in which it proposes adjustments, while authored, verifiable game logic retains control over what the agent actually does.

3 The Framework

The framework proposed has three components. The first is an architecture that places the language model within a game agent without letting it compromise the agent's real-time behavior. The second is a heuristic for deciding whether adaptation suits a given game at all. The third is a mechanism for controlling how much

adaptation is allowed once it has been deemed appropriate.

3.1 A Three-Tier Architecture

We organize a game agent into three tiers, separated by the timescale on which they operate and by who or what is responsible for their decisions.

1. The reflex tier governs behavior at the scale of milliseconds: animation, collision and hitbox resolution, target tracking, and the immediate execution of an action once it has been chosen. Decisions at this tier must be fast and consistent, and they must be correct every time. For the reasons set out in Section 2.3, an LLM has no place in this tier. It is too slow to meet the timing budget of a real-time loop, and its output is too variable to guarantee the consistency this tier requires.
2. The tactical tier governs behavior at the scale of seconds: choosing which action or short sequence of actions to perform in the current situation. This tier draws on a fixed set of designer-authored strategies, which we refer to as the agent's tactical vocabulary. Each strategy is written, tested, and balanced by the game's designers, exactly as in a conventional game agent. The tactical tier is where the agent's actual conduct is decided, and it remains under authored control at all times.
3. The strategic tier governs behavior between encounters, on the slowest timescale of the three. The language model operates only at this tier. The model receives an aggregated player profile, a summary of how the player has behaved over prior encounters, and uses it to re-weight the tactical vocabulary: raising the priority of some authored strategies and lowering that of others for the next encounter. The model does not act in real time, and it does not act directly on the game. It changes only the relative emphasis among strategies that the designers have already authored.

This separation is the central safety property of the framework. The language model never invents an action, never executes one, and never participates in the moment-to-moment control of the agent. It proposes a re-weighting of an authored vocabulary, and the tactical and reflex tiers retain full responsibility for what the agent actually does. This arrangement follows the division of labor described in Section 2.3: the model only makes suggestions, while authored and tested game logic determines what the agent does. In the worst case, an unsound suggestion can do no more than shift emphasis among strategies that are all individually safe, since every available action has been authored and tested in advance.

3.2 The Stable-Model versus Living-World Spectrum

The architecture of Section 3.1 describes where adaptation happens and how it is constrained, but not whether it should happen at all. That decision depends on the game. We propose a single heuristic for making it: ask whether the genre rewards the player for building a stable mental model of a fixed system, or for responding to a world that is unpredictable and alive. The first kind of game we call stable-model; the second, living-world. Where a game falls on this spectrum predicts whether strategic adaptation will help or harm it.

At the stable-model end, the player's satisfaction comes from studying a system that holds still long enough to be learned. A "souls-like" boss enemy is the clearest case. The term refers to the punishing, pattern-based combat popularized by the *Dark Souls* series (FromSoftware), in which an enemy's attacks and timings form a fixed challenge that the player is expected to learn through repeated attempts. Mastery is the experience the game offers. A competitive opponent is similar, in that fairness depends on both sides facing rules that do not silently change. For games of this kind, strategic adaptation is corrosive. If the agent quietly re-weights its strategies in response to the player, the system the player is trying to learn becomes a moving target, and the act of mastering it loses its meaning. The very mechanism described in Section 3.1 works against this kind of genre.

At the other end of the spectrum, the living-world case, the player's satisfaction comes from the sense that the world notices them and responds. A companion character that adjusts to the player's habits, or a narrative agent that shifts its behavior as it comes to know the player, is more compelling for doing so. Here the same strategic re-weighting is not a flaw but the point. Adaptation makes the world feel responsive and particular to this player, which is exactly what that kind of genre promises.

This division has a precedent in existing taxonomies of player modeling. Yannakakis and Togelius (2025) distinguish adaptation aimed at player behavior, such as dynamic balancing and matchmaking, from adaptation aimed at player experience, such as an experience tailored through non-player characters. The behavior-oriented cases tend toward our stable-model pole, where what is adjusted is the challenge a player is trying to master, while the experience-oriented cases tend toward our living-world pole, where what is adjusted is how the world feels. Our contribution is not to redraw this distinction but to read it normatively: to treat a game's position on the spectrum as a prediction of whether adaptation should be applied at all.

Most games are not purely of one kind, and a single game may contain both a system meant to be mastered and a world meant to feel alive. For example a "souls-like" game with a companion character has a stable-model combat system alongside a living-world

relationship, and the right decision may differ for each. The spectrum is therefore best applied not to a whole game but to the particular system an agent governs, and to the source of satisfaction that system is meant to provide.

3.3 The Governor

Deciding that a game sits toward the living-world end of the spectrum does not settle how much adaptation it should receive. A companion that re-invents itself after every conversation is as unconvincing as one that never changes at all. The spectrum tells a designer whether to adapt; it does not tell them how far. For that we introduce a governor: a small set of controls that bound the strategic tier's influence over the agent. The governor has two "knobs".

The first is the cap. The strategic tier re-weights the tactical vocabulary, but the cap places a limit on how much weight any single strategy may receive. With a low cap, no authored strategy can come to dominate the agent's behavior, so the agent continues to draw on its full set of actions even as the emphasis shifts. With a high cap, the strategic tier is free to concentrate the agent's behavior around a few strategies that suit the player. The cap therefore controls how strongly adaptation is allowed to reshape the agent in any one direction.

The second is the cadence. The strategic tier acts between encounters, but the cadence sets how often it is permitted to act: after every encounter, after several, or only at designer-chosen moments. A slow cadence means the agent changes rarely, leaving the player a stable target across many encounters before any adjustment occurs. A fast cadence means the agent changes often, responding quickly to how the player is behaving. The cadence therefore controls how readily the player perceives the agent as changing at all.

Together, the two "knobs" convert the spectrum's qualitative judgment into a quantitative setting. A stable-model system is governed tightly: a low cap, so the agent never reshapes itself into a different opponent, and a slow cadence, so what little adaptation occurs does not disturb a player. The limiting case is a cap and cadence set so low that adaptation is switched off entirely, which is the correct setting for a game that must not read its player at all. A living-world system is governed loosely: a higher cap and a faster cadence, so the world can respond visibly and often. Most systems fall between these settings, and the governor is what lets a designer place them there rather than choosing only between adaptation and its absence.

This is the sense in which the problem is calibration rather than capability. The capability, an LLM re-weighting an authored vocabulary, is the same across every game. What differs is how the governor is set, and setting it well is the act of matching adaptation to what a particular system is for.

4 Illustrative Example

To show how the framework operates, we walk the same mechanism through two cases. In both, a game agent is built on the three tiers of Section 3.1, and in both, the strategic tier re-weights an authored tactical vocabulary from a player profile. The cases differ only in where their games sit on the spectrum of Section 3.2, and therefore in how the governor of Section 3.3 is set. The mechanism is identical; the calibration is not.

4.1 A Boss in a Stable-Model Game

Consider a boss enemy in a game whose challenge is meant to be studied and overcome. The agent's tactical vocabulary is a set of authored strategies, each a pattern of attacks and responses that a player can learn to read and counter. Over repeated attempts, a player profile accumulates a summary of how this player behaves: which strategies they handle well and which they struggle against.

If the strategic tier were allowed to act freely on this profile, it would re-weight the vocabulary toward whatever the player handles least well, sharpening the boss against the player's specific weaknesses between attempts. This is precisely the wrong outcome for the genre. The player is trying to learn a fixed system, and an agent that shifts to exploit them turns that system into a moving target. Effort that should yield mastery instead yields a boss that is always one step ahead, and the satisfaction the game is built to provide is lost.

The framework's response is not to remove the mechanism but to govern it tightly. The cap is set low, so no single strategy can come to dominate and the boss remains the same opponent from attempt to attempt. The cadence is set slow, or to zero, so the boss does not change underneath a player who is in the middle of learning it. In the limiting case the governor switches adaptation off, and the agent behaves exactly as a conventional, authored boss would. For a stable-model game, that is the correct setting, and the framework reaches it not by lacking the capability to adapt but by choosing not to apply it.

4.2 A Companion in a Living-World Game

Consider instead a companion character in a game whose appeal is a world that responds to the player. The agent is built identically: the same three tiers, the same kind of authored tactical vocabulary, the same strategic tier re-weighting it from a player profile. Here the profile summarizes the player's habits and tendencies over time, and the vocabulary is a set of authored behaviors the companion can express.

When the strategic tier re-weights this vocabulary toward the player, the companion comes to behave in ways that fit how this particular person plays. This is the outcome the genre wants. The player's sense that

the companion notices and responds to them is the experience on offer, and the adaptation is what produces it.

The framework governs this case loosely. The cap is set higher, so the companion's behavior can concentrate meaningfully around the player rather than only shifting at the margins. The cadence is set faster, so the companion responds within a timescale the player will notice and attribute to being understood. The mechanism is unchanged from the boss; only the governor's settings differ, and they differ because the source of satisfaction differs.

4.3 The Same Mechanism, Calibrated Differently

The two agents are the same construction. Each confines its language model to the strategic tier, each reweights an authored vocabulary, and each is bounded by a cap and a cadence. What separates them is a judgment about what their respective games are for, expressed as two settings of the governor. The boss is governed tightly because mastery requires a stable target; the companion is governed loosely because responsiveness requires a moving one.

5 Discussion and Limitations

The framework is a design heuristic and an architecture, not an empirical result. It claims that the value of strategic player adaptation depends on what a game is for, that this can be predicted from where a system sits on the stable-model versus living-world spectrum, and that the resulting decision can be expressed through the governor's two controls. The one claim it makes firmly is the architectural safety property: because the LLM only re-weights an authored vocabulary, it cannot produce behavior the designers did not write and test.

The framework is conceptual and, in this paper, unvalidated. We present no implementation, no metrics, and no player study, and its central prediction, that adaptation harms stable-model systems and helps living-world ones, is argued from the structure of those genres rather than demonstrated. It is also worth stating that the LLM here is not learning in any machine-learning sense. It is not trained or updated by play; it is re-prompted with a player profile and asked to reweight authored options, so the adaptation is a change of emphasis, not a change in the model.

One question the framework raises but does not settle is whether players enjoy being read. Yannakakis and Togelius (2025) observe that a game could, in principle, come to know more about how a person plays than the person knows themselves. Our framework takes that capability as given and asks where it should be used, but it does not establish how players feel about it, and the same adaptation may read as attentive to one player and intrusive to another.

Finally, the framework does not introduce adaptive game AI, player modeling, or dynamic difficulty adjustment, all of which are mature areas surveyed in Section 2. Its contribution is the narrower, normative question of when an existing capability, made newly inexpensive by LLMs, should be applied at all.

6 Conclusion and Future Work

This work presents a conceptual framework for deciding when strategic player adaptation, now inexpensive to build with LLMs, is appropriate for a game. The framework combines a three-tier architecture that confines the LLM to a slow, strategic role and keeps it out of the real-time loop, a stable-model versus living-world heuristic that predicts whether adaptation helps or harms a given system, and a governor that turns that prediction into concrete settings for how strongly and how often the agent adapts. Its recurring claim is that the design problem is calibration rather than capability: the same mechanism serves one game and damages another, and the difference lies in how it is governed.

Future work could replace the illustrative cases of Section 4 with a working implementation, allowing the boss enemy and companion agents to be built rather than described. Player studies could then test the framework's central prediction directly, measuring whether tightly governed adaptation preserves the experience of a stable-model system while loosely governed adaptation improves a living-world one. Such studies would also address the open question of whether players enjoy being read, which the framework raises but cannot settle. We see the present framework as the conceptual groundwork for that validated follow-up, and as a first treatment of a design space we would call, in its fuller form, that of stable models and living worlds.

References

Cowley, B. U., & Charles, D. (2016). Adaptive artificial intelligence in games: Issues, requirements, and a solution through behavlets-based general player modelling. *arXiv preprint arXiv:1607.05028*.

Gallotta, R., Todd, G., Zammit, M., Earle, S., Liapis, A., Togelius, J., & Yannakakis, G. N. (2024). Large language models and games: A survey and roadmap. *IEEE Transactions on Games*, 1–18. https://doi.org/10.1109/TG.2024.3461510

Hu, S., Huang, T., Ilhan, F., Tekin, S., Liu, G., Kompella, R., & Liu, L. (2024). A survey on large language model-based game agents. *arXiv preprint arXiv:2404.02039*.

Kambhampati, S. (2024). Can large language models reason and plan? *Annals of the New York Academy of Sciences*, *1534*(1), 15–18. https://doi.org/10.1111/nyas.15125

Kambhampati, S., Valmeekam, K., Guan, L., Verma, M., Stechly, K., Bhambri, S., Saldyt, L., & Murthy, A. (2024). LLMs can't plan, but can help planning in LLM-modulo frameworks. *Proceedings of the 41st International Conference on Machine Learning (ICML)*.

Yannakakis, G. N., & Togelius, J. (2025). *Artificial intelligence and games* (2nd ed.). Springer.

Zohaib, M. (2018). Dynamic difficulty adjustment (DDA) in computer games: A review. *Advances in Human-Computer Interaction*, *2018*(1), 5681652. https://doi.org/10.1155/2018/5681652

Conversing with Judita and Werther: Gamified Cognitive Agents for Croatian-Language Literature Learning

Markus Schatten
University of Zagreb Faculty of Organization and Informatics
Artificial Intelligence Laboratory
Pavlinska 2, 42000 Varadin, Croatia
markus.schatten@foi.unizg.hr

Marinela Schatten, Elvira Gotal
School of Construction, Science and Mining
Hallerova aleja 3, 42000 Varadin, Croatia
{marinela.schatten,elvira.gotal}@skole.hr

Abstract. *This work-in-progress paper presents the design, implementation, and preliminary evaluation of two gamified cognitive agents for literature learning, letting students converse in Croatian with Maruli's Judita and Goethe's Werther. Following Design Science Research, the browser-based artifact combines speech recognition, retrieval-augmented large language model dialogue, speech synthesis, and viseme-based facial and behavioral animation. Open models remain limited for Croatian, requiring a cloud-based model. A focus group with ten vocational-school students indicated increased engagement and better understanding of character motivation, archaic language, and historical context; students stressed agents should complement, not replace, reading. We contribute design insights for embodied, game-like literary agents.*

Keywords. gamified cognitive agents, literature learning, conversational agents, large language models, text-to-speech, speech recognition, viseme animation, Croatian language, design science research

1 Introduction

Artificial intelligence has increasingly become a part of contemporary educational technologies, ranging from intelligent tutoring systems and adaptive learning environments to conversational agents and large language model based assistants. In this broader landscape, cognitive agents are especially relevant because they do not merely provide textual answers, but are designed to communicate with users through human-oriented modalities such as natural language, speech recognition, speech synthesis, animated embodiment, and contextual knowledge representation. Such agents can therefore act as social and conversational interfaces that support learning through dialogue, feedback, guidance, and affective engagement (Kuhail et al., 2023; Moreno et al., 2001; Sikström et al., 2022; Zawacki-Richter et al., 2019). Recent developments in Large Language Models (LLMs) have further expanded the possibilities for natural language interaction in education, while at the same time raising important questions regarding reliability, transparency, learner autonomy, and the appropriate pedagogical role of such systems (Kasneci et al., 2023).

The motivation for the work presented in this paper stems from the observation that literature education, particularly when canonical or historically distant works are involved, often requires students to overcome several simultaneous barriers: unfamiliar vocabulary, archaic or non-standard language, complex narrative structures, implicit historical context, and psychological distance from fictional characters. These barriers are especially visible in the reading of older Croatian literature, such as Maruli's *Judita*, but may also occur in foreign classics, such as Goethe's *The Sorrows of Young Werther*, where students must interpret emotional states, motives, social norms, and narrative perspective. Traditional educational support usually takes the form of lectures, notes, summaries, written interpretations, or classroom discussion. While these approaches remain essential, they may not always provide an immediate, personalized, and engaging way for students to ask questions, test interpretations, or explore the inner world of a literary character.

Gamification provides one possible response to this challenge. Commonly understood as the use of game design elements in non-game contexts (Deterding et al., 2011), gamification has been widely investigated as a way of supporting motivation, engagement, and participation, although its effects depend strongly on context and design quality (Hamari et al., 2014). In the context of literature learning, however, gamification need not be limited to points, badges, leader-

boards, or quizzes. A more meaningful game-inspired approach can be achieved through role-play, character embodiment, interactive fiction, and conversational exploration. From this perspective, enabling a student to speak directly with a literary character is a form of educational role-playing: the fictional world becomes partially interactive, while the student is invited to investigate the work by asking questions, comparing interpretations, and reflecting on the character's responses.

This paper continues our previous research on cognitive agents, gamification, and educational applications of artificial intelligence. In earlier work, we introduced the B.A.R.I.C.A. infrastructure as an architecture for Croatian-speaking cognitive agents, including a university student support agent capable of communicating through spoken Croatian (Schatten, Okrea uri, & Peharda, 2021). We further explored the combination of cognitive agents and gamification in a conceptual telemedicine framework, showing how agent-based interaction and gamified mechanisms can be orchestrated in service-oriented systems (Schatten, Okrea uri, & Protrka, 2021). The infrastructure was later generalized toward smart mobility applications, where B.A.R.I.C.A. was presented as an open-source cognitive-agent infrastructure for Croatian-language interaction in various domains (Schatten et al., 2022). In education, we reported a case study on the development of gamified cognitive agents for primary school pupils, where students and teachers participated in the design process and the agents were used to support learning in Croatian school settings (Schatten et al., 2023). More recently, we investigated the integration of cognitive agents into visual novels, demonstrating how agent-based dialogue can extend static branching narratives toward more dynamic and immersive interactive storytelling (Saari & Schatten, 2024). The practical use of cognitive agents in primary education was also presented to the wider educational community through CARNET's CUC 2024 conference (Schatten et al., 2024).

The present work builds on these foundations, but shifts the focus toward high-school literature learning and the gamified embodiment of literary characters. Instead of developing a general-purpose school assistant or integrating an agent into a fictional game world, we implement two cognitive agents that assume the roles of well-known literary figures: Judita from Marko Maruli's *Judita* and Werther from Johann Wolfgang von Goethe's *The Sorrows of Young Werther* (Goethe, 1774; Maruli, 1521). The agents communicate in Croatian and are intended to support students who have already encountered the literary works by allowing them to ask questions, clarify unclear parts of the plot, discuss motives and emotions, and compare the agent's perspective with their own interpretation of the text. In this sense, the goal is not to replace reading, but to create an additional interpretive and motivational layer that can be used before, during, or after reading.

From a technical perspective, the work also extends our previous architecture by combining a speech recognition–LLM–text to speech (TTS) pipeline with retrieval-augmented access to literary texts, browser-based animation, and viseme synchronization for Croatian speech. Retrieval-augmented generation is especially relevant in this setting because literary agents must not only speak fluently, but also remain grounded in the source work and its character logic (Lewis et al., 2020). At the same time, our initial objective was to keep the solution as open and browser-based as possible, making it easier for schools to deploy without dedicated infrastructure. This objective is aligned with the broader development of open and multilingual language models, including open foundation models and recent European multilingual models (Martins et al., 2024; Touvron et al., 2023). Nevertheless, our current implementation still relies on a commercial LLM API because lightweight open models did not yet provide sufficiently reliable Croatian-language literary dialogue in the intended deployment setting. This limitation is important because it highlights a practical gap between the promise of open educational artificial intelligence (AI) and the current requirements of high-quality interaction in smaller languages.

Methodologically, the work follows a design science research (DSR) approach, which is appropriate when the research objective is the construction and evaluation of an innovative artifact that addresses a relevant problem in a real context (Peffers et al., 2007). The artifact presented in this paper consists of two embodied gamified cognitive agents, their supporting backend and frontend architecture, and a preliminary evaluation with high-school students. The study therefore contributes in three main ways. First, it demonstrates how cognitive agents can be used as literary characters rather than only as assistants, tutors, or information providers. Second, it provides implementation insight into Croatian-language speech-based literary agents with animated embodiment. Third, it reports preliminary student reactions that help clarify the pedagogical value and limitations of such systems, especially the tension between increased engagement and the need to preserve reading as an interpretive and imaginative activity.

The rest of this paper is organized as follows. Section 2 provides an overview of related work on cognitive agents, gamification, conversational agents in education, and interactive storytelling. Section 3 presents the DSR methodology used in this study. Section 4 describes the system architecture and implementation of the Judita and Werther agents. Section 5 presents the preliminary focus-group evaluation with high-school students. Finally, Section 6 discusses the main findings, limitations, and directions for future research.

2 Related Work

The work presented in this paper is positioned at the intersection of cognitive agents, educational chatbots, gamification, interactive storytelling, and recent LLM-based role-playing systems. Cognitive and pedagogical agents have a long tradition in educational technology, where they have been investigated as tools for guidance, feedback, motivation, and social interaction (Baylor, 1999; Moreno et al., 2001; Sikström et al., 2022). Educational chatbots represent a closely related research direction. Recent reviews show that chatbots have been used in various educational contexts to answer questions, support practice, and provide immediate feedback, although their effectiveness depends on the quality of instructional design, interaction, and evaluation (Kuhail et al., 2023). The emergence of LLMs has further increased the potential of conversational systems in education, but has also introduced challenges related to accuracy, overreliance, transparency, and pedagogical control (Kasneci et al., 2023).

Gamification is commonly understood as the use of game design elements in non-game contexts (Deterding et al., 2011). In education, it has been studied as a means of increasing motivation and engagement, although its outcomes depend strongly on context and design quality (Dicheva et al., 2015; Hamari et al., 2014). In relation to text-based learning, Smith et al. (2021) proposed a model for automated educational game design from eBooks, showing how textual learning materials can be transformed into game-like educational activities. This is relevant to the present work because both approaches start from literary or textual material and attempt to create a more interactive learning experience. However, while that work focuses on generating educational games and activities from text, our approach focuses on embodied conversational agents that allow students to interact directly with literary characters.

The present paper is also a continuation of our previous research on the B.A.R.I.C.A. cognitive agent architecture. Earlier work introduced a Croatian-speaking cognitive agent for university student support (Schatten, Okrea uri, & Peharda, 2021), proposed the combination of cognitive agents and gamification in telemedicine (Schatten, Okrea uri, & Protrka, 2021), and generalized the infrastructure toward smart mobility applications (Schatten et al., 2022). In education, we reported a case study on gamified cognitive agents for primary schools, where students and teachers participated in the design of Croatian-speaking agents and used them in experimental school sessions (Schatten et al., 2023). Most closely related to the present paper is our recent work on integrating cognitive agents into visual novels, where agent-based dialogue was used to transform static dialogue structures into more dynamic and immersive narrative interaction (Saari & Schatten, 2024). The current work extends this direction from an original visual-novel setting to canonical literature learning.

Interactive storytelling and character-based interaction provide an additional foundation. Visual novels and interactive fiction traditionally rely on pre-authored dialogue and branching choices, whereas AI-based characters can support more flexible natural-language interaction (Cavallaro, 2009; Riedl & Bulitko, 2013; Schatten, Peharda, & Rasonja, 2021). Work on autonomous characters and multimodal conversational agents, such as cognitive modeling for games and animation (Funge, 1999) and the ODO multimodal chatbot for interactive theater (Bhushan et al., 2020), demonstrates the broader potential of believable agents in narrative media. More recently, LLM-based role-playing systems have explored how language models can imitate personas and fictional characters (Chen et al., 2024; Li et al., 2023). Such systems are promising, but they are usually entertainment-oriented and rarely evaluated in formal educational contexts, especially for smaller languages such as Croatian. Retrieval-augmented generation can help ground such agents in source texts (Lewis et al., 2020), which is particularly important when literary agents are expected to remain consistent with a novel, character, and pedagogical purpose.

Compared with previous work, the contribution of this paper is the implementation and preliminary evaluation of embodied, Croatian-speaking cognitive agents that take the roles of Judita and Werther. Rather than acting as general tutors or assistants, the agents function as gamified literary characters intended to complement reading, support interpretation, and enable students to explore motives, emotions, context, and character perspective through speech-based interaction.

3 Methodology

This study follows a DSR methodology, since the main research contribution is the design, implementation and preliminary evaluation of an educational software artifact (Peffers et al., 2007). The problem identification phase was motivated by the observed difficulty of motivating students to read and interpret literary works, especially when the works are linguistically, historically or emotionally distant from their everyday experience. In this context, literacy is understood not only as the ability to read a text, but also as the ability to understand characters, motives, context and style. The problem is particularly relevant for older or canonical texts such as Maruli's *Judita* and Goethe's *The Sorrows of Young Werther*, where students may struggle with archaic language, complex emotions, and unfamiliar cultural contexts.

The objective of the solution was to develop a gamified interaction in which students could communicate with literary characters directly, using spoken Croatian

language. The artifact was expected to support, rather than replace, reading by enabling students to ask questions, clarify unclear passages, explore character motivation and prepare for classroom discussion. Additional objectives were to provide embodied agents with speech animation, to ground the agents in the corresponding literary works, to make the system reusable for multiple agents, and to move as much of the interaction as possible toward browser-based deployment.

The design and development phase was iterative. Several prototypes were developed and tested informally by students, a vocational school teacher of Artificial Intelligence, and a school librarian. These iterations focused on four main issues: the quality of character prompts, the reliability of Croatian-language conversation, the usefulness of retrieval-augmented access to the literary texts, and the believability of the animated speech interface. Early prototypes were revised when open lightweight LLMs produced insufficient Croatian-language responses or when the interaction flow was not sufficiently natural for classroom use. Later iterations refined the viseme animation, idle expressions, RAG ingestion, agent configuration, and the speech-recognition–LLM–TTS pipeline.

The demonstration phase consisted of presenting the final artifacts during a regular elective Artificial Intelligence course in a vocational high school. In addition, the artifacts were presented at two independent professional meetings of informatics teachers, a county-level professional meeting and an inter-county event. Feedback from these professional presentations was positive and confirmed the perceived relevance of the approach for classroom experimentation, although these presentations were not treated as formal evaluation. The formal evaluation reported in this paper was based on a focus group of students, described in Section 5.

4 Implementation

The implemented system consists of a multi-agent backend, a browser-based animated frontend, a local RAG store for each agent, and speech input/output components. The overall architecture is shown in Fig. 1. Each agent is stored as a separate folder containing an `agent.json` configuration file, visual assets, audio assets, RAG documents, and a local SQLite RAG store. In the current prototype, the two literary agents are *Judita*, grounded in *Judita*, and *Werther*, grounded in the Croatian translation of *The Sorrows of Young Werther*. The same directory structure also supports additional agents, which makes the system extensible.

The backend exposes REST-like endpoints for each agent, including health, RAG status and chat interaction. At runtime, the server loads the selected agent configuration, retrieves relevant context from the local RAG store when available, sends the user question and system instructions to the LLM, and returns a cleaned response suitable for spoken output. Conversation state is preserved through response identifiers, allowing follow-up questions within a session. The RAG component indexes source documents by extracting text, chunking it, embedding the chunks, and storing them in SQLite. During interaction, the user question is embedded and compared with stored chunks, after which the most relevant passages are added to the model input. This design allows the same backend to serve multiple character agents while keeping their literary knowledge bases separate.

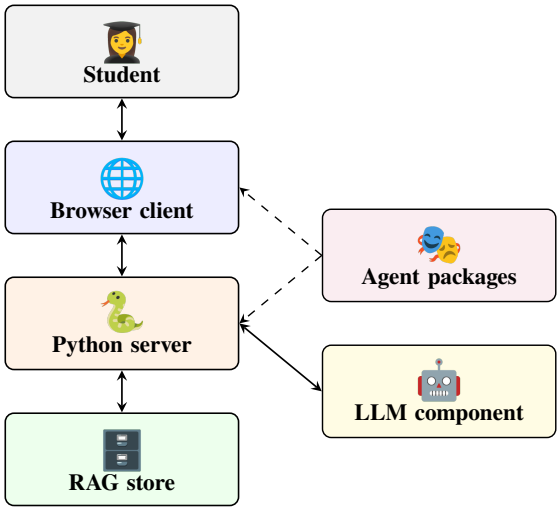


Figure 1. System architecture of the implemented multi-agent platform

Prompt engineering played an important role in shaping both agents as pedagogically useful literary characters rather than generic conversational assistants. For each agent, the system prompt defined the character identity, language, tone, stylistic constraints, relation to the source text, and expected educational function, while also limiting answers to short, speech-friendly Croatian utterances suitable for TTS output. The prompts instructed the agents to remain in character, answer from the character's perspective, avoid contradicting the literary work, and support students' understanding of motives, emotions, context, and themes. Special attention was required for Werther, since the original character is strongly associated with emotional suffering and suicide. Therefore, his prompt explicitly prohibits the romanticization of suffering, obsessive love, violence, self-harm, or suicide, and defines a safety override: if a student expresses suicidal ideation or self-harm intent, the agent must immediately leave the literary role, respond calmly and supportively, encourage the student to seek help from a trusted adult or emergency service, and avoid poetic descriptions or any practical details related to self-harm. This configuration was necessary to preserve the educational and literary value of the role-play while reducing the risk of unwanted or harmful advice during interaction with students.

The frontend is implemented in PyGame and transformed for web deployment using PygBag/WebAssembly. The web layer collects the student's spoken question through the browser speech-recognition interface, sends it to the backend, receives the generated answer, and forwards the text to the TTS component. The generated speech is then played by the PyGame frontend, which simultaneously animates the character. The characters (see Fig. 2 (a) and (b)) were initailly designed by students and later refined by artificial intelligence (namely ChatGPT 5.5). Since the application targets Croatian speech, the lip-sync module maps Croatian letters and digraphs to a reduced set of visemes (see example spritesheet in Fig. 2 (c)). The implemented viseme set includes rest position, vowels, bilabial consonants, labiodentals, sibilants, postalveolar sounds, dental/alveolar sounds, velar sounds and Croatian-specific digraphs such as *lj*, *nj* and *d*. Additional overlays implement blinking, eyebrow movements, gaze shifts and soft smiles, making the agents appear more natural during both speaking and idle states.

The current implementation therefore combines open browser technologies, reusable agent folders, local RAG stores and animated speech synthesis. The main compromise concerns the LLM component. Although the initial goal was a fully in-browser and open implementation, current lightweight open models did not provide sufficiently reliable Croatian-language literary dialogue. For this reason, the present prototype uses a commercial LLM API, while the remaining architecture is designed so that the language model component can be replaced when suitable local or browser-based models become available.

5 Evaluation

The evaluation was conducted after the classroom demonstration of both agents. The focus group consisted of ten vocational high-school students enrolled in the second and third grade, aged 15–17. The group was gender-balanced, with five male and five female students. All participants had previously been assigned both literary works. Since all students were enrolled in the elective Artificial Intelligence course, the sample is likely biased toward students who are more open to technology and AI-based learning tools than the general student population.

Students interacted with both agents for approximately 45 minutes during their regular Artificial Intelligence class. After the interaction, they were asked to write a short essay about their experience. The essays were analyzed using an LLM-assisted thematic analysis. To make this step transparent, the following prompt was used:

> Analyze the following Croatian student essays about interacting with two literary cognitive agents, Judita and Werther. Identify recurring themes, positive reactions, negative reactions, perceived educational benefits, perceived risks, preferred use cases, and comments on whether the agents should replace or complement reading. Do not invent data. Treat the sample as small and exploratory. Report findings in concise academic language and mention possible bias because the students attend an elective Artificial Intelligence course.

The analysis showed that most students perceived the interaction as novel, interesting and more dynamic than ordinary summaries or notes. Several students emphasized that asking direct questions to a character helped them understand motives, emotions and decisions, especially in the case of Werther. The agents were also perceived as useful for works with archaic language, difficult style or unfamiliar historical context. This was especially relevant for *Judita*, but students also mentioned other demanding school readings as possible candidates for similar treatment.

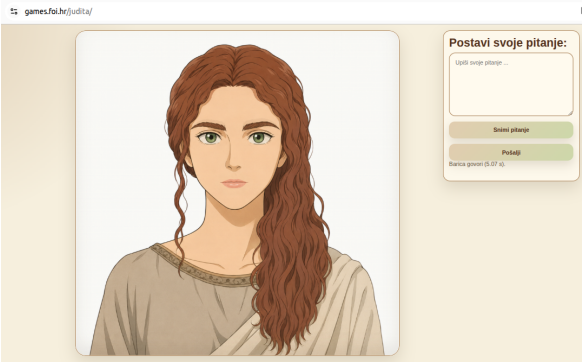
A second important finding is that students generally did not see the agents as a replacement for reading. The dominant view was that the agents are most useful as a complement before a test, after reading, during revision, or while clarifying difficult parts of the work. Several students explicitly warned that reading supports imagination, personal interpretation and attention to literary style, while the agent provides a ready-made representation of the character. This confirms that the pedagogical value of the artifact lies primarily in scaffolding interpretation and discussion, not in substituting the original text.

The essays also revealed limitations. One student considered the approach mostly unnecessary or distracting, arguing that motivated readers do not need such support and that the central novelty is only the imitation of a fictional character. Some students noted that relying on an agent could lead to incomplete knowledge if students ask only a narrow set of questions. One student found the interaction unusual and somewhat difficult to understand. These comments are important because they show that the artifact should be embedded carefully into teaching practice, with clear instructions that it is an auxiliary interpretive tool.

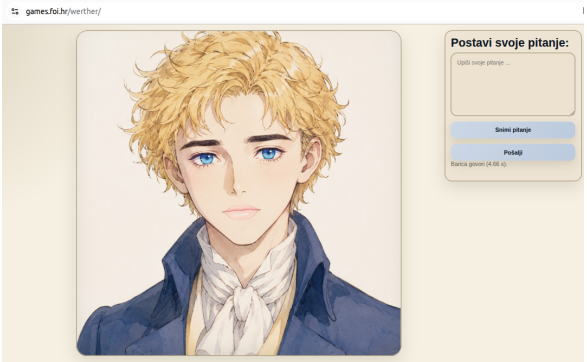
Overall, the preliminary evaluation suggests that embodied literary agents can increase engagement and help students approach literary characters from a new perspective. The results are exploratory due to the small and biased sample, but they support further investigation with larger groups, comparison classes and more structured evaluation instruments.

6 Conclusion and Future Research

This paper presented the design, implementation, and preliminary evaluation of two gamified cognitive agents for supporting Croatian-language literature learning. The agents embody Judita from Maruli's *Judita* and Werther from Goethe's *The Sorrows of Young Werther*, allowing students to interact with literary



(a) Judita



(b) Werther

(c) Judita viseme sprite sheet

Figure 2. Implemented literary agents and an example viseme sprite sheet

characters through spoken Croatian dialogue. The implemented artifact combines speech recognition, LLM-based dialogue, retrieval-augmented access to literary texts, TTS, viseme-based lip synchronization, and additional facial animations in order to provide a more natural and engaging interaction.

The work continues our previous research on cognitive agents and gamification in education, but shifts the focus from general educational support and interactive storytelling toward literature learning. The preliminary evaluation with high-school students suggests that such agents can make literary works more approachable, especially when students struggle with older language, complex emotional states, or historical context. Students particularly valued the possibility of asking direct questions and receiving answers from the perspective of a character. At the same time, the evaluation clearly indicates that the agents should not be understood as a replacement for reading. Their most appropriate role is complementary: they can support interpretation, revision, preparation for classroom discussion, and clarification of difficult parts of the text.

Several limitations need to be acknowledged. The evaluation was exploratory and included only ten students, all of whom were enrolled in an elective Artificial Intelligence course. This may have influenced their openness toward the system and limits the generalizability of the findings. In addition, the current implementation still depends on a commercial LLM API, although the original goal was to create a fully browser-based and serverless solution using open components.

Future research will therefore proceed in two main directions. First, we plan to test emerging minimalistic and lightweight LLMs with respect to their Croatian-language capabilities, literary reasoning, role consistency, and feasibility for browser-based execution. The goal is to gradually replace the backend-dependent LLM component and move toward a fully serverless implementation that schools could deploy without dedicated infrastructure or external commercial services. Second, we plan to evaluate the system with larger and more diverse groups of students, including students

who are not enrolled in AI-related courses. Future evaluations should include pre- and post-tests, comparison groups, teacher assessments, and more structured instruments for measuring comprehension, motivation, engagement, and perceived usefulness. Such studies will make it possible to better determine when and how gamified literary agents can meaningfully contribute to literature education.

Acknowledgement

This research has been financed by NextGenEU project AI-ARENA Framework for Ethical use of AI Agents in Multidisciplinary Research of the Digital Society (Okvir za etiku upotrebu AI agenata u multidisciplinarnim istraivanjima digitalnog drutva).

The authors acknowledge that ChatGPT 5.5 by OpenAI has been used to assist in editing and streamlining sections of this paper.

References

Baylor, A. L. (1999). Intelligent agents as cognitive tools for education. *Educational Technology*, *39*(2), 36–40.

Bhushan, R., Kulkarni, K., Pandey, V. K., Rawls, C., Mechtley, B., Jayasuriya, S., & Ziegler, C. (2020). ODO: Design of multimodal chatbot for an experiential media system. *Multimodal Technologies and Interaction*, *4*(4), 68. https://doi.org/10.3390/mti4040068

Cavallaro, D. (2009). *Anime and the visual novel: Narrative structure, design and play at the crossroads of animation and computer games*. McFarland.

Chen, N., Deng, Y., & Li, J. (2024). The oscars of ai theater: A survey on role-playing with language models. *arXiv preprint arXiv:2407.11484*. https://arxiv.org/abs/2407.11484

Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). From game design elements to gamefulness: Defining gamification. *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments*, 9–15. https://doi.org/10.1145/2181037.2181040

Dicheva, D., Dichev, C., Agre, G., Angelova, G., et al. (2015). Gamification in education: A systematic mapping study. *Educational Technology & Society*, *18*(3), 75–88.

Funge, J. D. (1999). *Ai for games and animation: A cognitive modeling approach*. A K Peters.

Goethe, J. W. v. (1774). *Die leiden des jungen werthers*. Weygand.

Hamari, J., Koivisto, J., & Sarsa, H. (2014). Does gamification work? a literature review of empirical studies on gamification. *Proceedings of the 47th Hawaii International Conference on System Sciences*, 3025–3034. https://doi.org/10.1109/HICSS.2014.377

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kuhail, M. A., Alturki, N., Alramlawi, S., & Alhejori, K. (2023). Interacting with educational chatbots: A systematic review. *Education and Information Technologies*, *28*(1), 973–1018. https://doi.org/10.1007/s10639-022-11177-3

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, *33*, 9459–9474.

Li, C., Leng, Z., Yan, C., Shen, J., Wang, H., Mi, W., Fei, Y., Feng, X., Yan, S., Wang, H., Zhan, L., Jia, Y., Wu, P., & Sun, H. (2023). ChatHaruhi: Reviving anime character in reality via large language model. *arXiv preprint arXiv:2308.09597*. https://doi.org/10.48550/arXiv.2308.09597

Martins, P. H., Fernandes, P., Alves, J., Guerreiro, N. M., Rei, R., Alves, D. M., Pombal, J., Farajian, A., Faysse, M., Klimaszewski, M., Colombo, P., Haddow, B., de Souza, J. G. C., Birch, A., & Martins, A. F. T. (2024). EuroLLM: Multilingual language models for europe. *arXiv preprint arXiv:2409.16235*. https://arxiv.org/abs/2409.16235

Maruli, M. (1521). *Judita*.

Moreno, R., Mayer, R. E., Spires, H. A., & Lester, J. C. (2001). The case for social agency in computer-based teaching: Do students learn more deeply when they interact with animated pedagogical agents? *Cognition and Instruction*, *19*(2), 177–213. https://doi.org/10.1207/S1532690XCI1902_02

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, *24*(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

Riedl, M. O., & Bulitko, V. (2013). Interactive narrative: An intelligent systems approach. *AI Magazine*, *34*(1), 67–77. https://doi.org/10.1609/aimag.v34i1.2449

Saari, S., & Schatten, M. (2024). Integrating cognitive agents into visual novels. *Proceedings of the 35th Central European Conference on Information and Intelligent Systems*, 27–32.

Schatten, M., Okrea uri, B., & Peharda, T. (2021). A cognitive agent for university student support. *Proceedings of the 2021 IEEE Technology & Engineering Management Conference-Europe (TEMSCON-EUR)*, 198–203. https://doi.org/10.1109/TEMSCON-EUR52034.2021.9488627

Schatten, M., Okrea uri, B., Peharda, T., & Tomii, I. (2022). A cognitive agent's infrastructure for smart mobility. *Transportation Research Procedia*, *64*, 199–204. https://doi.org/10.1016/j.trpro.2022.09.024

Schatten, M., Okrea uri, B., & Protrka, R. (2021). Conceptual architecture of a cognitive agent for telemedicine based on gamification. *Central European Conference on Information and Intelligent Systems*, 3–10.

Schatten, M., Peharda, T., & Rasonja, J. (2021). A game engine layer for the implementation of massively multiplayer on-line interactive fiction. *Central European Conference on Information and Intelligent Systems*, 11–16.

Schatten, M., Schatten, M., & Inkret-Martinevi, L. (2023). Development of gamified cognitive agents for primary schools – a case study. *Proceedings of the 34th Central European Conference on Information and Intelligent Systems*, 3–8.

Schatten, M., Schatten, M., & Inkret-Martinevi, L. (2024). Kognitivni agenti u osnovnom kolstvu – iskustva i primjeri [CROSBI ID: 853268]. *26. CARNET-ova konferencija za korisnike CUC 2024: AI u obrazovanju – novi poetak ili poetak kraja?*

Sikström, P., Valentini, C., Sivunen, A., & Kärkkäinen, T. (2022). How pedagogical agents communicate with students: A two-phase systematic review. *Computers & Education*, *188*, 104564. https://doi.org/10.1016/j.compedu.2022.104564

Smith, G. G., Okrea uri, B., & Ricija, M. (2021). Automated educational game design from ebooks. *Proceedings of the 32nd Central European Conference on Information and Intelligent Systems*, 17–24.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*. https://arxiv.org/abs/2302.13971

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? *International Journal of Educational Technology in Higher Education*, *16*(1), 39. https://doi.org/10.1186/s41239-019-0171-0

EduGameGen: A Framework for Educational Game Generation with Large Language Models

Markus Schatten
University of Zagreb Faculty of Organization and Informatics
Artificial Intelligence Laboratory
Pavlinska 2, 42000 Varadin, Croatia
markus.schatten@foi.unizg.hr

Abstract. *This work-in-progress paper presents EduGameGen, a design science artifact for prompt-driven creation and evaluation of educational game prototypes in PyGame using large language models (ChatGPT). It extends the Role–Context–Command–Format prompt structure by decomposing the context into priority-ordered educational, technical/usability, and game-design aspects, and includes a teacher-friendly rubric aligned with these aspects for non-technical evaluators. An artificial, researcher-led evaluation comprised 24 iterative prompt-refinement attempts across five educational game types. Final prototypes achieved a mean score of 32.60/36 with no weak rubric scores, suggesting structured context engineering and rubric-based iteration support rapid prototyping, while classroom evaluation remains necessary for assessing learning effectiveness.*

Keywords. educational games, prompt engineering, ChatGPT, PyGame, PygBag, vibe coding, design science research, RCCF prompt format, educational game evaluation

1 Introduction

Educational games can enrich teaching through interaction, feedback, narrative, challenge, and repeated practice. However, their classroom value depends on more than playfulness: a useful educational game must be aligned with learning outcomes, appropriate for learners' age and prior knowledge, technically usable, and feasible within the time constraints of a lesson. This is a demanding combination for teachers, who often know their subject, students, language, and classroom context well, but may not have the time, programming skills, or artistic resources needed to create custom games. As a result, the practical adoption of game-based learning and gamification remains limited by the availability of games that match concrete teaching needs (Clark et al., 2016; Dicheva et al., 2015; Hamari et al., 2014).

The need for custom-tailored games is particularly clear in primary and secondary education. Teachers may want a short game for introducing cyberbullying, practicing Python syntax, discussing traffic safety, or exploring mechanical systems. Such games should reflect the curriculum, learner characteristics, cultural context, accessibility needs, and available teaching materials. Off-the-shelf educational games rarely satisfy all of these requirements simultaneously. Generative artificial intelligence and Large Language Models (LLMs) therefore create a new opportunity: through prompt-based or "vibe coding" workflows, teachers may be able to describe a learning situation in natural language and obtain a playable prototype. Yet educational game generation is not only a coding task. It requires coordinated control over educational content, technical constraints, game mechanics, visual assets, and deployment.

This paper proposes *EduGameGen*, a design science artifact for prompt-driven educational game prototyping with ChatGPT and PyGame. The framework builds on the Role–Context–Command–Format (RCCF) prompt structure and expands its context component into priority-ordered educational, technical/usability, and game-design aspects, following the broader idea that structured prompt engineering can improve the quality and controllability of generated outputs (Korzyski et al., 2023; Sahoo et al., 2024; White et al., 2023). In its extended form, EduGameGen supports a multiphase workflow: generation of a short game development document (GDD), PyGame code, and graphics specification; optional generation and quality control of visual assets; and final integration for local execution or browser deployment using PygBag (Pygame Community, 2026; pygame-web contributors, 2026).

The study continues the author's previous work on gamification, educational games, design science approaches to game-based systems, and cognitive agents in learning environments (Saari & Schatten, 2024; Schatten & Schatten, 2020; Schatten, 2024; Schatten, Peharda, et al., 2022; Schatten & Schatten, 2019; Schatten et al., 2021, 2023; Tomii & Schatten, 2020). Its goal is not to claim that LLMs can replace teach-

ers or game developers, but to investigate whether a structured prompt framework can help transform teachers' instructional intentions into small, playable, and evaluable prototypes. The paper contributes: (1) the EduGameGen prompt-construction framework; (2) its extension toward graphics specification, quality control, integration, and PygBag-compatible deployment; and (3) a teacher-friendly rubric for evaluating generated prototypes along educational, technical/usability, and game-design dimensions. The artifact is evaluated through an artificial, researcher-led design science research (DSR) evaluation based on 24 iterative prompt-refinement attempts across five educational game types.

2 Related Work

EduGameGen builds on research in educational games, prompt engineering, and AI-assisted game creation. Prior game-based learning and gamification studies emphasize that game elements are educationally useful only when they are aligned with learning activities, feedback, and instructional goals rather than added as decoration (Clark et al., 2016; Dicheva et al., 2015; Hamari et al., 2014). Recent work on LLMs in education extends this argument by suggesting that playful and game-based learning can help preserve practice, motivation, and human judgment while using generative AI, but also stresses the need for domain expertise and critical evaluation of generated outputs (Huber et al., 2024; Kasneci et al., 2023). EduGameGen follows this line by treating the teacher not as a passive recipient of generated code, but as the author of the educational context and evaluator of the resulting prototype.

A second stream concerns LLMs in and for games. Surveys describe rapidly growing work on LLM-supported game design, narrative generation, NPCs, player modelling, and game development tools (Gallotta et al., 2024; Sweetser, 2024). Related educational work has used LLMs as interactive agents inside meaningful games; for example, Prompt-Gaming explored an LLM-evaluating agent in an energy game and reported potential for easy prototyping of educational games (Isaza-Giraldo et al., 2024). In contrast, the present work does not embed an LLM inside the running game. Instead, it uses the LLM as a design and development partner before deployment, producing static PyGame/PygBag prototypes that can be evaluated and revised before classroom use.

The closest studies are recent papers on ChatGPT-assisted educational game creation. Borisova et al. propose a methodology for creating educational games with ChatGPT and discuss examples and risks in school computer-technology education (Borisova et al., 2024). Hadzhikoleva et al. similarly describe an AI-driven, iterative process in which teachers formulate requirements, use multiple prompt cycles, and test the resulting game code (Hadzhikoleva et al., 2024). Spasi et al. focus explicitly on teachers without coding experience and show how a carefully designed prompt can generate a browser-based memory game (Spasi et al., 2025). EduGameGen extends these works by formalizing the prompt as a priority-ordered educational, technical/usability, and game-design context; adding a multiphase workflow for GDD generation, graphics specification, quality control, integration, and web deployment; and linking prompt construction to a teacher-friendly evaluation rubric.

Methodologically, the paper is situated in DSR, where artifacts are built and evaluated with respect to a problem context (Hevner et al., 2004; Peffers et al., 2007). This continues the author's previous work on educational games, gamification, cognitive agents, and design-oriented game systems (Saari & Schatten, 2024; Schatten & Schatten, 2020; Schatten, 2024; Schatten, Peharda, et al., 2022; Schatten & Schatten, 2019; Schatten et al., 2021, 2023; Tomii & Schatten, 2020). The present contribution differs by focusing on prompt construction and evaluation as the central artifact, rather than on a single educational game or game-based system.

3 EduGameGen Framework

EduGameGen is a prompt-construction framework for generating small educational games with LLMs. It adapts the RCCF structure to educational game development: the *Role* defines the model as an educational game designer, instructional designer, and Python/PyGame developer; the *Command* requests a complete playable prototype; and the *Format* fixes the expected output, including the concept, code, controls, running instructions, and self-checklist. The main contribution is the *Context* component, which is decomposed into optional educational, technical/usability, and game-design aspects, as described in Table 1.

All context fields are optional. This is important because teachers may begin with only a topic, a learning outcome, or a short description of a classroom need. Empty fields are treated as unspecified and may be inferred by the model. However, the fields are ordered by priority: when a prompt contains many clues or conflicting requirements, earlier items should guide the generation more strongly than later ones. This makes the framework usable both as a minimal prompt scaffold and as a detailed specification instrument.

The educational context expresses what the game should teach and to whom. Its highest-priority elements are the learning outcome and content, since these constrain the correctness and relevance of the generated game. The technical/usability context limits the implementation space. In the present work, this is intentionally restricted to PyGame to keep the generated games inspectable, modifiable, and suitable for teachers with limited technical support. The game-design context turns the learning activity into gameplay by

Table 1. Priority-ordered EduGameGen context model.

Context group	Optional aspects, ordered from higher to lower priority
Educational	learning outcome; learning materials/content; topic; learner profile; prior knowledge; learner diversity and accessibility considerations; required feedback; assessment/progress logic; cognitive level; misconceptions; subject/domain.
Technical/usability	external dependencies; external assets; code structure; runtime environment; interaction method; screen and interface requirements; accessibility requirements; testing/running instructions.
Game design	core mechanic; link between mechanic and learning outcome; goal; rules; feedback loop; challenge; win/loss conditions; progression; reward/scoring; session length; genre; aesthetics/theme; narrative.

defining player actions, goals, rules, feedback, and progression. The explicit requirement to state the link between the core mechanic and the learning outcome is intended to prevent purely decorative gamification, such as adding points or a story without changing the learning activity.

The current version of EduGameGen further extends the basic RCCF prompt into a three-phase workflow. In the first phase, the model produces a short GDD, a complete PyGame prototype, and, when external graphics are desired, a separate graphics specification. During this phase, only placeholder graphics are used so that the logic of the game can be checked independently from the final visual assets. In the second phase, the specified graphics are generated and quality-controlled against the GDD and graphics specification. In the third phase, the final assets are integrated into the game, the prototype is rechecked, and the deliverable files are prepared.

This multiphase structure was introduced because educational games often require both executable logic and coherent visual presentation. For example, a visual-novel game about cyberbullying may require consistent characters, scenes, speech bubbles, and alternative endings. Separating the generation of code and visual assets reduces the risk that graphics generation distracts from educational alignment or that attractive images hide incomplete gameplay. The final integration step may embed generated assets directly into the Python file, for example by using base64 encoding, which supports the goal of producing self-contained prototypes.

Finally, EduGameGen treats browser deployment as a design requirement rather than a later technical afterthought. The generated games are expected to run locally and, where possible, in a web browser through PygBag, which packages Python/PyGame projects for execution in modern browsers (pygame-web contributors, 2026). This enables the creation of static, client-side educational games that do not require a custom backend. For schools, this is practically important: a game can be hosted as ordinary web content on a school website or learning platform, making it easier for teachers to share custom-tailored prototypes with pupils without installing Python or maintaining a server application.

Table 2. Generated educational game types.

Game	Description
Cyberbullying visual novel	A fifth-grade visual novel about cyberbullying. The story starts when pupils share humiliating meme images of a classmate. The player, also a pupil in the class, chooses how to react; the branching narrative leads to multiple endings.
Portals rogue-like	A teacher-facing dungeon crawler projected in class. When the teacher-controlled player touches a monster, the class answers a question about educational portal websites for second-grade Informatics. Correct answers remove monsters; finding the treasure after at least three correct answers gives the class a reward.
Traffic RPG	A second-grade role-playing game (RPG) about traffic rules and signs. The pupil-player solves quests, collects hidden traffic-related items, answers quizzes, gains experience, and levels up; each level changes the player's class and appearance.
Python 3D puzzle	A programming game focused on Python `for` loops. The player controls 3D objects by writing code with predefined `move` and `rotate` functions, aiming to pass objects through holes across ten progressively harder tasks with hints.
Cogwheels sandbox	A mechanical sandbox for experimenting with cogwheels of different sizes, direct connections, and straight or crossed belts. Any cogwheel can become a motor; the simulation shows rotation speeds and detects mechanical jams.

4 Evaluation Method

EduGameGen was evaluated through an artificial, formative DSR evaluation rather than immediate classroom deployment (Peffers et al., 2007). The goal was not to measure learning gains, but to examine whether the framework can support the generation of prototypes that are technically feasible, pedagogically plausible, and sufficiently game-like for later pilot use with teachers and learners.

The evaluation used 24 generated prototypes across five game types. For each game type, the prompt was iteratively refined until a usable prototype was obtained. A prototype was considered acceptable only if none of the twelve rubric criteria received a score below adequate. The five final games, shown in Fig. 1, were intentionally diverse in topic, genre, audience, and interaction style, as summarized in Table 2.

The rubric follows the three EduGameGen context dimensions. Educational quality includes clarity of the learning objective, correctness of content, alignment between learning and activity, and feedback. Technical/usability quality includes whether the prototype runs, ease of use, readability/accessibility, and stability. Game-design quality includes clarity of goals and rules, challenge/progression, engagement, and quality of the core mechanic. Each criterion was scored from 0 to 3, resulting in a maximum of 36 points. The evaluation was researcher-led and should therefore be interpreted as evidence of prototype feasibility and design

Table 3. Final prototype scores.

Prototype	E	T	G	Total	Class
Cyberbullying visual novel	12	11	12	35	Strong
Portals rogue-like	12	12	11	35	Strong
Traffic RPG	11	11	11	33	Strong
Python 3D puzzle	9	10	9	28	Usable
Cogwheels sandbox	10	10	12	32	Strong

Table 4. Prompt-refinement effort by game type.

Game type	Attempts	First	Final	Gain
Cyberbullying visual novel	8	18	35	+17
Portals rogue-like	2	30	35	+5
Traffic RPG	9	21	33	+12
Python 3D puzzle	4	18	28	+10
Cogwheels sandbox	1	32	32	0

plausibility rather than evidence of learning effectiveness.

5 Results

Across all 24 attempts, the mean score was 25.46 out of 36. Non-final attempts achieved a mean score of 23.58, whereas the final accepted prototypes achieved a mean score of 32.60. Four final prototypes were classified as strong, and one was classified as usable, as shown in Table 3. All final prototypes met the stricter acceptance criterion of having no individual rubric criterion below adequate.

The number of required attempts differed notably across game types, as presented in Table 4. The cogwheels sandbox reached an acceptable result in one attempt, while the portals rogue-like required two attempts. More complex or less straightforward game types required more refinement: the Python 3D puzzle required four attempts, the cyberbullying visual novel required eight attempts, and the traffic RPG required nine attempts. The largest improvement was observed for the cyberbullying visual novel, which increased from 18 to 35 points, followed by the traffic RPG, which increased from 21 to 33 points.

The most frequent weak points across all attempts were ease of use, game mechanic quality, readability/accessibility, clarity of goals and rules, and learning–activity alignment. In contrast, all attempts received strong scores for being runnable prototypes and for stability during play. This suggests that the main bottleneck was not generating executable PyGame code, but shaping the generated artifacts into coherent, usable, and educationally aligned games. The screenshots in Fig. 1 illustrate the diversity of genres and interfaces produced by the same framework, ranging from code-oriented puzzles to visual-novel and sandbox mechanics.

6 Discussion

The findings return to the central motivation of this paper: teachers often need short, curriculum-specific games, but do not usually have the time, programming experience, or graphical resources required to build them from scratch. EduGameGen addresses this gap by turning a teacher's instructional intention into a structured prompt specification. The five prototypes show that the same framework can support substantially different classroom purposes: a branching story for discussing cyberbullying, a teacher-led dungeon crawler for whole-class questioning, an RPG for traffic-safety quests, a code-based 3D puzzle for Python loops, and a mechanical sandbox for experimentation. This diversity is important because it suggests that the framework is not tied to a single genre or subject area, but to a reusable way of specifying educational, technical/usability, and game-design context.

Compared with prior work on ChatGPT-assisted educational game creation (Borisova et al., 2024; Hadzhikoleva et al., 2024; Spasi et al., 2025), EduGameGen makes three practical advances. First, it formalizes the prompt context as a priority-ordered and optional structure rather than as a single monolithic prompt. This allows a teacher to start with minimal information, such as a topic and learning outcome, while still supporting detailed specification when materials, learner characteristics, accessibility needs, game mechanics, visual style, or deployment constraints are known. Second, it explicitly connects the construction of the prompt with evaluation through the same three dimensions. The rubric therefore does not merely score finished games, but helps identify which part of the prompt should be refined. Third, the framework extends beyond code generation by including GDD creation, graphics specification, optional image generation and quality control, and final integration.

The evaluation also clarifies the role of LLMs in this process. The results do not support a simple one-shot view of educational game generation. Several prototypes required multiple iterations before reaching the acceptance threshold, and the most frequent weak points concerned ease of use, game mechanic quality, readability/accessibility, clarity of goals and rules, and learning–activity alignment. These weaknesses are precisely the aspects that teachers and instructional designers must be able to inspect. At the same time, all attempts were runnable and stable, indicating that the main challenge was less the generation of PyGame code and more the transformation of code into a coherent educational game. In this sense, EduGameGen advances the idea of prompt engineering as a digital competence (Korzyski et al., 2023; Sahoo et al., 2024; White et al., 2023): the teacher is not expected to program the game manually, but must still articulate educational intent, judge the output, and iteratively improve the specification.

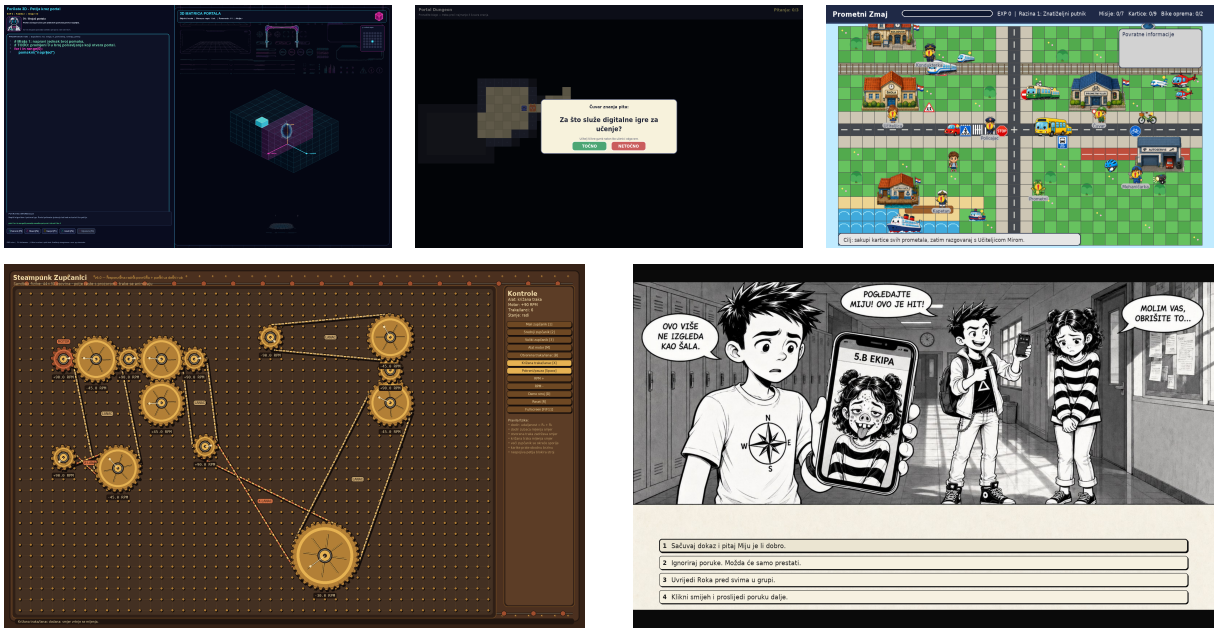


Figure 1. Screenshots of the final generated prototypes: Python 3D puzzle, portals rogue-like, traffic RPG, cogwheels sandbox, and cyberbullying visual novel.

The browser-deployment requirement further strengthens the practical relevance of the framework. By targeting PyGame prototypes that can be packaged with PygBag, EduGameGen supports games that run as static client-side applications without a custom backend. This is particularly suitable for schools, where installation rights, server administration, and platform heterogeneity can become barriers to classroom use. A generated game can therefore be distributed through a school website or learning platform, making it easier for teachers to share tailor-made prototypes with pupils and colleagues.

Finally, the results should be interpreted as evidence of feasibility and design plausibility, not as evidence of learning effectiveness. The evaluation was artificial and researcher-led, and the prototypes were not tested with teachers or learners. Nevertheless, this is consistent with an early DSR cycle (Hevner et al., 2004; Peffers et al., 2007): the study demonstrates an artifact, applies it across diverse cases, and identifies concrete refinement needs. The next step is naturalistic evaluation in classrooms, where the framework should be studied not only in terms of generated game quality, but also in terms of teacher workload, ease of prompt authoring, learner engagement, and learning outcomes.

7 Conclusion and Future Research

This paper presented EduGameGen, a design science artifact for prompt-driven educational game prototyping with ChatGPT and PyGame. The framework responds to the need for lightweight, teacher-accessible ways of creating custom educational games that are tailored to a specific subject, learner group, classroom situation, and learning outcome. Rather than treating prompt writing as an ad hoc activity, EduGameGen structures the process through an RCCF-based prompt format, an optional and priority-ordered context model, a three-phase generation workflow, and a rubric for formative evaluation.

The main contribution is the connection between prompt construction, artifact generation, and rubric-based refinement. The context model decomposes the prompt into educational, technical/usability, and game-design aspects, while allowing any field to remain unspecified when the teacher does not have the time or expertise to define it. The three-phase workflow extends the framework beyond code generation by including GDD creation, optional graphics generation and quality control, and final integration. By targeting PyGame prototypes that can be deployed through PygBag, the framework also supports browser-playable games that can be distributed as static web applications, which is especially relevant for school websites and learning platforms.

The artificial evaluation across 24 prompt-refinement attempts and five game types provided encouraging initial evidence. The final prototypes achieved a mean score of 32.60 out of 36, and all satisfied the acceptance criterion of having no weak rubric score. The results indicate that the framework can support diverse educational game genres, including visual novels, rogue-likes, RPGs, programming puzzles, and sandbox simulations. At the same time, the evaluation showed that useful prototypes usually emerge through iteration. The most frequent weaknesses concerned usability, game mechanics, readability/accessibility, clarity of goals and rules, and

learning–activity alignment, while technical execution and stability were less problematic.

The study has several limitations. The evaluation was researcher-led and artificial, the number of game types was limited, and the results do not measure actual learning outcomes. Furthermore, the framework was evaluated only with PyGame-based prototypes and within a small set of Croatian educational scenarios. The findings should therefore be interpreted as evidence of feasibility and design plausibility, not as evidence that the generated games improve learning.

Future research will focus on naturalistic classroom evaluation. Teachers should use EduGameGen to design games for their own lessons, after which both teacher and student feedback can be collected to refine the framework. Such studies should examine ease of prompt authoring, teacher workload, classroom usability, student engagement, accessibility, and perceived educational value. In later stages, the games should also be evaluated in terms of learning outcomes and compared with alternative teaching activities. Additional work should investigate inter-rater reliability of the rubric, refine the context fields and priority rules, and explore how the graphics-generation and PygBag deployment phases can be made more robust for everyday school use.

A EduGameGen Prompt Template

The following compact template, provided in Listing 1, gives the main EduGameGen structure for reuse. In normal use, only the **Context** block should be edited; all its fields are optional and may be left empty.

Listing 1. Copyable EduGameGen prompt template.

```
# MULTI-PHASE PROMPT FOR CREATING AN EDUCATIONAL PYGAME GAME

## MAIN GOAL
Create a small, complete, and playable educational game prototype in Python/PyGame
        ↪ .

The work is performed in up to three phases:
1. Phase 1 - Game concept, GDD, code, and graphics specification
2. Phase 2 - Graphics generation and quality control
3. Phase 3 - Integration, final checking, and delivery of files

Do not tell the user which phase you are in unless explicitly asked.

# MULTI-PHASE WORKFLOW RULES

## Workflow decision
If the user asks for generated graphics, then:
1. Perform Phase 1 only.
2. Do not generate images in Phase 1.
3. Use only programmatically created placeholder graphics in PyGame during Phase
        ↪ 1.
4. Create a file named generate_graphics.md containing the graphics specification
        ↪ and image-generation prompts.
5. After finishing the GDD, code, and graphics specification, inform the user that
        ↪  the program code is ready.
6. Ask whether the user wants to continue with graphics generation.
7. Do not continue to Phase 2 until the user explicitly confirms.
8. Do not offer the .py and .md files for download during Phase 1.

If the user does not ask for generated graphics, then:
1. Skip Phase 2.
2. After Phase 1, proceed directly to Phase 3.
3. In Phase 3, deliver the final files and summary.

Treat graphics generation as requested if the user mentions generated visual
        ↪ elements, sprites, scenes, illustrations, comic panels, backgrounds,
        ↪ character images, graphics, or other external visual resources.

# GENERAL RULES

* Do not generate images during Phase 1 or Phase 3.
* If visual elements are needed during game development, use PyGame placeholder
        ↪ graphics.
* If generated graphics are used, generate them only in Phase 2.
* Respect any visual, narrative, or genre style specified by the user.
* If educational, technical, and aesthetic requirements conflict, prioritize the
        ↪ learning outcomes.
* The game must be complete, understandable, and playable.
* Do not leave obvious TODO notes, unfinished systems, or unimplemented GDD
        ↪ elements.
* After generating the code, check whether everything specified in the GDD has
        ↪ been implemented. If not, fix the code before answering.
* The code must be in a single Python file unless the user explicitly requests
        ↪ otherwise.
* Avoid external dependencies other than PyGame.

# PHASE 1

## ROLE
Act as an expert educational game designer, instructional designer, and Python/
        ↪ PyGame developer. If the user specifies a particular visual, narrative
        ↪ , or genre style, also act as an expert in that style. Your goal is to
        ↪  create a small playable educational game prototype that aligns game
        ↪ mechanics with learning outcomes.

## CONTEXT
# Only edit this Context block. All fields are optional. Leave unknown or
        ↪ irrelevant fields empty.

### Educational context
* Subject/domain:
* Topic:
* Learning outcome(s):
* Cognitive level / learner age:
* Learning materials or content to include, if any:
* Learner profile:
* Prior knowledge:
* Learner diversity or accessibility considerations:
* Common misconceptions:
* Required educational feedback:
* Assessment or progress logic:

### Technical context
* Code structure: one file
* External dependencies: preferably none except PyGame
* External resources / generated assets:
* GUI or screen requirements:
* Interaction method:
* Runtime environment:
* Browser deployment: the game should be runnable locally and in a web browser
        ↪ through PygBag, without requiring a backend.
* Accessibility requirements:
* Testing or running instructions:

### Game-design context
* Genre:
* Core mechanic:
* Link between mechanic and learning outcome(s):
* Goal:
* Rules:
* Challenge:
* Progression:
* Feedback loop:
* Reward or scoring system:
* Win/loss condition:
* Aesthetics/theme:
* Session length:
* Narrative:

### Constraints and priorities
* Must-have requirements:
* Nice-to-have requirements:
* Avoid:
* If there is a conflict, prioritize:

## COMMAND
Before generating code, develop a short and realistic Game Design Document (GDD)
        ↪ and save it as gdd.md.

Create a complete, executable PyGame prototype of the educational game described
        ↪ in gdd.md. The game should be simple enough to understand quickly, but
        ↪  complete enough to be playable. Include educational content, player
        ↪ interaction, feedback, scoring or progress tracking, and a clear end
        ↪ condition.

The game should be runnable locally and in a web browser through PygBag. For local
        ↪  execution, use fullscreen mode unless otherwise specified. For
        ↪ browser execution, set an appropriate fixed resolution.

After generating the code, check whether all GDD elements have been implemented.
        ↪ If anything is missing, revise the code before the final response.
        ↪ Repeat this check until the implementation matches the specification.

If graphics generation was requested, do not offer generated .py or .md files for
        ↪ download in Phase 1. Offer downloads only in Phase 3 after graphics
        ↪ generation and integration.

# PHASE 2

If graphics generation was requested in Phase 1, open generate_graphics.md and
        ↪ generate all graphics according to the specified prompts. If the
        ↪ number of required images exceeds the system limit, generate the
        ↪ maximum possible number, then ask the user whether to continue. If
        ↪ generation continues in several rounds, use the first generated set as
        ↪  a reference to preserve consistency of characters, style, proportions
        ↪ , and important objects.

After generating the graphics, inspect each image and perform quality control
        ↪ against generate_graphics.md and gdd.md. Check whether each image can
        ↪ be integrated into the game. If any image does not match the prompt,
        ↪ has the wrong resolution, or is unsuitable for integration, adjust the
        ↪  prompt and regenerate the image.

After graphics generation is complete, inform the user and ask whether to start
        ↪ the final integration. Continue to Phase 3 only after the user
        ↪ confirms.

# PHASE 3

## FORMAT
Based on the previous phases, return the final answer in this structure:
```

```
1. Brief game concept
2. Short explanation of how the game mechanic supports the learning outcome(s)
3. Complete Python/PyGame code in a single file; provide a download link instead
        ↪ of printing the full code. If generated graphics were requested,
        ↪ integrate the graphics into the code by converting all images to
        ↪ base64 and embedding them directly in the .py file. If other necessary
        ↪  files exist, convert and embed them as well. Before integration,
        ↪ inspect all graphics and adjust the code to ensure appropriate size,
        ↪ resolution, cropping, and visual quality.
4. Controls and running instructions
5. Short checklist confirming whether the educational, technical, and game-design
        ↪ requirements were satisfied
```

Acknowledgement

This research has been financed by NextGenEU project AI-ARENA Framework for Ethical use of AI Agents in Multidisciplinary Research of the Digital Society (Okvir za etiku upotrebu AI agenata u multidisciplinarnim istraivanjima digitalnog drutva).

The authors acknowledge that ChatGPT 5.5 by OpenAI has been used to assist in editing and streamlining sections of this paper.

References

Borisova, M., Hadzhikoleva, S., & Hadzhikolev, E. (2024). ChatGPT as a tool for creating educational games for school education in computer technology. *Proceedings of the International Conference on Virtual Learning*, *19*, 323–333. https://doi.org/10.58503/icvl-v19y202427

Clark, D. B., Tanner-Smith, E. E., & Killingsworth, S. S. (2016). Digital games, design, and learning: A systematic review and meta-analysis. *Review of Educational Research*, *86*(1), 79–122. https://doi.org/10.3102/0034654315582065

Dicheva, D., Dichev, C., Agre, G., & Angelova, G. (2015). Gamification in education: A systematic mapping study. *Educational Technology & Society*, *18*(3), 75–88.

Gallotta, R., Todd, G., Zammit, M., Earle, S., Liapis, A., Togelius, J., & Yannakakis, G. N. (2024). Large language models and games: A survey and roadmap. *arXiv preprint arXiv:2402.18659*. https://doi.org/10.48550/arXiv.2402.18659

Hadzhikoleva, S., Gorgorova, M., Hadzhikolev, E., & Pashev, G. (2024). AI-driven approach to educational game creation. *Web Proceedings of the 16th International Conference ICT Innovations*, 112–123. https://proceedings.ictinnovations.org/attachment/paper/601/ai-driven-approach-to-educational-game-creation.pdf

Hamari, J., Koivisto, J., & Sarsa, H. (2014). Does gamification work? a literature review of empirical studies on gamification. *Proceedings of the 47th Hawaii International Conference on System Sciences*, 3025–3034. https://doi.org/10.1109/HICSS.2014.377

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, *28*(1), 75–105. https://doi.org/10.2307/25148625

Huber, S. E., Kiili, K., Nebel, S., Ryan, R. M., Sailer, M., & Ninaus, M. (2024). Leveraging the potential of large language models in education through playful and game-based learning. *Educational Psychology Review*, *36*(1), 25. https://doi.org/10.1007/s10648-024-09868-z

Isaza-Giraldo, A., Bala, P., Campos, P. F., & Pereira, L. (2024). Prompt-gaming: A pilot study on LLM-evaluating agent in a meaningful energy game. *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3613905.3650774

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Korzyski, P., Mazurek, G., Krzypkowska, P., & Kurasinski, A. (2023). Artificial intelligence prompt engineering as a new digital competence: Analysis of generative AI technologies such as ChatGPT. *Entrepreneurial Business and Economics Review*, *11*(3), 25–37. https://doi.org/10.15678/EBER.2023.110302

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, *24*(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

Pygame Community. (2026). *Pygame: Python game development*. Retrieved July 6, 2026, from https://www.pygame.org/

pygame-web contributors. (2026). *Pygbag: Pygame wasm*. Retrieved July 6, 2026, from https://pygame-web.github.io/wiki/pygbag/

Saari, S., & Schatten, M. (2024). Integrating cognitive agents into visual novels. *Proceedings of the 35th Central European Conference on Information and Intelligent Systems*, 27–32.

Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., & Chadha, A. (2024). A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*. https://doi.org/10.48550/arXiv.2402.07927

Schatten, M., & Schatten, M. (2020). A comparative study of gamification in programming education in a croatian high school. *2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO)*, 700–704. https://doi.org/10.23919/MIPRO48935.2020.9245237

Schatten, M. (2024, February). *O_hai 4 games - d4.4. case study 4 - hologame* (tech. rep. No. AIL202401) (Technical Report No. AIL202401 – First release, February 2024, edit April 2024. This work has been supported in full by the Croatian Science Foundation under the project number IP-2019-04-5824.). Laboratorij za umjetnu inteligenciju, Fakultet organizacije i informatike, Sveuilite u Zagrebu. Varadin, Croatia. https://urn.nsk.hr/urn:nbn:hr:211:989793

Schatten, M., Okrea uri, B., & Peharda, T. (2021). An agent-based game engine layer for interactive fiction. *Highlights in Practical Applications of Agents, Multi-Agent Systems, and Social Good. The PAAMS Collection*, 385–389. https://doi.org/10.1007/978-3-030-85739-4_38

Schatten, M., Peharda, T., et al. (2022). Towards a streamed holographic 3d game engine. *Central European Conference on Information and Intelligent Systems*, 17–22.

Schatten, M., & Schatten, M. (2019). Gamification of game programming education: A case study in a croatian high school. *Central European Conference on Information and Intelligent Systems*, 13–18.

Schatten, M., Schatten, M., & Inkret-Martinevi, L. (2023). Development of gamified cognitive agents for primary schools – a case study. *Proceedings of the 34th Central European Conference on Information and Intelligent Systems*, 3–8.

Spasi, A. J., Mladenovi, M., Nikoli, B., & Stoi, L. (2025). ChatGPT-assisted educational game development: A prompt engineering technique for teachers without coding experience. *Proceedings of the 1st International Scientific Conference Education and Artificial Intelligence (EDAI 2024)*, 153–164. https://doi.org/10.46793/EDAI24.153S

Sweetser, P. (2024). Large language models and video games: A preliminary scoping review. *arXiv preprint arXiv:2403.02613*. https://doi.org/10.48550/arXiv.2403.02613

Tomii, I., & Schatten, M. (2020). A conceptual network analysis of gamification practices in primary and secondary education. *2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO)*, 1091–1097. https://doi.org/10.23919/MIPRO48935.2020.9245225

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with ChatGPT. *arXiv preprint arXiv:2302.11382*. https://doi.org/10.48550/arXiv.2302.11382

Cyber-Physical and Embedded Systems

Maroš Valášek, Tomáš Ranuša, Gabriel Gašpar, Roman Budjač
Performance Evaluation of Python Web Frameworks for Resource-Constrained Embedded Systems

Gabriela Krizanova, Michal Kopcek, Roman Pikna
Behavioural model of a bulk material hopper for virtual commissioning

Zvonimir Šimić, Danijel Topić, Marina Dubravac, Lucija Gujić
Flexibility in the low voltage PV-rich distribution network with demand response and BESS

Tadej Šeruga, Igor Bernik
Cyber Security and Resilience of Artificial Intelligence Systems in the Nuclear Sector: A Literature Review

Kolok Peter
A Pulse-Based Semiconductor Reverse Breakdown Voltage Meter Using Inductive Flyback Energy Transfer

Marek Tebeľák, Lukáš Čechovič, Ondrej Karpiš, Ján Kapitulík
Predictive Control Algorithm for Pneumatic Positioning Cyber-Physical Systems Using Standard ON/OFF Valves

Thi Kim Anh Vo
Memory-Optimized Persistent Sheaf Laplacians for Real-Time Biomedical Signal Anomaly Detection on Resource-Scarce Platforms

Andrej Pavešić, Renata Mekovec
Mobile application development: securing user data

Performance Evaluation of Python Web Frameworks for Resource-Constrained Embedded Systems

Maroš Valášek, Tomáš Ranuša
University of Zilina
Faculty of Electrical Engineering
and Information Technology
Univerzitna 8215/1, 010 26 Zilina, Slovakia
{valasek, ranusa}@stud.uniza.sk

Gabriel Gašpar, Roman Budjač
University of Zilina
Research centre
Univerzitna 8215/1, 010 26 Zilina, Slovakia
{gabriel.gaspar, roman.budjac}@uniza.sk

Abstract. *The aim of this work is to analyze the performance characteristics of two popular Python web frameworks, FastAPI and Django. These frameworks are deployed on a Raspberry Pi 5 single-board computer. Testing is conducted using the wrk tool, a custom Python concurrent benchmarking script, and an asynchronous WebSocket client based on the websockets library. The study examines the impact of different WSGI/ASGI web servers on throughput, latency stability, and system resource consumption. The results show that FastAPI is a faster and less system-intensive choice compared to Django, although the violin plots indicate that while FastAPI has a lower average latency, Django exhibits a lower maximum latency.*

Keywords. Python, framework, embedded systems, resource-constrained

1 Introduction

With the growing number of IoT devices and embedded *systems*, there is a need for efficient data processing directly at the edge of the network. These devices have limited computing power, making it critical to choose a software architecture that maximizes throughput with minimal resource consumption. This necessitates the careful selection of lightweight yet robust web *frameworks* capable of handling high request volumes within such constrained environments. This paper evaluates the performance of Python web *frameworks*, specifically FastAPI(FastAPI, n.d.) and Django(Django Software Foundation, n.d.), for their suitability in such resource-constrained settings, focusing on metrics crucial for real-time edge processing (Dey et al., 2025). Specifically, the evaluation encompasses key performance indicators such as request throughput, response times, and resource utilization (e.g., CPU, memory) for basic operations (Bednarz and Miłosz, 2025).

The aim of this work is to analyze the performance characteristics of two popular Python web *frameworks*: FastAPI and Django. While both are Python-based and widely utilized for backend development (KP and N, 2023), they differ significantly in their architectural paradigms and performance profiles (Choma et al., 2023), which are critical considerations for IoT deployments.

FastAPI is a modern, high-performance web framework for building Application Programming Interfaces (APIs) with Python, which natively supports asynchronous programming via asyncio(Python Software Foundation, n.d.), Python's built-in library for asynchronous programming that enables concurrent handling of I/O operations through an event loop without blocking threads and the WebSocket protocol. FastAPI, known for its high performance and asynchronous capabilities, offers a compelling solution for microservices and real-time data processing, while Django, a more mature and feature-rich framework, provides a comprehensive suite for complex web applications (Poliakov and Yeremenko, 2024).

Django, on the other hand, is a high-level Python web framework following the Model-View-Template architectural pattern. Historically designed for a synchronous processing model via Web Server Gateway Interface (WSGI), Django was integrated with the Django Channels package to extend it with the Asynchronous Server Gateway Interface (ASGI) protocol and support WebSockets. This integration allows Django to handle long-lived connections and asynchronous tasks, positioning it as a contender for applications requiring real-time functionalities, despite its traditional synchronous core (Ruchi et al., 2026). While FastAPI is recognized for its API-first design and asynchronous capabilities, providing low-latency communication (Mabotha et al., 2025), Django's extensive features and scalability often come with increased CPU utilization compared to more lightweight *frameworks* (Ruchi et al., 2026, Tusa and Clayman, 2021).

While both *frameworks* facilitate web application

development, their fundamental architectural differences, particularly regarding synchronous versus asynchronous operations, profoundly affect their performance in resource-constrained environments (Bednarz and Miłosz, 2025). This study examines how these architectural paradigms translate into measurable performance indicators, including throughput, latency, and resource utilization, on a Raspberry Pi 5, a representative platform for edge computing applications (Choma et al., 2023, Ruchi et al., 2026). This comparative analysis seeks to provide empirical data to guide developers in selecting the optimal Python web framework for IoT and embedded *systems*, where computational efficiency and responsiveness are paramount.

The insights from this analysis help determine which framework is better suited for resource-limited edge deployments, particularly those requiring efficient data handling and real-time responsiveness.

2 Related works

Several studies have evaluated the performance of Python web *frameworks*. Some research focuses on comparing *frameworks* in standard environments, while other studies compare databases running on these *frameworks*, noting that much larger performance differences often arise specifically due to the database layer. Our study isolates the performance of the *frameworks* themselves for IoT-specific tasks directly on ARM architecture typical for single-board computers. For instance, previous works have explored general web application development with Python, specifically highlighting *frameworks* like Django and Flask for their simplicity and readability in web server applications (KP and N, 2023). However, these studies often overlook the unique constraints and performance implications of deploying such *frameworks* on resource-limited embedded *systems*, which is a primary focus of the current research. Furthermore, while many evaluations focus on general server-side performance, our investigation delves into the specific performance characteristics pertinent to IoT contexts, such as real-time responsiveness and efficiency under constrained computational resources. For example, (Morabito et al., 2017) highlights how management of HTTP requests can significantly impact gateway performance, indicating the necessity of evaluating *frameworks* within the context of their communication protocols. Other research explores the use of Flask for microservices on edge devices, finding it suitable for object detection functionalities due to its ease of integration and testing across diverse hardware (Alqahtani et al., 2024). However, these studies often concentrate on high-level application functionality rather than low-level performance metrics relevant to continuous, resource-intensive IoT operations. A study comparing Flask and web2py for AI microservices, for instance, found Flask to have lower latency and higher throughput, underscoring the importance of empirical evaluations for framework selection in specialized contexts (Idowu and Omiyale, 2026). Conversely, while some studies discuss the general use of Python in embedded *systems* for programmability and code mobility (Morabito et al., 2017), they do not typically provide comprehensive benchmarking of web *frameworks* tailored for specific real-time performance metrics like inference latency, CPU utilization, and memory footprint in TinyML applications (Dey et al., 2025). Django and FastAPI were selected as representatives of two opposing architectural paradigms, a mature synchronous framework and a modern, natively asynchronous one, enabling a direct comparison of synchronous versus asynchronous processing. Although Flask is a widely used lightweight framework, it predates FastAPI and lacks native asynchronous and WebSocket support, making FastAPI a more suitable candidate for our real-time, sensor-data streaming use case. While Flask and web2py have been examined in prior studies, our focus was on the two most widely adopted frameworks offering native WebSocket support. This paper extends upon this body of work by conducting a rigorous comparative analysis of FastAPI and Django specifically on a Raspberry Pi 5, a ubiquitous platform for IoT and edge computing, to provide empirical data on their performance characteristics under scenarios typical of embedded *systems*.

3 Methodology and setup

This section outlines the experimental design, hardware and software configurations, and the metrics employed for evaluating the performance of FastAPI and Django. Our methodology encompasses controlled experimental conditions to ensure reproducibility and statistical validity, focusing on isolating the performance impacts of each framework on critical *system* resources. The experimental setup includes a Raspberry Pi 5 acting as the host for the web *frameworks*, meticulously configured to monitor CPU, memory, and network I/O during various load conditions. This approach facilitates a direct comparison of how each framework handles varying request loads and data processing demands, thereby offering insights into their suitability for edge computing environments.

3.1 Hardware specification

A Raspberry Pi 5 was used, which is a single-board computer based on a 64-bit ARM Cortex-A76 Broadcom BCM2712 processor with a frequency of 2.4 GHz. This unit is equipped with 8 GB of LPDDR4X SDRAM, providing ample memory for complex computational tasks and multitasking, typical of edge AI applications. This configuration enables robust processing capabilities suitable for demanding machine

learning inferences at the edge (Emanuilov and Dimov, 2025).

Both the server and client Raspberry Pi 5 units ran Raspberry Pi OS (Raspberry Pi Ltd, n.d.) (64-bit, Debian-based), while the development workstation used for some tests ran Windows 11 Pro (Microsoft, n.d.). The detailed technical specifications of the hardware used in this study are summarized in Table 1 (Raspberry Pi Ltd, n.d.). The Raspberry Pi 5's PCIe interface is accessible through an FPC (Flexible Printed Circuit) connector, allowing for external expansion.

Table 1. Technical Specifications of Raspberry Pi 5 (8 GB version)

Parameter	Specification
Model	Raspberry Pi 5 (8 GB RAM variant)
SoC	Broadcom BCM2712
CPU	Quad-core ARM Cortex-A76 @ 2.4 GHz
GPU	VideoCore VII (OpenGL ES 3.1, Vulkan 1.2)
Memory	8 GB LPDDR4X SDRAM (up to 34 GB/s)
Storage	microSD (UHS-I), PCIe 2.0 external support
PCIe	1 × PCIe 2.0 lane (FPC connector)
USB	2 × USB 3.0, 2 × USB 2.0
Networking	Gigabit Ethernet (PoE+ support via HAT)
Wireless	Wi-Fi 802.11ac, Bluetooth 5.0 / BLE
Display	2× micro-HDMI (dual 4Kp60)
Camera	2× MIPI CSI (4-lane)
DSI	2× MIPI DSI (4-lane)
GPIO	40-pin header (backward compatible)
Power	5 V USB-C (recommended 5 V/5 A)
Dimensions	85 mm × 56 mm

3.2 Software specification

The work examines the impact of different WSGI/ASGI web servers (Gunicorn, Daphne, Uvicorn). Specifically, we deployed FastAPI applications using Uvicorn, a high-speed ASGI web server, and Django applications using Gunicorn with Uvicorn workers to ensure optimal asynchronous handling and performance comparability between the *frameworks* (Durán et al., 2024).

For FastAPI, Uvicorn was used as the web server, which is a lightweight ASGI server based on the uvloop and httptools *libraries*. This choice leverages Uvicorn's asynchronous capabilities to efficiently handle concurrent connections, crucial for high-throughput edge deployments.

For the Django framework, three web servers were initially tested: Gunicorn (Gunicorn, n.d.), Daphne (Django Software Foundation, n.d.), and Uvicorn (Uvicorn, n.d.). Gunicorn is a traditional WSGI server that delivered the highest HTTP throughput; however, it lacks native WebSocket support, rendering it unsuitable for applications that necessitate real-time communication. Daphne and Uvicorn are both ASGI servers and were assessed for their compatibility and performance with Django's asynchronous capabilities.

Daphne proved to be much slower and did not support multiple workers, so it was excluded from the test. Consequently, Uvicorn was selected as the deployment server for Django, running directly with multiple workers through the ASGI interface provided by Django Channels. This configuration was chosen as the optimal deployment setup for Django to facilitate a fair performance comparison with FastAPI, which is also served by Uvicorn. This configuration ensures that both *frameworks* operate under similar asynchronous serving mechanisms, thereby minimizing server-side performance discrepancies attributable to the underlying WSGI/ASGI implementation rather than the framework logic itself.

System statistics were collected using the psutil (psutil developers, n.d.) *library* in a separate thread, preventing blocking of HTTP requests. The *system* resource utilization metrics, including CPU load, memory consumption, and network throughput, were logged at regular intervals to provide a granular view of performance under various synthetic load conditions. Initial measurements showed a latency of 120 ms despite a 7 ms ping because the psutil function was blocking further requests, this was resolved by moving psutil to a secondary thread. This concurrent monitoring strategy allowed for precise, non-intrusive data collection, enabling an accurate assessment of framework performance under realistic load conditions.

The software stack consisted of Python 3.13.5, FastAPI 0.137.2, Uvicorn 0.49, Django 6.0.6, Django Channels 4.3.2, psutil 7.2.2, websockets 16.0, seaborne 0.13.2, matplotlib 3.10.8 pandas 2.3.3 and numpy 2.3.5. All communication was conducted over the HTTP/1.1 protocol, the default for both Uvicorn and the wrk benchmarking tool (Wrk developers, n.d.).

3.3 Test scenarios

The server was implemented in the main.py file and consisted of four HTTP endpoints and a WebSocket handler:

- GET ping: The simplest possible endpoint returning only "message": "pong" with no processing or *system* info. This endpoint serves as a baseline for measuring raw network and framework overhead, isolating the effects of data serialization and deserialization.
- GET hello: Similar to ping but additionally returns *system* information from the Pi (CPU, RAM, temperature, timestamp). It simulates a real endpoint returning device status. The only difference from the ping endpoint is the additional overhead of retrieving and serializing cached system information, which is continuously updated by a background thread.
- POST echo: Receives a JSON payload with a random name, number, and a 64-character string and returns it back. It simulates sending sensor data to the server, typical for research center projects. This endpoint specifically tests the framework's efficiency

in handling data ingress, parsing, and egress under varying payload sizes, directly relevant to IoT data acquisition pipelines.

- GET heavy: The server calculates the sum of the squares of numbers from 0 to 10000 before responding. It simulates an endpoint that must process data, such as sensor data aggregation. This computational task evaluates the *frameworks'* capacity for CPU-intensive operations, reflecting scenarios common in edge analytics or localized machine learning inference.
- WS ws: A persistent two-way connection in which the client can continuously stream sensor data to the server independently of individual server requests. The server immediately returns each received message as an echo response. This simulates a live sensor-data stream where new data are pushed by the client as they become available, rather than being repeatedly requested by the server. This WebSocket interaction assesses the *frameworks'* real-time communication capabilities and their efficiency in maintaining continuous, low-latency data streams, which is critical for interactive edge applications.

The detailed performance comparison across all tested web framework configurations, including GET, POST, and WebSocket throughput alongside latency percentiles, is presented in Table 3, where req/s denotes requests per second, WS represents WebSocket throughput, and p50 and p99 correspond to the median and 99th percentile latency, respectively. The best-performing results in the table are highlighted in bold. The data was collected using a custom Python concurrent benchmark script running 12 parallel workers over a duration of 10 minutes per configuration.

The results of Django + Daphne are included in Tables 3 and **??** solely to demonstrate its performance inadequacy during the initial evaluation phase and to justify its exclusion from the primary benchmark, rather than as a viable deployment configuration.

Table 2. Performance Comparison of Web Framework Configurations

Configuration	GET req/s	POST req/s	WS req/s	p50 [ms]	p99 [ms]
Django + Daphne	82	82	52	37	88
Django + Uvicorn	165	82	**55**	23	50
FastAPI + Uvicorn	**187**	**93**	54	**19**	**45**

3.4 Benchmarking tools

To evaluate the performance characteristics of each framework under heterogeneous workloads, a combination of industry-standard benchmarking tools and custom-developed test scripts was employed to ensure consistent and objective data collection. An initial benchmark consisted of a 10-minute stress test executed via a Python script on a Windows-based *system*, utilizing the concurrent futures module to generate parallel tasks. This approach was designed to simulate high-concurrency conditions and to assess sustained *system* performance under load. However, experimental configurations with 50, 100, 300, and 500 worker threads were constrained by operating *system* throughput, with a maximum observed rate of approximately 190 requests per second. To address these constraints, subsequent experiments were conducted using Apache JMeter 5.5, deployed within a dedicated benchmarking container. This setup enabled the generation of controlled, scalable workloads, facilitating the evaluation of server response times under varying network conditions while minimizing client-side bottlenecks.

To further mitigate the limitations of the Windows environment, an auxiliary Raspberry Pi device was used to run the wrk utility. This configuration allowed for a more accurate assessment of server performance under unconstrained load conditions. The wrk tool, implemented in C, can generate high request rates, reaching tens of thousands of requests per second. Additionally, its extensibility via Lua scripting enables customized request generation, response handling, and reporting, thereby providing flexibility for a wide range of performance testing scenarios.

All benchmarking scripts, including the concurrent Python script, the asynchronous WebSocket client, and the visualization scripts, were developed by the authors. Each request was logged to a CSV file with its timestamp, test type, latency, and system metrics, and subsequently processed using the pandas and statistics libraries. The complete set of benchmark results, including raw CSV outputs and supplementary measurements, is available in the project GitHub repository DePolarBear, n.d.

4 Results and discussion

This section presents the empirical results and a comparative analysis of the performance metrics, focusing on latency, throughput, and resource utilization, derived from the rigorous benchmarking of FastAPI and Django under various simulated loads.

4.1 Throughput

Using the wrk tool, we achieved up to 1660 req/s with 12 workers compared to the 180 req/s limit on the Windows PC. This significant increase underscores the impact of client-side resource limitations on benchmark accuracy and validates the necessity of utilizing robust load generation tools and environments. The utilization of wrk, known for its efficiency in generating high-volume traffic due to its asynchronous I/O model and C-based implementation, allowed for a more realistic evaluation of server capabilities under sustained

throughput. Furthermore, this disparity in throughput highlights the critical importance of selecting appropriate benchmarking tools and environments to accurately assess the maximum transaction processing capacity of an application.

The detailed results of the ping and request throughput benchmark across all tested configurations are presented in Table 3, where all measurements were obtained using the wrk HTTP benchmarking tool with 4 threads, 100 concurrent connections, and a test duration of 30 seconds per endpoint. The best-performing results in the table are highlighted in bold.

Table 3. Ping and Request Throughput Benchmark

Configuration	ping req/s	hello req/s	heavy req/s	localhost ping
Django + Daphne	299	289	223	295
Django + Uvicorn	634	622	549	568
FastAPI + Uvicorn	**1401**	**1597**	**1133**	**1660**

FastAPI with Uvicorn reached 1,660 req/s on the localhost ping, while Django with Uvicorn reached 568 req/s. This substantial difference in throughput for a trivial endpoint suggests fundamental architectural distinctions in request handling and processing overhead between the two *frameworks*. Specifically, FastAPI's asynchronous nature and Uvicorn's efficient ASGI server likely contribute to its superior performance in I/O-bound operations, whereas Django's more synchronous foundation, even when paired with an ASGI server, may introduce additional overhead.

On the "hello" endpoint, FastAPI processed 1,597 req/s, whereas Django handled 622 req/s. This indicates that even with minimal additional processing for retrieving *system* information, FastAPI maintains a significant performance advantage, suggesting superior efficiency in managing concurrent requests and integrating *system*-level data.

During the WebSocket test with 200 clients and 20,000 messages, FastAPI achieved a throughput of 8,751.5 req/s, while Django reached 7,606.7 req/s. This demonstrates FastAPI's marginally superior efficiency in managing persistent, real-time communication streams, even though both *frameworks* exhibit robust performance under high WebSocket load. This finding suggests that while both *frameworks* are competent in handling synchronous and asynchronous tasks, FastAPI's design, often leveraging asynchronous programming models, provides a slight edge in scenarios requiring high concurrency and real-time data exchange.

The comprehensive WebSocket load test results across varying numbers of concurrent clients for both *frameworks* are presented in Table 4, where higher throughput values indicate better performance, lower latency values indicate better responsiveness, p50, p95, and p99 represent latency percentiles, max denotes the worst observed latency, and the best values per column are highlighted in bold. All measurements were obtained using a custom Python asyncio benchmark script executed on the second Raspberry Pi acting as the client, with each configuration tested using 10 messages per client.

4.2 Latency

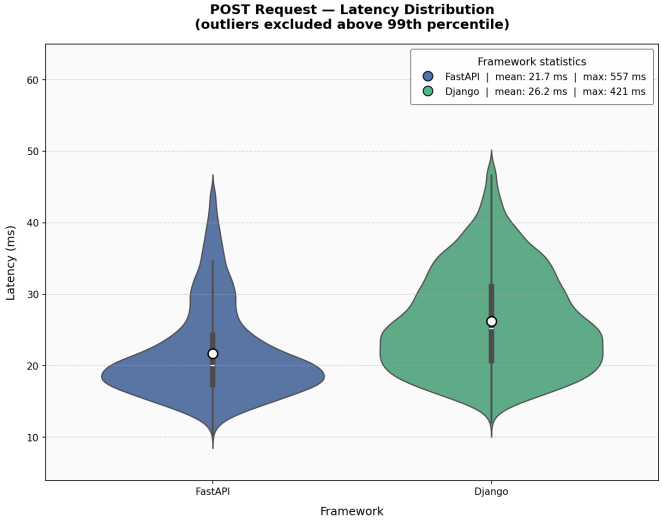
We visualized the latency distribution using violin plots. These plots offer a comprehensive view of response-time variability under different load conditions, thereby providing more insight into performance consistency than simple mean or median metrics (Hintze and Nelson, 1998).

All plots were generated using matplotlib, with violin plots additionally relying on seaborn. The 99th percentile latency represents the value below which 99% of all measured latencies fall, characterizing the worst-case response time while excluding extreme outliers.

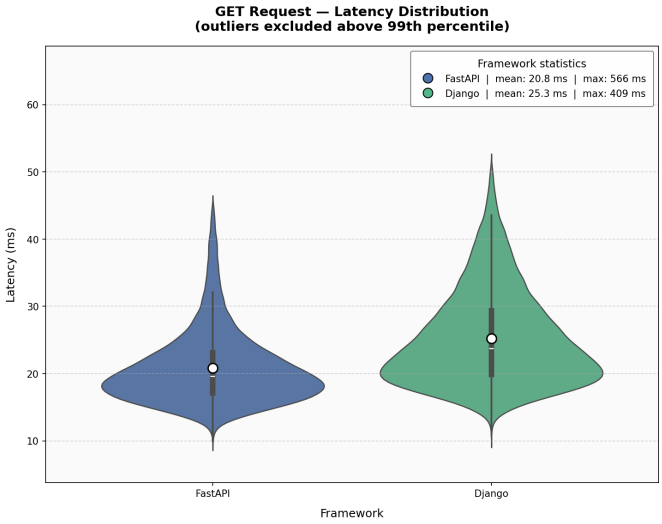
The plots show that FastAPI has a lower average latency, but Django has a lower maximum latency. The latency distributions for GET, POST, and WebSocket requests are visualized using violin plots in Fig. 1 a), b), c) respectively, with outliers above the 99th percentile excluded to better illustrate the core distribution of response times. In the overall benchmark, the average latency for FastAPI was 16.38 ms with a maximum of 565.79 ms. In contrast, the average latency for Django was 19.34 ms, but its maximum latency was lower at 420.99 ms. This discrepancy suggests that while FastAPI generally offers faster responses, Django exhibits greater stability under peak load conditions, potentially due to different internal queuing mechanisms or resource management strategies. These observed trade-offs between CPU and RAM utilization, alongside thermal performance, provide critical insights for deployment strategies, particularly for resource-constrained environments or applications requiring sustained high performance.

4.3 Resource utilization

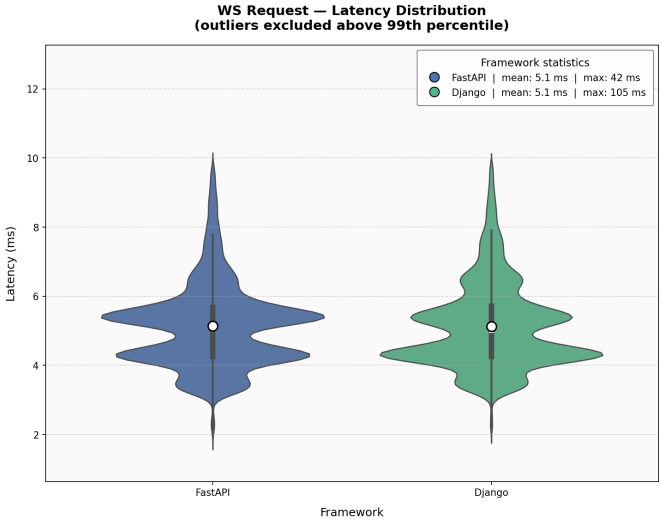
In the case of FastAPI, there is lower CPU utilization but higher RAM utilization compared to Django. These metrics were collected over a 10-minute benchmark session using a custom Python concurrent script with 12 parallel workers, continuously logging CPU utilization, RAM consumption, and CPU temperature at 1-second intervals via the psutil *library*, resulting in approximately 600 samples per configuration. No thermal throttling was observed during any of the test runs, as the CPU temperature remained below 48 °C across all configurations, well under the Raspberry Pi 5 throttling threshold of 85 °C. The average CPU utilization for Django with Uvicorn was 15.5% (max 29.6%), whereas FastAPI averaged 11.8% (max 18.9%). This indicates that FastAPI, despite its higher throughput,



(a)



(b)



(c)

Figure 1. Latency distribution of requests for FastAPI and Django (outliers above 99th percentile excluded) a) POST requests b) GET requests c) WS requests.

Table 4. Performance Comparison of FastAPI and Django Under Load

Framework	Clients	Requests	Throughput [req/s]	p50 [ms]	p95 [ms]	p99 [ms]	max [ms]
FastAPI	10	1000	**1255.5**	**5.59**	**9.11**	**10.96**	**13.95**
FastAPI	50	5000	**5414.7**	**7.46**	**10.23**	**11.82**	**35**
FastAPI	100	10000	**6777.7**	**10.03**	**15.25**	**19.97**	**115.45**
FastAPI	200	20000	**8751.5**	**16.35**	**26.27**	**31.92**	**171.12**
Django	10	1000	1051.0	7.38	10.80	13.58	16.21
Django	50	5000	4229.3	9.14	14.40	20.54	42.56
Django	100	10000	6164.6	13.00	20.13	26.64	88.70
Django	200	20000	7606.7	18.63	31.05	43.56	206.13

is more CPU-efficient, likely due to its asynchronous architecture and optimized I/O handling, which minimizes blocking operations.

Conversely, average RAM usage was 13.8% (1113.18 MB) for Django and 15.5% (1247.31 MB) for FastAPI. The average temperature was slightly lower for FastAPI (42.2 °C) than for Django (43.2 °C). This suggests that while FastAPI leverages memory more intensively, possibly for maintaining asynchronous task states or larger event loops, its optimized CPU usage contributes to overall cooler operation, indicating a more efficient thermal profile under load.

The overall *system* resource utilization comparison between FastAPI and Django, including average and peak CPU and RAM consumption recorded during the benchmark, is presented in Table 5, where lower values indicate more efficient resource usage and the best values per column are highlighted in bold.

Table 5. Benchmark Comparison of FastAPI and Django

Metric	FastAPI	Django
Avg CPU [%]	**11.82**	15.49
Max CPU [%]	**18.9**	29.6
Avg RAM [%]	15.47	**13.82**
Max RAM [%]	15.8	**14.7**
Avg RAM [MB]	1247.31	**1113.18**
Max RAM [MB]	1274.8	**1188.3**

5 Conclusions

After the tests, FastAPI proved to be faster and less *system*-intensive choice compared to Django. This holds true for both synchronous and asynchronous operations, aligning with findings that asynchronous architectures generally outperform synchronous ones in terms of lower tail latencies under high concurrency. We tested "plain" FastAPI against Django, which basically has everything implemented, which may be a reason for Django's higher latency. It should be noted that during testing, Django was deployed with its default configuration, meaning that built-in middleware components such as session management, authentication, Cross-Site Request Forgery (CSRF) protection, and the admin panel were not explicitly disabled. These components introduce additional processing overhead on every request, even when not actively utilized, and represent a limitation of this study. Therefore, the results should be interpreted in the context of a default Django deployment compared to a minimalist FastAPI setup, rather than as an absolute performance comparison between the two *frameworks*. For our purposes of controlling and receiving data from embedded *systems*, we chose FastAPI for its speed and modularity, which make it adaptable to any task. This decision is further supported by evidence highlighting the benefits of asynchronous architectures in high-transaction *systems*, where minimizing blocking I/O leads to superior tail latencies.

Acknowledgments

This research was supported by the project VEGA through the Research on motion data analysis methods for applications in diagnosis and therapy of gnostically relevant symptoms under Grant 1/0147/25.

References

Alqahtani, D. K., Cheema, A., & Toosi, A. N. (2024). Benchmarking deep learning models for object detection on edge computing devices. *arXiv (Cornell University)*, (242). https://doi.org/10.48550/arxiv.2409.16808

Bednarz, B., & Miłosz, M. (2025). Benchmarking the performance of python web frameworks. *Journal of Computer Sciences Institute*, *36*, 336–341. https://doi.org/10.35784/jcsi.7738

Choma, D., Chwaleba, K., & Dzieńkowski, M. (2023). The efficiency and reliability of backend technologies: Express, django, and spring boot. *Informatyka Automatyka Pomiary w Gospodarce i Ochronie*

Środowiska, *13*, 73–78. https://doi.org/10.35784/iapgos.4279

DePolarBear. (n.d.). Framework-Benchmark-Data [Accessed: 2026-06-30].

Dey, A., Srivastava, S. P., Singh, G., & Pettit, R. G. (2025). Real-time performance benchmarking of tinyml models in embedded systems (pico: Performance of inference, cpu, and operations). *arXiv (Cornell University)*, 279–284. https://doi.org/10.48550/arxiv.2509.04721

Django Software Foundation. (n.d.). Daphne [Accessed: 2026-06-29].

Durán, F., Martinez, M., Lago, P., & Martínez-Fernández, S. (2024). Insights into resource utilization of code small language models serving with runtime engines and execution providers. https://doi.org/10.48550/ARXIV.2412.15441

Emanuilov, S., & Dimov, A. (2025). A quantitative framework for evaluating architectural patterns in ml systems. https://doi.org/10.48550/ARXIV.2501.11543

FastAPI. (n.d.). FastAPI. Retrieved June 30, 2026, from https://fastapi.tiangolo.com/

Gunicorn. (n.d.). Gunicorn - Python WSGI HTTP Server for UNIX [Accessed: 2026-06-30].

Hintze, J. L., & Nelson, R. D. (1998). Violin plots: A box plot-density trace synergism. *The American Statistician*, *52*(2), 181–184. https://doi.org/10.1080/00031305.1998.10480559

Idowu, O. F., & Omiyale, A. D. (2026). Comparative evaluation of flask and web2py for ai microservices: An empirical benchmark on model-inference workloads. *ABUAD Journal of Engineering Research and Development (AJERD)*, *9*, 12–27. https://doi.org/10.53982/ajerd.2026.0901.02-j

KP, B., & N, D. (2023). Leveraging django: A web framework for efficient infrastructure monitoring tools. *International Journal for Research in Applied Science and Engineering Technology*, *11*, 407–412. https://doi.org/10.22214/ijraset.2023.55414

Mabotha, E., Mabunda, N., Ali, A., & Khan, B. (2025). Exploring dynamic restful api implementation in iot environments using docker. *Scientific Reports*, *15*(34267). https://doi.org/10.1038/s41598-025-16460-0

Microsoft. (n.d.). Download Windows 11 [Accessed: 2026-06-30].

Morabito, R., Farris, I., Iera, A., & Taleb, T. (2017). Evaluating performance of containerized iot services for clustered devices at the network edge. *IEEE Internet of Things Journal*, *4*, 1019–1030. https://doi.org/10.1109/jiot.2017.2714638

Poliakov, M., & Yeremenko, B. (2024). Infocommunication system of the electronic estimation of adolescents' special abilities. *Management of Development of Complex Systems*, 139–145. https://doi.org/10.32347/2412-9933.2024.58.139-145

psutil developers. (n.d.). psutil documentation [Accessed: 2026-06-30].

Python Software Foundation. (n.d.). asyncio — Asynchronous I/O [Accessed: 2026-06-30].

Raspberry Pi Ltd. (n.d.). Raspberry Pi 5 [Accessed: 2026-06-29].

Ruchi, R., Saroj, S., Joshi, H., & Thareja, D. Y. (2026). A comparative analysis of fastapi and django frameworks: Evaluating performance, scalability, and concurrency efficiency in modern web applications. *International journal of scientific research in engineering and management*, *10*, 1–9. https://doi.org/10.55041/ijsrem58536

Tusa, F., & Clayman, S. (2021). The impact of encoding and transport for massive real-time iot data on edge resource consumption. *Journal of Grid Computing*, *19*(32). https://doi.org/10.1007/s10723-021-09577-9

Uvicorn. (n.d.). Uvicorn [Accessed: 2026-06-30].

Wrk developers. (n.d.). wrk - a HTTP benchmarking tool [Accessed: 2026-06-30].

Behavioural model of a bulk material hopper for virtual commissioning

Michal Kopcek, **Gabriela Krizanova, Roman Pikna**

Slovak University of Technology in Bratislava

Faculty of Materials Science and Technology in Trnava

Institute of Applied Informatics, Automation and Mechatronics

Jána Bottu 25, 91724 Trnava, Slovakia

{michal.kopcek, gabriela.krizanova, roman.pikna}@stuba.sk

Abstract. *This paper presents the design, implementation, and experimental validation of a behavioural model of a bulk material process hopper for virtual commissioning at the machine level. The mathematical model is based on the geometry of a truncated pyramid-shaped hopper and Bernoulli's equation for mass flow rate calculation. It was implemented in the SIMIT environment and integrated with TIA Portal. Validation using a physical hopper model at SKARTEK s.r.o. achieved an average correlation of 0.977 between mass response curves. The model enables the development of control algorithms without physical prototypes, in line with Industry 4.0 principles.*

Keywords. behavioural model, bulk material hopper, virtual commissioning, SIMIT, TIA Portal

1 Introduction

Nowadays, the development of virtual models is becoming increasingly widespread in industry with the aim of improving the efficiency, flexibility, and quality of manufacturing systems. One of the key areas is virtual commissioning, which enables the simulation and testing of production equipment prior to its physical implementation, thereby significantly reducing commissioning time and minimizing the risk of errors in the real system (Tjahjono et al., 2017). Bulk material hoppers constitute an integral part of automated production processes in the construction, food processing, chemical, and agricultural industries.

A current trend in industry is the increase of production capacity while simultaneously reducing the time required for the design and commissioning of production lines. This leads to a higher risk of errors during the design, implementation, and debugging of systems, resulting in production delays and increased costs. Identifying and eliminating an error already during the initial design phase is more economically advantageous than resolving it during the physical implementation stage. For these reasons, simulation tools are being developed to eliminate as many deficiencies as possible before the physical implementation of the system (Kopček et al., 2022).

To address these challenges and systematically evaluate the proposed approach, this paper formulates two key research questions (RQs) that guide the theoretical derivation, implementation, and experimental validation:

RQ1: How accurately can a SIMIT-based behavioral model reproduce the transient discharge dynamics (mass flow rate and weight profile over time) of a truncated pyramid-shaped bulk material hopper under real-time PLC control?

RQ2: To what extent does native integration between the SIMIT simulation platform and TIA Portal facilitate machine-level virtual commissioning in terms of signal mapping, execution synchronization, and model reusability compared to non-native simulation environments?

The aim of the presented work was to design and implement a behavioural model of a truncated pyramid-shaped process hopper in the SIMIT environment and to verify its validity through experimental comparison with a physical model built at SKARTEK s.r.o. The model is intended to serve as a digital twin for the development of PLC control algorithms, thereby eliminating the need for repeated construction of physical prototypes for different hopper variants (Pikna, 2025).

2 Theoretical background

2.1 Industry 4.0 and digital twin

The term Industry 4.0 refers to the fourth industrial revolution enabled by the introduction of the Internet of Things (IoT) into manufacturing environments. The vision of Industry 4.0 emphasizes global networks of machines within a smart factory environment, capable of autonomous information exchange and control of

their own production operations (Lechler et al., 2019; Tjahjono et al., 2017). Since collaboration among suppliers, manufacturers, and customers is essential for the transparency of the entire value chain, it is necessary to analyze the impact of Industry 4.0 on the supply chain as a whole.

To establish the theoretical basis for answering **RQ1**, it is essential to examine how digital twin architectures support virtual commissioning (VC). A digital twin could be defined as a virtual representation of a physical object, process, or system that enables its simulation, analysis, and optimization in real time (Kopček et al., 2022; Schroeder et al., 2016). In industrial automation, VC utilizes this twin to validate programmable logic controller (PLC) code prior to the physical installation of equipment, shifting software debugging from the shop floor to the virtual sandbox (Reinhart & Wünsch, 2007). This model-in-the-loop (MiL) and software-in-the-loop (SiL) paradigm drastically reduces project implementation risks, prevents mechanical collisions, and shortens deployment cycles (Hoffmann et al., 2021).

The commissioning process of an automation system involves several consecutive phases: preparation and installation of hardware and software, functional testing, correction and tuning, personnel training, transition to real operation, and subsequent support and maintenance (Barth & Fay, 2013). Thanks to behavioral simulation tools, most of these phases can be carried out virtually before physical deployment, which significantly reduces project risks and costs (Pikna, 2025). In the context of bulk material handling, behavioral models must accurately capture continuous mass flow rates and discrete actuator dynamics to serve as reliable validation targets for control logic (Klüber & Stark, 2020).

2.2 Simulation platform - SIMIT

Addressing **RQ2** requires a comparative evaluation of available simulation platforms regarding their capability to interface with industrial control hardware. Currently, several software platforms are used for creating behavioural models of industrial equipment, including MATLAB/Simulink, FlexSim, Factory I/O, and SIMUL8. Table 1 summarizes the main characteristics, primary applications, advantages, and limitations of these platforms.

Table 1. Comparison of prominent industrial simulation platforms for virtual commissioning

Platform	Main focus	Key characteristic
MATLAB / Simulink	Mathematical and control modelling	High analytical accuracy and extensive modelling capabilities
FlexSim	Manufacturing and logistics simulation	Advanced 3D visualization and process optimization
Factory I/O	Automation education and training	Fast creation of virtual automation scenarios
SIMUL8	Discrete-event simulation	Statistical analysis of process performance
SIMIT	Virtual commissioning	Native integration with Siemens automation systems

For the purposes of this work, SIMIT from Siemens was selected due to its native, shared-memory integration with SIMATIC systems and TIA Portal (Siemens AG, 2024). Unlike general-purpose tools like MATLAB/Simulink or discrete-event simulators like SIMUL8 that rely on intermediate OPC UA wrappers—which introduce communication latencies and tag configuration overhead—SIMIT operates as a synchronized behavioral extension of the virtual controller (PLCSIM Advanced / S7-1200).

SIMIT supports real-time process simulation and allows direct control via SIMATIC S7 PLCs through standard communication interfaces such as PROFINET, Modbus, and PROFIBUS (Siemens AG, 2024). Its modular block architecture enables seamless implementation of hydraulic, pneumatic, and mass-balance equations while supporting external C++ or MATLAB/Simulink co-simulations if required. This native integration directly satisfies the requirements of **RQ2** by ensuring deterministic execution synchronization and streamlined signal mapping.

3 Mathematical model of the hopper

3.1 Hopper geometry

The analyzed hopper is shaped as a **square pyramidal frustum** with unequal top and bottom bases (Pikna, 2025). The areas of the top base S_{top} and the bottom base S_{bot} are calculated according to the following equations (eq. 1, eq. 2):

$$S_{top} = a \times b \quad (1)$$

$$S_{bot} = m \times n \quad (2)$$

Where a, b are the sides of the top base, and m and n are the sides of the bottom base. The total volume of the hopper is calculated according to the formula (eq. 3) for the volume of a frustum of a pyramid:

$$V = \frac{H}{3} \times (S_{bot} + \sqrt{(S_{bot} \times S_{top})} + S_{top} \quad (3)$$

where H represents the maximum fill level of the material in the hopper.

In the physical model, the bottom base is identical to the discharge gate, while the model considers both dimensions separately for greater versatility.

3.2 Mass flow rate

The basis for calculating the material mass flow rate during discharge is Bernoulli's equation (eq. 4), which expresses the conservation of energy for a flowing fluid:

$$p + \frac{1}{2} \times \rho \times v^2 + \rho \times g \times h = const. \quad (4)$$

where p is the pressure [Pa], ρ is the density of the material in the vessel [kg/m³], v is the flow velocity [m/s], g is the acceleration due to gravity [m/s²], and h is the level height [m].

Using mathematical derivations, we express the equation for the flow rate through the gate (eq. 5). The area of the gate is expressed as follows:

$$S_{gate} = x \times y \quad (5)$$

where x and y are the maximum dimensions of the gate. By combining eq. 4 and eq. 5, the mathematical expression for the flow rate Q_v can be derived (eq. 6):

$$Q_v = S_{gate} \times \sqrt{(2 \times g \times h)} \quad (6)$$

Finally, the mass flow rate $\dot{m}$ is calculated (eq. 7) as follows:

$$\dot{m} = Q_v \times \rho \quad (7)$$

These equations form the computational core of the behavioural model and are implemented within its logical section in the SIMIT environment using standard mathematical blocks.

4 Physical model of the hopper

4.1 Physical model description

SKARTEK s.r.o. has developed a physical model (Fig. 1) of a process hopper as a scaled-down and simplified version of a real-world concrete bucket. This physical model is used to develop and verify control algorithms prior to their deployment on actual industrial equipment. Its main functional units consist of a pneumatic gate actuator, a sensor system, and a PLC control system. (Pikna, 2025)

4.2 Pneumatic actuator

The opening and closing of the discharge gate are performed by a FESTO pneumatic cylinder, type ADN-25-130-A-PPS-A (536350). It is controlled by a FESTO solenoid valve, type VUVG-L10-P53C-T-M5-1P3, which is operated by a digital output of a Siemens S7-1200 PLC. The electrical and pneumatic schematics of the physical model were designed and functionally verified using the Festo FluidSim software.

4.3 Sensors

Three SICK sensors are used to monitor the pneumatic cylinder's position: two digital limit sensors for the end positions (home position HP and working position WP), which send a binary signal to the PLC, and one analog continuous position sensor. The weight of the hopper, including its contents, is measured by an EMSYST EMS 100-5kN strain gauge load cell. The control program implements a TARE function, which subtracts the weight of the empty hopper and displays the net weight of the filled material.

4.4 Control system

A Siemens S7-1200 PLC (type 1214C), programmed in the TIA Portal environment, was selected to control the entire process. The program includes the valve control logic, sensor signal processing, material weight calculation, and a Trace function designed for real-time recording of monitored variables with the option to export to CSV format. The same TIA Portal project was also utilized to interface with the behavioural model in SIMIT.

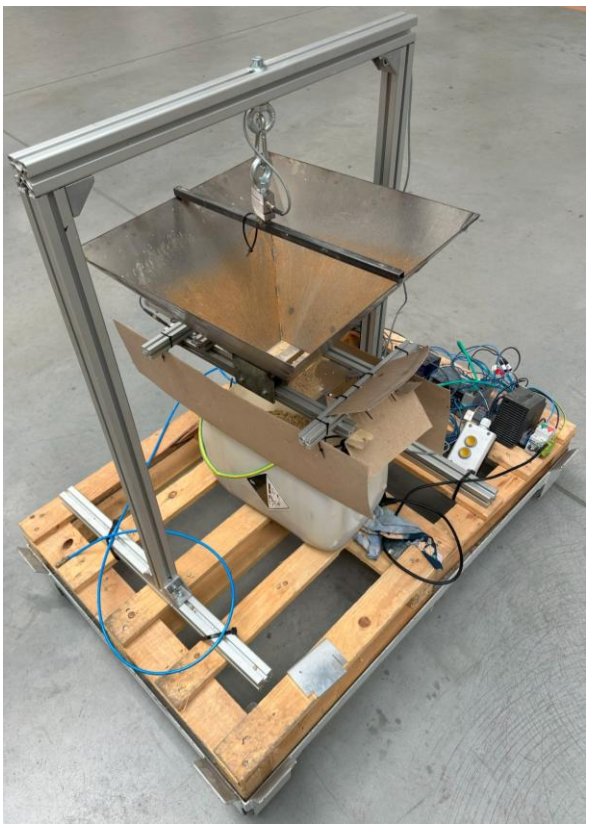


Figure 1. Physical model of the hopper

5 Behavioural model implementation (SIMIT)

5.1 Model architecture

The behavioural model of the hopper is divided into two functional sections: the visual section (Frontend) and the logical section (Backend). This architectural separation ensures model clarity, allows for independent modifications, and facilitates reuse by different users without requiring deep knowledge of the internal logic. The connection between SIMIT and TIA Portal is established through a standard communication interface, with both systems operating using shared PLC variables.

5.2 Frontend

The **Frontend** contains a graphical representation of the hopper and all control elements linked to TIA Portal. Here, the user configures the geometric dimensions of the hopper (top and bottom base sides, height, and gate dimensions) and the material density through editable input fields. The following control elements are available:

- RESET – Virtual filling of the hopper to maximum capacity (calculates the missing weight and adds it to the current value).
- btnOpen / btnClose – Manual opening and closing of the gate in MAN mode.
- Slider – Adjusting the gate opening percentage from 0 to 100% in manual mode.
- Mode selector switch MAN/AUTO – toggles between manual and automated control modes.

In automatic mode (AUTO), the user enters the target batch weight and the gate opening percentage. Upon pressing btnOpen, the gate opens and automatically closes once the target weight is reached. The current filling level of the hopper is continuously displayed via a bar graph and a numerical value in kilograms. Data for validation were collected exclusively in automated mode, thereby eliminating human error associated with manually closing the gate at the correct moment.

5.3 Backend

The Backend performs all mathematical model calculations in real-time. The computational chain consists of the following steps:

1. Calculation of the top and bottom base areas according to the specified dimensions.
2. Calculation of the maximum hopper volume.
3. Calculation of the current gate area as a function of the opening percentage.
4. Calculation of the volumetric and mass flow rates.
5. Continuous update of the material level (height) and weight within the hopper.

All operations are implemented using standard SIMIT blocks: sum, product, quotient, and the SQRT (square root) block. The logic incorporates a ramp block for gradual gate opening, ensuring that the gate does not open instantaneously upon pressing btnOpen. Instead, it opens progressively, aligning with the finite speed of the physical pneumatic actuator.

6 Data collection and experiments

Data collection from both the physical and behavioural models was conducted in TIA Portal using the Trace function. This feature allows for the selection of monitored variables, real-time visualization of their trends, and the export of measured data into a CSV format suitable for further analysis. Two key variables were monitored for both models: the current material weight in the hopper (rWeightNETTO) and the gate opening percentage (rSpoutOpen).

Data Collection Setup and Protocol: Measurements on the physical model were conducted for two initial material weights (5 kg and 10 kg) and four gate opening levels (25%, 50%, 75%, and 100%). To compensate for physical stochasticity, five consecutive repetitions were executed for each configuration, yielding 40 physical measurement runs in total. The physical Siemens S7-1200 PLC sampled and logged sensor data at a high-resolution cycle time of 10ms directly into internal memory trace buffers.

Dataset Characteristics, Volume, and Integrity: Each experimental run recorded data over a duration of 8 to 12 seconds, resulting in approximately 800 to 1,200 data rows per run for the physical setup (totaling ~32,000 recorded data points across all physical trials). Inspection of the raw CSV exports confirmed 0% missing values (no dropped frames or missing timestamps), attributable to the local deterministic logging buffer of the SIMATIC controller.

Raw physical measurements exhibited characteristic signal noise (transient fluctuations of ±0.15kg) caused by mechanical vibrations transmitted to the single-point load cell suspension (EMS 100-5kN) and minor granular settling of the gravel bulk material (density ρ = 1,550kg/m³). Rather than applying destructive low-pass filtering that could alter temporal dynamics, ensemble arithmetic averaging across the 5 runs per configuration was applied to compute a representative mean physical time series. An example of the measured time profile is shown in Fig. 2.

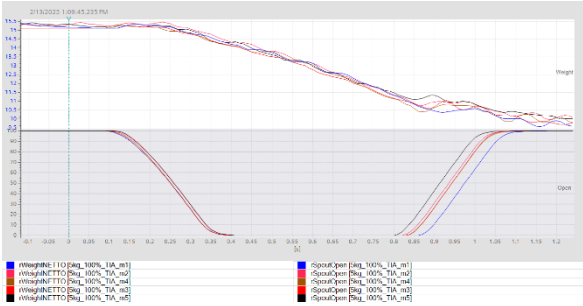


Figure 2. Discharge of 5 kg batch from the physical model at 100% opening.

Simulation Data and Synchronization: In the SIMIT behavioural model, a single execution run per configuration was sufficient, as the mathematical model operates deterministically without external environmental noise. The simulated PLC in SIMIT executed at a fixed sampling slice of 100ms.

Following export, physical datasets (10ms sampling) were synchronized with virtual datasets (100ms sampling) by performing linear interpolation on physical time series to align with a unified 100ms time grid. Temporal alignment was established by defining $t = 0$ms as the exact instant when the gate actuation signal (rSpoutOpen > 0%) was triggered. This standardized dataset provided a rigorous basis for evaluating model accuracy in **RQ1** and interface fidelity in **RQ2**.

7 Results

The consistency between the physical and behavioural models was quantitatively evaluated using the Pearson correlation coefficient for the time series of weight and gate opening. Table 2 presents selected measured values for the reference configuration: a 5 kg batch with 100% gate opening.

Table 2. 5 kg batch at 100% gate opening – selected values (Pikna, 2025)

Time [ms]	Weight physical [kg]	Weight virtual [kg]	Opening physical [%]	Opening virtual [%]
-800	15.32	15.10	100.00	100.00
-400	15.15	15.10	100.00	100.00
0	15.14	15.10	100.00	100.00
400	13.81	14.05	0.00	0.00
800	10.84	12.70	48.45	0.00
1200	9.77	11.12	100.00	0.00
1600	9.92	9.94	100.00	100.00
2000	9.88	9.36	100.00	100.00

Weight Correlation: 0.947 | Gate Opening Correlation: 0.489

An examination of the time steps in Table 2 clarifies the physical mechanisms underlying the discharge process:

- **Static phase** (t = -800ms to 0ms): Both physical and virtual systems register static full-capacity weight (15.14kg physical vs 15.10kg virtual). The slight initial offset (0.04kg to 0.22kg) stems from residual mechanical settling of the bulk material against the hopper walls and load cell tare tolerance.
- **Initiation of discharge** (t = 0ms to 400ms): Upon command execution at t = 0ms, the valve actuates. By t = 400ms, physical material weight decreases to 13.81kg, while virtual weight decreases to 14.05kg. The physical system shows earlier material flow initiation due to immediate opening of the pneumatic seal before the continuous analog position sensor fully records movement.
- **Transient discharge phase** (t = 800ms to 1200ms): At t = 800ms, physical gate opening reaches 48.45% and physical weight drops to 10.84kg. In contrast, the virtual model records 12.70kg. This lag occurs because the virtual model utilizes a linear ramp block to emulate gate motion, whereas the physical FESTO pneumatic cylinder undergoes non-linear pressure buildup, accelerating faster during the initial stroke. By t = 1200ms, physical gate opening reaches 100% while the virtual gate ramp finishes slightly later, causing temporary variance in flow rates.
- **Completion and steady state** (t = 1600ms to 2000ms): As the gate reaches full closure/opening, physical weight stabilizes at 9.88kg and virtual weight at 9.36kg (representing the remaining batch inside the hopper). Both curves converge closely, achieving a high overall weight correlation r = 0.947).

Based on the analysis of all eight configurations, the following summary results were achieved:

- **Weight profile correlation:** range 0.947 – 0.994, arithmetic mean **0.977**.
- **Gate opening profile correlation:** range 0.412 – 0.956, arithmetic mean **0.679**.

A critical evaluation of these trends directly addresses **RQ1**:

- **Intermediate Openings** (50% and 75%): These configurations yielded the highest weight correlations (>0.985). At moderate flow cross-sections, gravel flow matches Bernoulli's ideal fluid assumption closely, maintaining steady quasi-continuous discharge without severe velocity spikes or mechanical clogging.
- **Small Opening** (25%): The lowest weight correlation occurred at 25% opening. At restricted cross-sections, internal granular friction, particle interlocking, and temporary arching (bridging) decrease the effective discharge velocity below theoretical fluid predictions—phenomena not captured by basic Bernoulli equations.

- **Gate Actuator Discrepancy** (100% opening / 5 kg batch): The lower gate opening correlation (0.412 - 0.489) is attributed to pneumatic cylinder dynamics. Under fast actuation, air compressibility, solenoid switching hysteresis, and mechanical friction cause non-linear acceleration that the simplified linear SIMIT ramp block does not replicate. However, because batch dosing algorithms rely primarily on cumulative material weight rather than sub-second valve position, this gate dynamic discrepancy does not impede control code validation.

8 Discussion

The high average correlation of weight profiles (0.977) directly answers **RQ1**, confirming that the mathematical model derived from Bernoulli's equation accurately describes the discharge dynamics of bulk material from a frustum-shaped hopper. The model correctly captures the dependence of the flow rate on both the material level and the gate opening percentage.

The lower average correlation during the gate opening and closing process (0.679) results from the simplified ramp model, which does not account for varying load pressures on the gate at different opening percentages. In the physical system, the pneumatic cylinder overcomes significantly higher material pressure at a 25% opening compared to a 100% opening, leading to distinct gate movement dynamics. However, this deviation does not fundamentally impact the model's utility for verifying dosing algorithms, where the precision of the final batch weight is decisive, rather than the gate's movement dynamics.

Another source of deviations was the mechanical design of the physical model: the hopper was attached to the support frame solely through the EMS 100-5kN load cell, causing all mechanical vibrations generated by the bulk material to be transmitted directly into the measurement path. The hopper was mechanically suspended at a single point, allowing for slight tilting during uneven material distribution during discharge. These effects are not accounted for in the behavioural model, which explains the noise visible at the start of the physical model's measurements.

Regarding **RQ2**, the practical implementation confirmed that native integration between SIMIT and TIA Portal significantly simplifies machine-level virtual commissioning. Shared PLC variable tables eliminated the need for manual tag mapping or middleware configuration. Signal exchange operated deterministically without observed communication latency during 100ms simulation steps. The key advantage of the proposed approach is the full parameterizability of the model: changing any geometric parameter of the hopper (e.g., a different height while maintaining the base dimensions) does not require a new physical model – it is sufficient to adjust the relevant input values in the Frontend. For hoppers of different geometric shapes, such as cylindrical or conical hoppers, a change in the computational equations in the Backend is sufficient without the need to redesign the entire model architecture. (Wu et al., 2023)

While the present study validates the core behavioral model using Bernoulli's equation across 8 different physical configurations, further mathematical enhancements are necessary to overcome observed edge-case deviations. Future iterations of the mathematical model will incorporate Discrete Element Method (DEM) physics to capture granular material dynamics (such as particle friction and arching effects during partial gate openings) as well as non-linear pressure dynamics within the pneumatic actuator.

9 Conclusion

The presented behavioural model of the bulk material hopper was successfully designed, implemented in the SIMIT environment, and experimentally validated by comparison with a physical hopper model at SKARTEK s.r.o.

In response to **RQ1**, experimental validation demonstrated an average weight profile correlation of 0.977 across eight measurement configurations, proving that the Bernoulli-based mathematical model provides sufficient accuracy for testing PLC dosing logic without physical prototypes.

In response to **RQ2**, the native coupling between SIMIT and TIA Portal demonstrated seamless tag synchronization, zero-latency signal exchange, and rapid parameter modification, proving that dedicated native simulation environments provide superior workflow efficiency for machine-level virtual commissioning compared to generic simulation tools.

The key advantage of the behavioural model over a physical prototype is significantly lower cost, shorter preparation time, and easy parameterization for various hopper variants. When developing projects involving multiple hoppers with different dimensions, it is sufficient to adjust the geometric parameters in the virtual model without the need to build new physical prototypes. This directly fulfills the principles of Industry 4.0 – shortening lead times, reducing development costs, and increasing the quality of the final control algorithm.

In the future, the authors are encouraged to expand this methodology through additional case studies involving different industrial hopper geometries (e.g., cylindrical, conical, and asymmetric hoppers) and various bulk materials with distinct flow properties. Extending the digital twin framework to multi-hopper production lines and integrating real-time DEM simulations will provide even stronger empirical evidence for automated virtual commissioning across broader Industry 4.0 applications.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used Google Gemini in order to assist with English language translation, grammatical refinement, formatting, and structural revision based on peer review feedback. After using this tool, the authors reviewed, critically evaluated, and edited the content as needed, and take full responsibility for the content of the published article.

References

Barth, M., & Fay, A. (2013). Automated generation of simulation models for virtual commissioning of automated systems. IFAC Proceedings Volumes, 46(9), 447–454. https://doi.org/10.3182/20130619-3-RU-3018.00028

Hoffmann, P., Thomas, C., & Zäh, M. F. (2021). Simulation-based virtual commissioning of automated manufacturing systems: A literature review. The International Journal of Advanced Manufacturing Technology, 114(9), 2589–2608. https://doi.org/10.1007/s00170-021-06899-7

Klüber, M., & Stark, R. (2020). Model-based virtual commissioning for bulk material processing systems. Journal of Industrial Information Integration, 18, 100142. https://doi.org/10.1016/j.jii.2020.100142

Kopček, M., Križanová, G., & Ružarovský, R. (2022). A brief survey of creating a digital twin architecture for virtual commissioning on the machine level. Research Papers Faculty of Materials Science and Technology Slovak University of Technology, 30(50), 71–80. https://doi.org/10.2478/rput-2022-0009

Lechler, T., Fischer, E., Metzner, M., & Mayr, A. (2019). Virtual commissioning – Scientific review and exploratory use cases in advanced production systems. Procedia CIRP, 81, 1125–1130. https://doi.org/10.1016/j.procir.2019.03.279

Pikna, R. (2025). Behavioural model of a bulk material hopper for virtual commissioning [Bachelor's thesis]. Slovak University of Technology in Bratislava, Faculty of Materials Science and Technology in Trnava.

Reinhart, G., & Wünsch, G. (2007). Economic application of virtual commissioning to mechatronic production systems. Production Engineering, 1(4), 371–379. https://doi.org/10.1007/s11740-007-0066-0

Siemens AG. (2024). Automation framework for SIMATIC PLC/HMI and SINAMICS drives. Retrieved from https://support.industry.siemens.com/cs/document/109817223/automation-framework-for-simatic-plc-hmi-and-sinamics-drives

Tjahjono, B., Esplugues, C., Ares, E., & Pelaez, G. (2017). What does Industry 4.0 mean to supply chain? Procedia Manufacturing, 13, 1175–1182. https://doi.org/10.1016/j.promfg.2017.09.191

Wu, Y., Zhang, M., Chen, Y., & Li, S. (2023). A comprehensive review of digital twin: Definitions, applications, and challenges. Journal of Manufacturing Systems, 68, 346–361. https://doi.org/10.1016/j.jmsy.2023.01.004

Flexibility in the low voltage PV-rich distribution network with demand response and BESS

Zvonimir Šimić, Danijel Topić, Marina Dubravac, Lucija Gujić
Josip Juraj Strosssmayer University of Osijek
Faculty of Electrical Engineering, Computer Science and Information Technology Osijek
Kneza Trpimira 2B, 31000 Osijek, Croatia
{zvonimir.simic, danijel.topic, marina.dubravac}@ferit.hr,
lucija.gujic@student.ferit.hr

Abstract. *With the increase of electricity generation from variable renewable energy sources, such as photovoltaic (PV) systems and wind power plants, the flexibility of the power system is required. In this paper, the flexibility of the power system is demonstrated on the example of a PV-rich low-voltage (LV) distribution network. Sources of the flexibility implemented in this example include demand response (DR) and battery energy storage systems (BESS). Both sources of flexibility on their own significantly contribute to increasing the flexibility in the LV distribution network, while the best results were achieved with the simultaneous implementation of both sources.*

Keywords. Flexibility, Photovoltaic, PV-rich distribution network, Demand response, Battery energy storage system

1 Introduction

The energy sector is responsible for 75 % of greenhouse gas emissions in the European Union (EU) (European Commission, 2023), and to reduce greenhouse gas emissions, of which carbon dioxide is the most abundant, conventional fossil fuel-based energy sources are being replaced by low-carbon renewable energy sources (RES). The most represented RES used for electricity generation are wind power plants with 37.5 % and solar photovoltaic (PV) systems with 27.5 % (Eurostat, 2026). Despite the clean electricity, the mentioned RESs are characterized by the unpredictability and variability of electricity generation. Unlike conventional power systems, where electricity production follows and adapts to electricity demand, in power systems with a high share of electricity generation from unpredictable and variable RESs, it is the entire power system that adapts to fast and unpredictable changes in electricity generation. The ability of the power system to react quickly to rapid changes in electricity generation and demand, even when events are not predicted, is defined as the flexibility of the power system (Rahman et al., 2024).

According to (Babatunde et al., 2020), the flexibility of the power system refers to the level to which a power system is capable of adapting the generation and consumption of electricity. The definition of the International Energy Agency (IEA, 2019) emphasizes that the flexibility of the power system should be reliable and cost-effective. According to the International Renewable Energy Agency (IRENA, 2018), flexibility must be implemented in all parts of the power system, from flexible electricity generation, more flexible transmission and distribution systems, to more flexible demand and more installed storage capacity.

The implementation of flexibility at all levels of the power system is not enough on its own but requires interaction between all levels of the power system (Salman et al., 2022). In addition to the integration of variable RESs, the drivers of the flexibility of the power system are fluctuating fuel prices, especially the price of natural gas, and the uncertainty of electricity consumption, which is becoming even more pronounced with the increasing number of electric vehicles (EVs) (Kaushik et al., 2022). The main sources of flexibility are demand side management (DSM) and demand response (DR), energy storage systems, and flexible electricity generation (Babatunde et al., 2020). Some other sources of flexibility include the diversity of RESs, a dynamic capability assessment of power lines, new ancillary services, and restructuring the electricity market (Rahman et al., 2024). Active power curtailment and electricity price mechanisms can also be considered as sources of flexibility of the power system (Abrahamsen et al., 2026).

The number of PV systems installed in Croatia in February 2025 was 27,540, with a total installed power of 835 MW (TehnoEko, 2025). In October 2025, this number was increased to 38,947 PV systems, and the total installed power increased to 1106 MW (Lider, 2025). In the majority, PV systems are connected to low voltage (LV) distribution networks, and the significant increase in PV systems requires the need for flexibility in LV distribution networks. Distribution System Operators (DSOs) can use flexibility to avoid or post-

pone the high investments in network infrastructure, mainly in new transformers and power lines (Mehinovic et al., 2025).

The sources of flexibility that can be applied in LV distribution networks are DR and energy storage systems. The applications of sources of flexibility are demonstrated in this paper through power flow simulations of a PV-rich LV distribution network. In this work, the impact of DR and energy storage is analyzed based on results of power flow simulations for different scenarios. DR is achieved by shifting the real measured consumption profiles of household appliances. In addition to a separate analysis of the impact of DR and energy storage on the flexibility of the power system, the combined impact is also analyzed and compared with isolated impacts.

In the next chapter, the LV distribution network case study, input data, and the sources of flexibility are explained. In Chapter 3, results are presented and analyzed. The last chapter is the conclusion.

2 Flexibility in a PV-rich LV distribution network

The flexibility of the power system is demonstrated by simulation of the power flows in a PV-rich LV distribution network. Fig. 1 represents the schematic diagram of the LV distribution network used in the case study. There are 4 radial feeders with 10 grid users in each feeder, or 40 grid users in total. Of the 40 grid users, 20 are prosumers, and 20 are consumers.

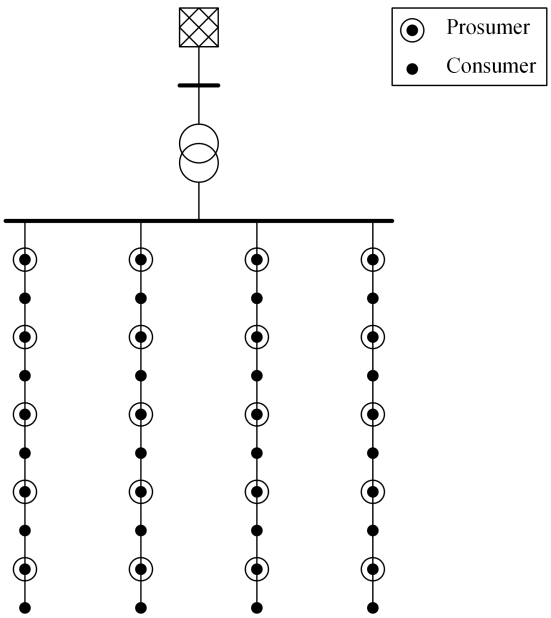


Figure 1. Low voltage distribution network

Consumers are grid users who buy electricity from the electricity supplier, while prosumers are grid users who produce electricity for their own needs from their own PV power plant. Each prosumer has a PV power plant connected behind the meter. In addition to a PV power plant, there is a battery energy storage system (BESS) that allows storing produced electricity and using it when electricity generation is not available.

Table 1 represents the parameters of the LV distribution network that include the rated power of each PV power plant. As the rated power of a PV power plant for each prosumer is 3 kW, the total installed PV power with 20 prosumers in the observed LV distribution network is 60 kW.

Table 1. Network parameters

Parameter	Value
Transformer ratio	20/0.4 kV
Transformer rated power	630 kVA
LV feeder length	120 m
LV cable type	NAVY 4x120 SE
PV rated power	3 kW
Total installed PV power	60 kW

2.1 Input data

The assumption that 10 households in one feeder have the same daily consumption profile is considered due to the simplification. As there are 4 feeders in the LV distribution grid, 4 different consumption profiles are represented. Different daily consumption profiles with a 15-minute time step are presented in Fig. 2. There is also a daily profile of electricity generation from a PV power plant on a clear, sunny day. The peak of electricity generation occurs in the middle of the day, while the peak of electricity consumption occurs in the evening. The electricity generation during a clear, sunny day is predictable and does not represent a challenge for distribution system operators in terms of the variability of the electricity generation as it is during cloudy days. A clear sunny day is selected as the worst scenario due to the highest difference between electricity generation and consumption.

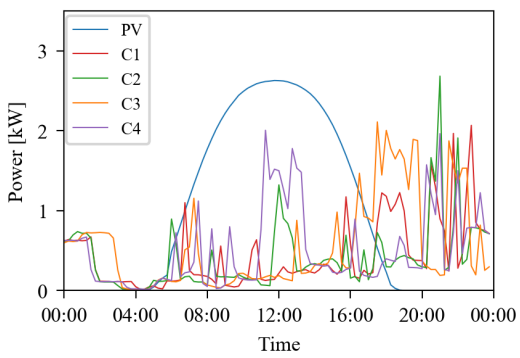


Figure 2. Daily consumption profiles and electricity generation from a PV power plant

2.2 Demand response

Time-shiftable devices are household appliances that can be used at any time of the day. The idea is to match the use of time-shiftable devices with the electricity generation. DR as a source of flexibility is shown in Fig. 3. A significant share of the electricity consumption in households is for domestic water heaters, which can also contribute to a demand-side flexibility (Seljom et al., 2024). Around 50 % of global energy consumption is for heating and cooling, and with smart control it can also contribute to demand-side flexibility (Parreño-Rodríguez et al., 2026).

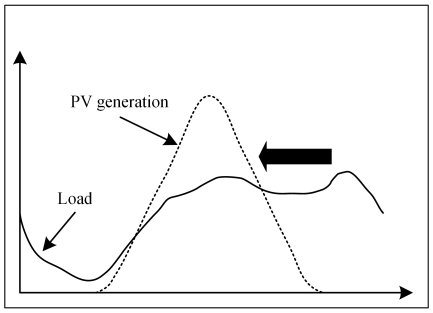


Figure 3. Demand response as the source of flexibility of the power system

DR in this study is achieved by shifting the time of use of the washing machine, tumble dryer, and dishwasher to a period with electricity generation. The daily electricity consumption for C1, C2, C3, and C4 are 10.3, 9.8, 13.4, and 12.1 kWh respectively. The electricity consumption of time-shiftable devices is 2.75 kWh, which represents 27, 28, 20, and 23 % of daily consumption, respectively. The daily electricity generation is 23 kWh, and the daily consumption of time-shiftable devices is equal to 12 % of the daily electricity generation. The daily consumption profiles of 4 different users, with a 15-minute time step, when the demand response is applied, are presented in Fig. 4. It can be seen that electricity consumption is better matched with electricity generation.

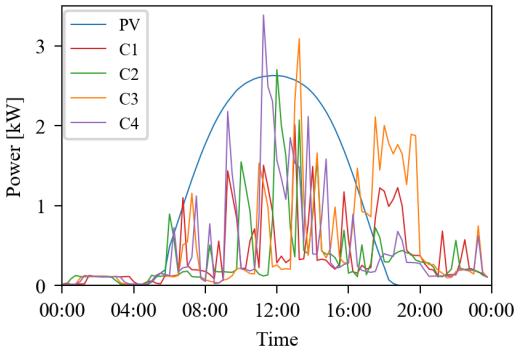


Figure 4. Daily consumption profiles with applied demand response

All prosumers and consumers participate in DR. Before DR, the start times for the washing machine and dishwasher were between 20:00 and 22:00, while the start time for the tumble dryer was after the end of the washing machine cycle. With DR, the washing machine start time is between 08:00 and 12:00, the dishwasher start time is between 12:00 and 14:00, while the tumble dryer start time is after the end of the washing machine cycle.

2.3 Energy storage

Energy storage is an important source of flexibility in the power system as it helps to balance variable electricity generation and consumption (Khalili et al., 2025). Different storage technologies are represented, such as pumped hydro, flywheels, compressed air, hydrogen, thermal storage, super capacitors, and batteries (Degefa et al., 2021). As a source of flexibility in the LV PV-rich distribution networks, BESSs are the most represented storage technology. Using BESS as a source of flexibility is presented in Fig. 5. The role of BESS is to store electricity generated from a PV power plant instead of exporting excess to the grid. The stored electricity can be used to supply electricity consumption instead of importing electricity from the grid.

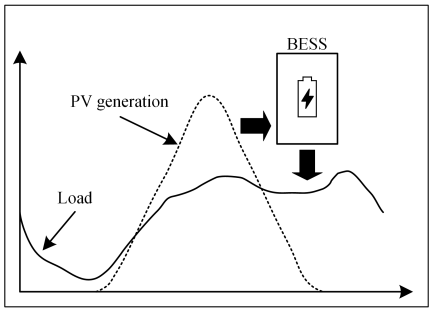


Figure 5. BESS as the source of flexibility of the power system

In this work, two different BESS capacities are used. A larger 10.24 kWh BESS is used when BESS is the only source of flexibility, and a smaller 5.12 kWh BESS is used when BESS and DR are sources of flexibility, Table 2 represents the parameters of BESS.

Table 2. BESS parameters

Parameter	Value
BESS capacity	100 Ah / 200 Ah
Rated voltage	51.2 V
Rated energy	5.12 kWh / 10.24 kWh
Maximum power	3 kW
Minimum SOC	0.2
Maximum SOC	1.0
Initial SOC	0.2
Final SOC	0.2

The rule-based algorithm of the BESS charging and discharging strategy is shown in the flow diagram in Fig. 6. The input data for the algorithm are the electricity generation and consumption data, while the output data are the BESS charging and discharging power.

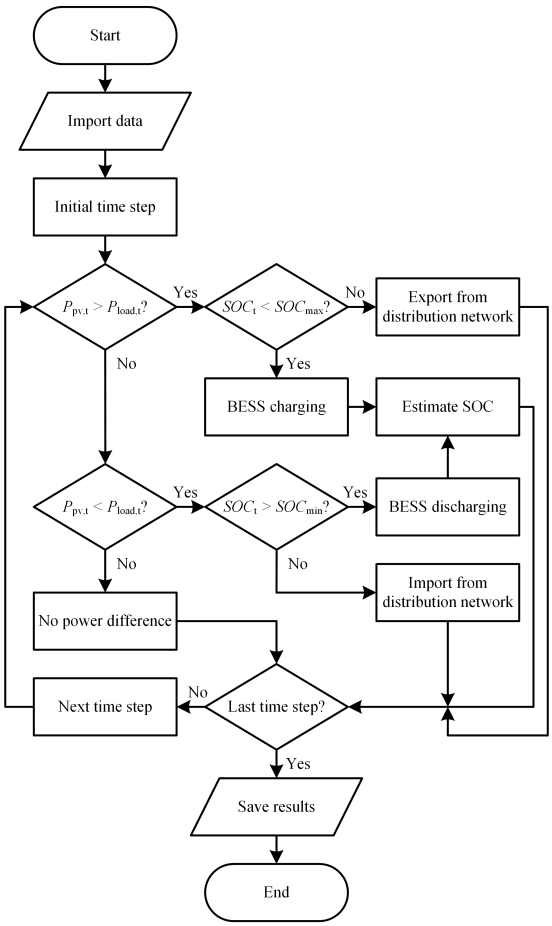


Figure 6. Flow diagram of the BESS charging and discharging strategy

2.4 Scenarios overview

Power flow simulations are performed for an initial scenario without PV power plants and for 4 different scenarios with integrated PV power plants, as presented in Table 3. In Scenario 1, there are no sources of flexibility. In Scenarios 2 and 4, the source of flexibility is the DR, while energy storage is the source of flexibility in Scenarios 3 and 4.

Table 3. Scenarios overview

Scenario	PV	DR	BESS
Initial	-	-	-
S1	+	-	-
S2	+	+	-
S3	+	-	+
S4	+	+	+

3 Results

In this section, the results are presented through five different scenarios described in the previous section. The scenarios are compared according to the electricity exchange with the main grid. Power flow simulations are conducted using Panda Power open-source software in a Python environment, with a 15-minute time step. In addition, the loading of the initial section of feeder 2 and the voltage at the end of feeder 2 are compared for different scenarios.

3.1 Scenario 1

The power of electricity generation and consumption in Scenario 1 is presented in Fig. 7. Between 06:30 and 17:00, more electricity is generated from PV power plants than is needed to cover the electricity demand.

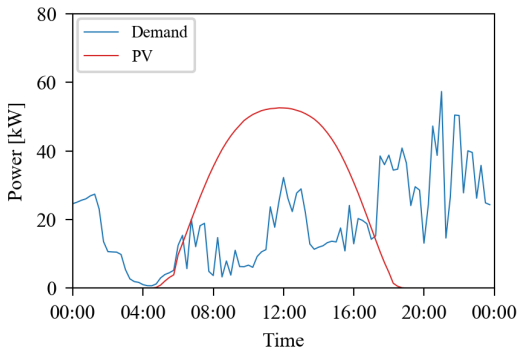


Figure 7. Electricity generation from PV power plants and electricity consumption in Scenario 1

Scenario 1 is compared to the initial scenario through a power exchange with the main grid, as shown in Fig. 8. In the initial scenario, the exchange power is positive for the entire day, which indicates that electricity is imported from the main grid for the entire day. In Scenario 1, negative power occurs between 06:30 and 17:00, which indicates that an excess of produced electricity is exported to the main grid.

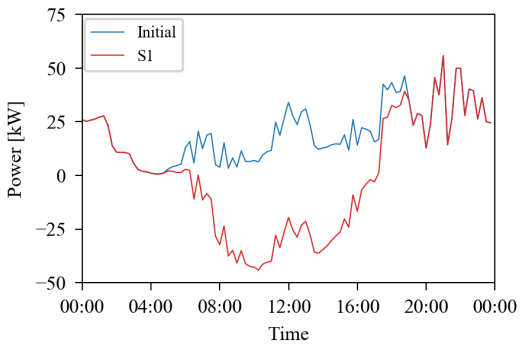


Figure 8. Exchange with the main grid in Scenario 1

3.2 Scenario 2

In Scenario 2, the DR is applied and part of the electricity consumption is shifted from evening to midday. The higher electricity consumption during the day and the lower electricity consumption in the evening are more in line with the daily electricity generation curve, as shown in Fig. 9.

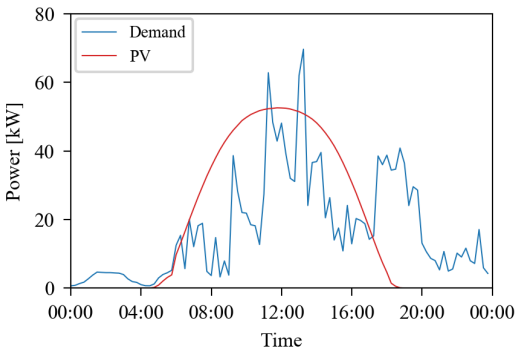


Figure 9. Electricity generation from PV power plants and electricity consumption in Scenario 2

With DR in Scenario 2, between 09:00 and 15:30, less electricity is exported than in Scenario 1, as presented in Fig. 10. In the evening, from 20:00 to 00:00, the amount of imported electricity is decreased compared to Scenario 1.

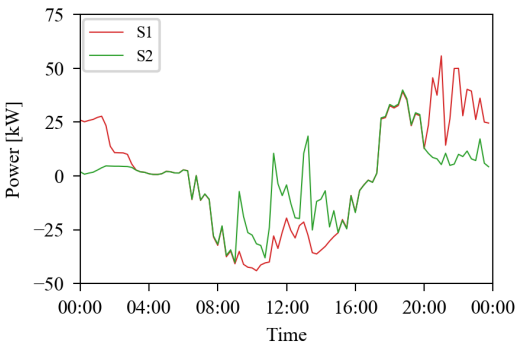


Figure 10. Exchange with the main grid in Scenario 2

3.3 Scenario 3

In Scenario 3, the source of flexibility is BESS. In Fig. 11, the positive power of BESS indicates charging, while the negative power of BESS indicates discharging. Charging during a period of high electricity generation and low electricity consumption helps reduce the amount of electricity exported to the main grid, which can be seen in Fig. 12 as a decrease in the power of exported electricity in Scenario 3. However, BESS discharge after 17:30 helps reduce the amount of imported electricity from the main grid when there is not enough electricity generation.

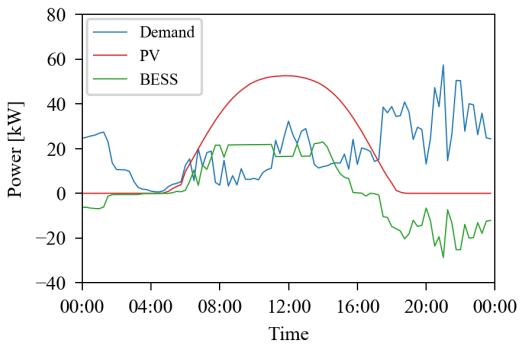


Figure 11. Electricity generation, consumption and BESS utilization in Scenario 3

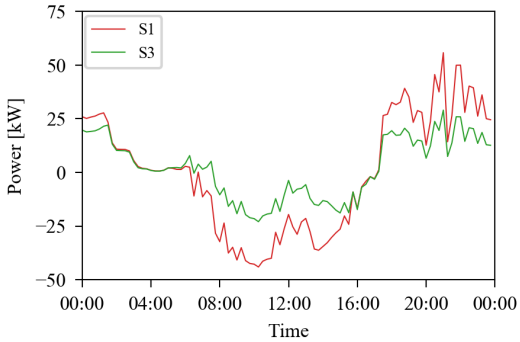


Figure 12. Exchange with the main grid in Scenario 3

3.4 Scenario 4

The sources of flexibility in Scenario 4 are DR and BESS. As electricity demand shifts to the period with high electricity generation, BESS is discharged between 17:30 and 20:00 when consumption is high, compared to Scenario 3, when BESS is discharged between 17:30 and 00:00, as can be seen in Fig. 13.

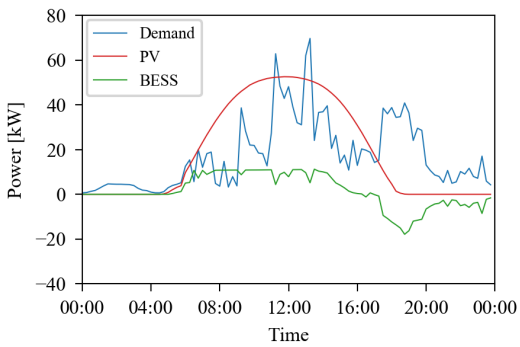


Figure 13. Electricity generation, consumption and BESS utilization in Scenario 4

From the comparison of the exchange power in Scenario 1 and Scenario 4, presented in Fig. 14, it can be seen that the amount of electricity exported to the main grid is significantly lower in Scenario 4 compared

to Scenario 1. However, the amount of electricity imported from the main grid in the period between 17:30 and 00:00 is significantly lower in Scenario 4 compared to Scenario 1.

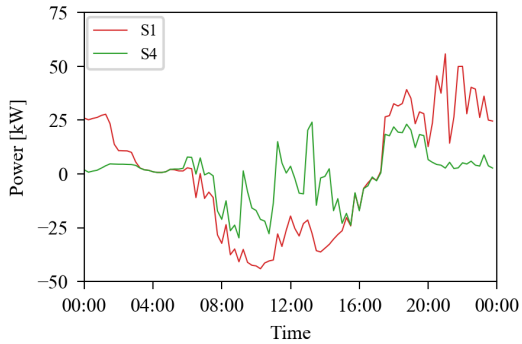


Figure 14. Exchange with the main grid in Scenario 4

The amount of electricity exchanged with the main grid for an entire day is presented in Table 4. In Scenario 2, the DR leads to a significant reduction in imported electricity. In Scenario 3, the use of BESS results in a greater reduction in exported electricity. However, the best results are achieved with DR and BESS as sources of flexibility. The total amount of imported electricity in Scenario 4 is 499 kWh, which is 58 % lower than in Scenario 1 and 74 % lower than in the initial scenario. The total amount of electricity exported in Scenario 4 is 398 kWh, which is a 63 % decrease compared to Scenario 1.

Table 4. Electricity exchange

Scenario	Import [kWh]	Export [kWh]
Initial	1954	0
S1	1193	1082
S2	637	687
S3	753	456
S4	499	398

Due to the large cross-sectional area of LV cables, the line loading is not high, as can be seen in Fig. 15. The highest line loading occurs in the evening when electricity consumption is high. The implementation of BESS in Scenario 3 slightly contributes to the reduction of the line loading, while the impact of DR on the line loading in Scenarios 2 and 4 is more significant. However, the line loading increase caused by PV integration can be significantly reduced with DR and also with BESS. Fig. 16 represents the voltage at the end of feeder 2, and due to the short length of the feeders, the voltage variations are not significant. The voltage drop in the evening improves slightly when BESS is applied in Scenario 3 and more significantly when DR is applied in Scenarios 2 and 4. The voltage increase caused by PV integration can be reduced with BESS and also with DR.

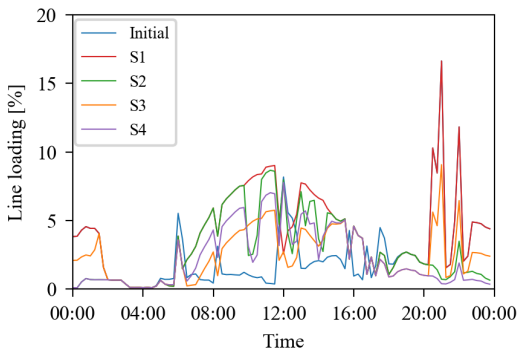


Figure 15. Loading of the first line section in feeder 2

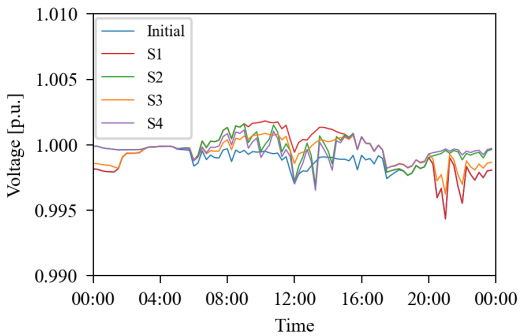


Figure 16. Voltage at the end of feeder 2

4 Conclusion

With the increase in variable and intermittent RESs, there is an increasing need for flexibility of the power system. The focus of this work is on the flexibility in the LV distribution network with a significant increase in the number of installed PV power plants. The sources of flexibility in the PV-rich LV distribution network in the case study are DR and BESS. The impact of flexible sources was evaluated according to the exchange of electricity with the main grid. Although the contribution of DR is more visible in the reduction of electricity imported from the main grid, the contribution of BESS is more significant in the reduction of electricity exported to the main grid. The contribution of DR and BESS together as sources of flexibility is significant in the reduction of imported and exported electricity.

In recent years, the number of EVs has increased, which has contributed to an increase in total electricity consumption. Future work may include EVs as sources of flexibility in LV distribution networks. As domestic water heaters and room heating and cooling have a significant share in electricity consumption, their contribution to flexibility in PV-rich LV distribution networks should be investigated in future work.

Acknowledgments

This work was funded by the Faculty of Electrical Engineering, Computer Science and Information Technology Osijek project FERIT ZP-2025-"Aktivne niskonaponske distribucijske mreže za održivu budućnost.

References

Abrahamsen, F. E., Norheim, I., & Uhlen, K. O. (2026). Flexibility and controllability in low-voltage distribution grids under high pv penetration. *Energies*, *19*(9). https://doi.org/10.3390/en19092072

Babatunde, O. M., Munda, J. L., & Hamam, Y. (2020). Power system flexibility: A review. *Energy Reports*, *6*(2), 101–106. https://doi.org/10.1016/j.egyr.2019.11.048

Degefa, M. Z., Sperstad, I. B., & Sæle, H. (2021). Comprehensive classifications and characterizations of power system flexibility resources. *Electric Power Systems Research*, *194*. https://doi.org/10.1016/j.epsr.2021.107022

European Commission. (2023). *Directive (eu) 2023/2413 of the european parliament and of the council*. https://eur-lex.europa.eu/eli/dir/2023/2413/oj/eng (Accessed: 25.05.2026).

Eurostat. (2026). *47% of eu's electricity came from renewables in 2025*. https://ec.europa.eu/eurostat/web/products-eurostat-news/w/ddn-20260319-2 (Accessed: 25.05.2026).

IEA. (2019). *Status of power system transformation 2019*. https://www.iea.org/reports/status-of-power-system-transformation-2019 (Accessed: 25.05.2026).

IRENA. (2018). *Power system flexibility for the energy transition*. https://www.irena.org/-/media/Files/IRENA/Agency/Publication/2018/Nov/IRENA_Power_system_flexibility_1_2018.pdf (Accessed: 25.05.2026).

Kaushik, E., Prakash, V., Mahela, O. P., Khan, B., El-Shahat, A., & Abdelaziz, A. Y. (2022). Comprehensive overview of power system flexibility during the scenario of high penetration of renewable energy in utility grid. *Energies*, *15*(2). https://doi.org/10.3390/en15020516

Khalili, S., Lopez, G., & Breyer, C. (2025). Role and trends of flexibility options in 100% renewable energy system analyses towards the power-to-x economy. *Renewable and Sustainable Energy Reviews*, *212*. https://doi.org/10.1016/j.rser.2025.115383

Lider. (2025). *Solarni boom u hrvatskoj: 80 posto kapaciteta drži biznis*. https://lidermedia.hr/zeleno-i-digitalno/solarni-boom-u-hrvatskoj-80-posto-snage-na-krovovima-drzi-biznis/ (Accessed: 26.05.2026).

Mehinovic, A., Suljanovic, N., & Zajc, M. (2025). Quantifying the impact of flexibility asset location on services in the distribution grid: Power system and local flexibility market co-simulation. *Electric Power Systems Research*, *238*. https://doi.org/10.1016/j.epsr.2024.111037

Parreño-Rodríguez, A., Molina-García, A., Skarmeta, A. F., & Ramallo-González, A. P. (2026). Demand response impact on medium and low-voltage networks with high renewable integration: A multi-scale perspective. *Sustainable Energy, Grids and Networks*, *46*. https://doi.org/10.1016/j.segan.2026.102201

Rahman, M. M., Dadon, S. H., He, M., Giesselmann, M., & Hasan, M. M. (2024). An overview of power system flexibility: High renewable energy penetration scenarios. *Energies*, *17*(24). https://doi.org/10.3390/en17246393

Salman, U. T., Shafiq, S., Al-Ismail, F. S., & Khalid, M. (2022). A review of improvements in power system flexibility: Implementation, operation and economics. *Electronics*, *11*(4). https://doi.org/10.3390/electronics11040581

Seljom, P., Rosenberg, E., & Haaskjold, K. (2024). The effect and value of end-use flexibility in the low-carbon transition of the energy system. *Energy*, *292*. https://doi.org/10.1016/j.energy.2024.130455

TehnoEko. (2025). *Pregled solarnih elektrana u hrvatskoj: Od kućanstava do velikih igrača*. https://www.tehnoeko.com.hr/energetika/pregled-solarnih-elektrana-u-hrvatskoj-od-kucanstava-do-velikih-igraca/a/10592 (Accessed: 26.05.2026).

Cyber Security and Resilience of Artificial Intelligence Systems in the Nuclear Sector: A Literature Review

Tadej Šeruga, Igor Bernik
University of Maribor
Faculty of Criminal Justice and Security
Kotnikova 8, SI-1000 Ljubljana, Slovenia
tadej.seruga@student.um.si,igor.bernik@um.si

Abstract. *This literature review examines the cyber security and resilience of artificial intelligence (AI) systems in the nuclear sector. Based on a structured review of publications from Scopus and Web of Science, the paper analyses how recent literature addresses AI systems as cyber-vulnerable and resilience-critical assets. The findings show that AI is often discussed as a tool for cyber security, while fewer studies examine the security of AI systems themselves. Key concerns include adversarial attacks, data manipulation, model robustness, verification and validation, explainability, secure lifecycle management, and nuclear-specific deployment constraints.*

Keywords. cyber security, resilience, artificial intelligence, machine learning, operational technology, nuclear sector

1 Introduction

The nuclear sector is one of the most safety- and security-sensitive areas of critical infrastructure. It includes nuclear power plants (NPPs), research reactors, fuel cycle facilities, radioactive waste management facilities, and other installations where nuclear or radioactive material is used, processed, stored, or controlled. Nuclear energy is increasingly discussed as a strategic low-carbon energy source that can contribute to energy security and decarbonisation objectives (Huang et al., 2023; Khadka et al., 2024).
At the same time, the safety-sensitive nature of nuclear activities means that failures, accidents, malicious acts, or cyber incidents may have consequences that extend beyond ordinary business disruption and include equipment damage, loss of safety functions, radiological concerns, environmental impact, injury or death, loss of sensitive information, and loss of public confidence (Khadka et al., 2024; Lezzi et al., 2025). Nuclear facilities are therefore operated under strict safety, security, and regulatory requirements.

Although cyber security in nuclear facilities includes the protection of digital assets, control networks, communication pathways, safety-related systems, and operational technology (OT) environments, it cannot be reduced to conventional information technology security.

Digital systems may support nuclear safety, nuclear security, physical protection, emergency preparedness, operational monitoring, and control functions. As a result, a cyber incident may affect not only the confidentiality, integrity, or availability of information, but also the behaviour of physical processes and the ability of the facility to maintain safe and secure operation (Ayodeji et al., 2023; Lezzi et al., 2025).

As the sector undergoes a gradual but significant digital transformation, it is also adopting more advanced digital and data-driven technologies such as artificial intelligence. In this paper, AI is used as a broad term covering data-driven and learning-based systems, including machine learning (ML) and deep learning. In the nuclear sector, AI is increasingly discussed in relation to monitoring, diagnostics, anomaly detection, predictive maintenance, digital twins, remote inspection, decision support, autonomous control, and cyber security (Arhouni et al., 2025; Cancila et al., 2024; Huang et al., 2023; Khadka et al., 2024; Mendoza et al., 2026; Park & Koo, 2026). Within cyber security, AI-based and machine-learning-based methods are being used for anomaly detection, intrusion detection, attack classification, threat hunting, vulnerability assessment, attack path analysis, risk assessment, and predictive intelligence (Blakely et al., 2025; Lezzi et al., 2025). AI for cyber security is therefore a growing research area in industrial control system (ICS) security and nuclear cyber security.

The nuclear sector adds several constraints that make this topic distinct from general AI cyber security or general ICS cyber security. Nuclear systems are characterised by strict safety requirements, regulatory oversight, long system lifecycles, limited tolerance for unverified changes, restricted access to operational data, and the need to distinguish between equipment faults, normal transients, and malicious cyber activity. These characteristics affect how AI systems can be

designed, tested, deployed, and trusted (Yockey et al., 2023).

Existing literature has already examined the use of AI-based methods for cyber security in nuclear power plants, particularly for threat detection, cyber risk assessment, and AI-based countermeasures (Lezzi et al., 2025). However, less attention has been given to the opposite perspective: the cyber security and resilience of AI systems themselves. This review therefore focuses on AI systems as cyber-vulnerable and resilience-critical assets in the nuclear sector.

2 Background

For the purpose of this review, the term AI system is used broadly and refers not only to a trained model, but also to the wider technical and organisational environment required for its development, operation and maintenance. This includes data, model architecture and parameters, software libraries, deployment platforms, interfaces with instrumentation and control (I&C) systems, monitoring mechanisms, update processes, and human interpretation of outputs. This broader view is important because cyber security risks may arise across the AI lifecycle, from data collection and model development to deployment, operation, maintenance, and retirement.

AI systems can support nuclear applications in different ways. Relevant applications include monitoring, diagnostics, predictive maintenance, digital twins, remote inspection, reactor design support, anomaly detection, cyber response, and operational decision support (Huang et al., 2023; Khadka et al., 2024; Mendoza et al., 2026). In some cases, AI may become part of operational or safety-relevant decision-making chains.

AI assurance in the nuclear sector is further complicated by data and explainability challenges. Many AI applications in nuclear reactor research rely on simulated data because real plant data are difficult to obtain. This may lead to problems such as data imbalance, limited representativeness, and uncertainty regarding generalisability to operational environments (Huang et al., 2023). At the same time, high-performing AI models may behave as black boxes, making them difficult to explain, validate, or justify in safety-sensitive contexts.

In the nuclear sector, cyber security is commonly addressed through structured computer security programmes subject to regulatory oversight. International guidance and national regulatory frameworks emphasise the need to identify, protect, and maintain computer-based systems and digital assets that support safety, security, and emergency preparedness functions. The International Atomic Energy Agency addresses computer security as part of nuclear security in its implementing guidance and provides more detailed technical guidance for nuclear facilities (IAEA, 2021a, 2021b).

A central concept in this context is the identification of cyber-relevant and critical (sensitive) digital assets. Critical digital assets may include digital components and systems used in instrumentation and control, reactor protection, engineered safety features, distributed control, feedwater control, turbine control, emergency diesel generator systems, physical protection, emergency preparedness, communication, monitoring, and network security (Eggers, 2021; IAEA, 2021b; Lezzi et al., 2025).

AI systems that support safety-relevant, security-relevant, emergency preparedness, operational monitoring, autonomous control, or cyber defence functions may therefore need to be assessed as cyber-relevant digital assets. Depending on their function and on the potential consequences of compromise, some AI systems may also need to be considered as critical digital assets. As AI adoption increases, governance, transparency, robustness, and cyber security of AI systems are also likely to become increasingly relevant to nuclear assurance and regulatory discussions.

Against this programme-based background, the growing use of AI introduces a new dimension to the existing challenges of cyber security in the nuclear sector. AI systems are not only tools that may support cyber security; they may also become cyber-vulnerable assets.

The cyber security of AI systems differs from conventional software cyber security in several ways. AI systems may be vulnerable to a set of threats such as manipulation of training data, poisoning attacks, evasion attacks, model manipulation, data drift, insecure model updates, trojan attacks and dependency vulnerabilities in AI software libraries (Khadka et al., 2024; Ndiaye et al., 2024; Yockey et al., 2023). In industrial control and nuclear environments, these threats are further complicated by physical process constraints. Manipulated data must often remain plausible within the expected range of sensor and actuator behaviour, and the consequences of incorrect model outputs may affect not only information assets but also physical process states.

Resilience is therefore a key concept in this review. In the context of AI systems in the nuclear sector, resilience refers to the ability of an AI-enabled function to maintain, recover, or support safe and secure operation under adverse cyber conditions. This includes the ability to detect manipulation, tolerate degraded or uncertain data, resist adversarial inputs, provide meaningful explanations, support operator decision-making, and avoid unsafe responses.

For these reasons, verification, validation, robustness testing, explainability, and lifecycle assurance are central to the cyber security and resilience of AI systems in the nuclear sector. AI components in nuclear I&C systems cannot be accepted only because they achieve high accuracy on a test dataset. They must also be evaluated for reliability, transparency, robustness, adversarial behaviour, lifecycle control, and compatibility with regulatory

expectations (Park & Koo, 2026). Similarly, secure lifecycle approaches for AI-enabled industrial and nuclear systems should include data and model management, risk assessment, adversarial testing, explainable AI (XAI), and resilience considerations (Ndiaye et al., 2024).

3 Methodology

This paper uses a structured literature review approach guided by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) to identify and analyse academic publications addressing the cyber security and resilience of AI systems in the nuclear sector. The review focuses on studies that discuss AI systems, adversarial threats, cyber security, resilience, verification and validation (V&V), explainability, or secure lifecycle concerns in nuclear, operational technology, I&C, ICS, or cyber-physical system (CPS) contexts.

Structured searches were conducted in Scopus and Web of Science on 6 June 2026. The same Boolean search expression was used in both databases and entered using the database-specific all-fields search syntax, namely ALL(...) in Scopus and an all-fields search in Web of Science. The search strategy was based on five conceptual blocks: operational technology and industrial control systems; cyber security and resilience; adversarial attacks and data manipulation; AI and machine learning; and the nuclear sector. Synonymous terms within each block were combined with OR, while the five blocks were combined with AND. The exact search string is provided in the footnote[1].

The review was guided by the following research questions:

- RQ1: What types of artificial intelligence systems are discussed in the context of cyber security and resilience in nuclear or nuclear-relevant contexts?
- RQ2: Which cyber threats, adversarial techniques, and attack surfaces affecting artificial intelligence systems are identified in the reviewed literature?
- RQ3: How does the reviewed literature address the resilience, robustness, verification, validation, explainability, and secure lifecycle of artificial intelligence systems in nuclear or nuclear-relevant environments?

Publications were included if they were published in English between 2021 and 2026, were available in full text, and addressed cyber security, resilience, adversarial threats, or secure lifecycle issues related to AI systems in the nuclear sector or in a closely related nuclear-relevant OT, I&C, ICS, or cyber-physical system context. After the Boolean search, records were filtered according to the formal eligibility limits of the review: publication year, language and document type. Eligible Scopus document types were articles, reviews and conference papers; eligible Web of Science document types were articles, proceedings papers and review articles. Additional contextual sources cited in the Introduction and Background were used to frame the review and were not counted as part of the final synthesis unless they met the full inclusion criteria.

The literature review was guided by PRISMA, which structures the review process into identification, screening, eligibility, and inclusion stages. An adapted PRISMA flow diagram was used to document the selection process, including the number of records identified, duplicates removed, records screened, full-text publications assessed, and publications included in the final review, as presented in Fig. 1.

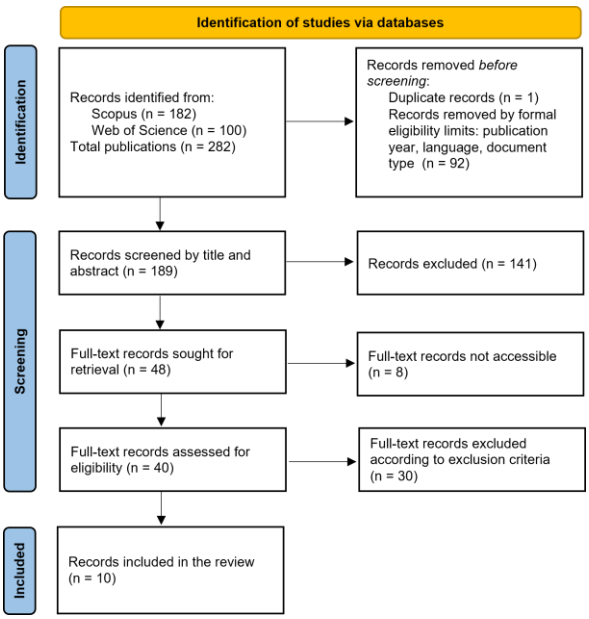


Figure 1. PRISMA flow diagram of the literature selection process

The initial search returned 182 records in Scopus and 100 records in Web of Science, resulting in 282 records in total. Before title and abstract screening, 92

[1]("OT" OR "operational technology" OR "I&C" OR "instrumentation and control" OR "instrumentation and control systems" OR "I&C systems" OR "industrial control system" OR "industrial control systems" OR "ICS" OR "SCADA" OR "distributed control system" OR "DCS" OR "cyber-physical" OR "cyber physical" OR "cyber-physical systems") AND ("cybersecurity" OR "cyber security" OR "computer security" OR "information security" OR "AI security" OR "ML security" OR "security testing" OR "adversarial testing" OR "secure lifecycle" OR "verification and validation" OR "V&V" OR "robustness" OR "resilience" OR "safety and security" OR "functional safety") AND ("adversarial attack" OR "adversarial attacks" OR "adversarial machine learning" OR "aML" OR "data poisoning" OR "poisoning attack" OR "feature poisoning" OR "evasion attack" OR "exploratory attack") AND ("AI" OR "artificial intelligence" OR "machine learning" OR "ML" OR "resilient machine learning" OR "rML" OR "AI model" OR "ML model" OR "large language model" OR "LLM") AND ("nuclear industry" OR "nuclear sector" OR "nuclear power plant" OR "NPP" OR "nuclear facility" OR "nuclear facilities" OR "nuclear reactor" OR "nuclear safety" OR "nuclear security")

records were removed by applying the formal eligibility limits of the review, namely publication year, language and document type. Duplicate checking was performed by comparing DOI where available, and by title, authors and publication year where DOI information was missing or incomplete. One duplicate record was identified. The low number of duplicates was manually validated using DOI and title-author-year checks and appears to reflect database-specific indexing and the narrow combination of AI, adversarial, OT/ICS and nuclear search terms. The remaining 189 records were screened by title and abstract, of which 141 were excluded because they did not address the combined focus of the review.

Title and abstract screening were used as the first relevance filter after the application of formal eligibility limits and duplicate removal. At this stage, publications were excluded when the title and abstract clearly showed that they did not address the combined focus of the review. This included records that focused on general cyber security without AI or ML, general AI without cyber security or resilience relevance, non-nuclear or non-nuclear-relevant domains, purely conventional IT security, general critical infrastructure without OT, I&C, ICS, CPS, or nuclear relevance, or topics unrelated to the research questions. When the title and abstract did not provide enough information to make a reliable exclusion decision, the publication was retained for full-text assessment.

The remaining 48 publications were selected for full-text retrieval. Eight publications were excluded because the full text was not accessible. The remaining 40 full-text publications were assessed for eligibility. After full-text assessment, 30 publications were excluded, resulting in ten publications included in the final synthesis.

Full-text publications were assessed using four linked relevance dimensions:

1. AI/ML relevance,
2. AI system cyber security and resilience relevance,
3. nuclear or nuclear-relevant context, and
4. contribution to the research questions.

AI/ML relevance required that a publication addressed an AI/ML model, AI component, AI-enabled function, autonomous or ML-enabled system, or AI-based detection, diagnosis, decision-support, or control function.

AI system cyber security and resilience relevance required that the cyber security or resilience issue was linked to the AI/ML system itself, to an AI component, or to an AI-enabled function. This included adversarial threats against AI models or data, manipulation of training or real-time process data used by AI systems, model tampering, evasion, poisoning, model robustness, explainability, V&V of AI components, secure AI lifecycle, AI security testing, deployment assurance of AI-enabled systems, or resilience mechanisms that protect or compensate for AI-enabled functions. General nuclear cyber security, cyber risk assessment, incident response, or ICS/SCADA security was not included unless it also provided evidence relevant to the assurance, robustness, validation, lifecycle security, or operational resilience of AI systems in nuclear or nuclear-relevant environments.

Nuclear relevance required either a direct nuclear-sector context, such as nuclear power plants, nuclear I&C systems, nuclear testbeds, reactor simulators, or nuclear cyber-physical systems, or a clearly nuclear-relevant OT, I&C, ICS, or CPS context with explicit transferability to nuclear AI assurance.

Contribution to the research questions required that the publication provided evidence relevant to at least one of the three research questions. The included publications were classified according to the contribution across the four relevance dimensions into one of three groups, namely as: core studies, conceptual core-supporting studies, or supporting studies. This classification was not used as a quality ranking, but as a relevance and contribution classification indicating how directly each publication supported the review focus and research questions.

Core studies were those that met all four dimensions at a high level and addressed AI/ML systems or AI components, cyber security or resilience of such systems, a direct nuclear-sector context, and a direct contribution to at least one research question. These studies treated AI systems, AI components, or ML-enabled nuclear applications as cyber-vulnerable, resilience-critical, or assurance-relevant assets.

Conceptual core-supporting studies had strong AI/ML and cyber security or resilience relevance, but their nuclear contribution was either conceptual, based on a nuclear-relevant use case, or transferable to nuclear AI assurance rather than a direct empirical assessment of AI systems in NPPs.

Supporting studies did not primarily examine the cyber security of AI systems themselves but were retained when they provided nuclear-specific evidence on AI-enabled detection, testbed-based validation, response, recovery, operational resilience, or deployment constraints needed to interpret how AI systems may be assured in nuclear cyber-physical environments, thereby contributing to at least one research question.

Publications were excluded from the final synthesis if they failed to meet the combined focus of the review. This included publications on general AI applications in nuclear engineering without a cyber security or resilience perspective; nuclear cyber security without AI or intelligent systems; general ICS, SCADA, CPS, or critical-infrastructure security without nuclear-specific contribution; AI-based intrusion detection or anomaly detection without implications for AI robustness, validation, explainability, lifecycle assurance, or nuclear deployment; general adversarial machine learning without nuclear or nuclear-relevant applicability; and AI governance, safety, or ethics

without both a cyber security and nuclear-sector connection.

The included publications were analysed using structured data extraction aligned with the three research questions. The extracted fields corresponded to the comparison dimensions in Table 1 and covered each publication's role, AI system or function, nuclear or nuclear-relevant context, threat or assurance focus, relevant attack surface or adversarial technique where applicable, and contribution to the research questions.

Because the final corpus included heterogeneous publication types, including surveys, conceptual frameworks, testbeds, simulation studies, and technical case studies, this assessment was used as a relevance and contribution appraisal rather than as a formal statistical quality or risk-of-bias assessment.

The relatively small final corpus reflects the narrow focus of the review rather than a broad lack of literature on either AI or nuclear cyber security. Many retrieved publications addressed AI for cyber security, general AI applications in the nuclear sector, general ICS or CPS cyber security, or AI assurance without a nuclear cyber security perspective. They were therefore excluded because this review specifically focuses on the cyber security and resilience of AI systems themselves in nuclear or nuclear-relevant environments.

4 Results

The final synthesis included ten publications: four core studies, two conceptual core-supporting studies, and four supporting studies. The included publications were not treated as a homogeneous body of evidence, but were compared according to their role, AI system or function, nuclear or nuclear-relevant context, threat or assurance focus, and contribution to the research questions. Table 1 shows three partially overlapping perspectives: core studies addressing AI systems as cyber-vulnerable or assurance-relevant assets; conceptual core-supporting studies providing transferable threat-modelling and trustworthiness concepts; and supporting studies showing how AI-enabled detection, response, testbeds, and operational resilience are constrained by nuclear-specific environments.

This distinction is important because only the core studies are used as direct evidence on the cyber security or assurance of AI systems themselves, whereas conceptual core-supporting and supporting studies are used more narrowly to interpret transferable threat concepts, deployment constraints, validation environments, response mechanisms, and nuclear-specific assurance needs.

Table 1. Summary of publications included in the final synthesis

Publication	Role	AI system / function	Context	Threat / assurance focus	Review contribution
Yockey et al. (2023)	C	ML digital twins, AutoML, ACS	Advanced reactor / NPP ACS	Data and model attack surfaces	Shows ACS ML components as cyber-relevant assets.
Khadka et al. (2024)	C	ML-enabled nuclear applications	Nuclear industry	Adversarial attacks, resilient ML	Maps attack surfaces, defences, barriers and opportunities for resilient ML.
Park and Koo (2026)	C	AI components in nuclear I&C systems	NPP I&C	V&V, robustness, transparency, software dependencies	Provides assurance considerations for AI components in nuclear I&C systems.
Ndiaye et al. (2024)	C	ML/AI fault detector, XAI	Industrial I&C with explicit nuclear relevance	Secure lifecycle, adversarial testing, XAI	Links adversarial testing and explainability to AI deployment in I&C systems.
Sundaram and Abdel-Khalik (2021)	CCS	Predictive model, covert cognizance	Nuclear reactor modelling	Trustworthiness, manipulation detection	Connects predictive model integrity with nuclear cyber-physical trustworthiness.
Al-Hawawreh et al. (2025)	CCS	ML-based IDS	Closed-loop CPS/ICS with NPP use case	Adversarial examples, evasion	Supports threat modelling for AI-enabled detection in nuclear-relevant CPS.
Yoo et al. (2025)	S	Autoencoder and RL	NPP under FDI attack	Detection, reconstruction, recovery	Shows AI-enabled resilience and recovery under false data injection.
Thiyagarajan and Hammad (2024)	S	Transformer-LSTM detector	Air-gapped NPP ICS	DoS, MITM, white-box detection	Supports deployment-related resilience in constrained nuclear ICS.
Zebrowitz et al. (2026)	S	ML cyber testbed	NPP network testbed	Generalisability, validation, XAI	Demonstrates the need for testbed validation and cross-environment evaluation.
Lee et al. (2021)	S	Hidden Markov model	NPP cyber response	Security state estimation, functional impact	Links cyber state estimation with response-oriented resilience.

Note. C = core study; CCS = conceptual core-supporting study; S = supporting study; ACS = autonomous control system; DoS = denial-of-service; FDI = false data injection; IDS = intrusion detection system; MITM = man-in-the-middle; NPP = nuclear power plant; RL = reinforcement learning; XAI = explainable AI.

4.1 Types of artificial intelligence systems addressed in the literature

With regard to RQ1, the reviewed publications show that AI systems in the nuclear cyber security context are discussed in several different forms. One direct example is the use of ML-based digital twins within autonomous control systems (ACS) for advanced reactors. In this setting, AI systems are not merely auxiliary analytical tools, but part of the control logic that supports malfunction detection, component state classification, forecasting, and control strategy selection (Yockey et al., 2023). This makes the AI component itself a cyber-relevant asset because manipulation of the data pipeline or model behaviour may affect the operation of the autonomous control system.

A second group of publications addresses AI components in nuclear I&C systems. These studies focus less on a specific attack scenario and more on whether AI components can be specified, verified, validated, tested, and accepted within the nuclear software assurance framework. This is important because AI components differ from traditional software due to their complexity, dependence on data, limited transparency, and less predictable behaviour (Park & Koo, 2026).

A third group focuses on AI-enabled cyber security functions, especially anomaly detection, intrusion detection, and cyber attack detection in nuclear or nuclear-relevant cyber-physical environments. Examples include transformer-LSTM models for Modbus TCP traffic in air-gapped nuclear ICS environments, ML-based intrusion detection system (IDS) models for closed-loop CPS environments, and machine-learning-based attack detection models evaluated across virtual and hybrid NPP network testbeds (Al-Hawawreh et al., 2025; Thiyagarajan & Hammad, 2024; Zebrowitz et al., 2026). These studies show how AI-enabled detection may be deployed in constrained nuclear or nuclear-relevant environments, even when their primary focus is AI for cyber security rather than the cyber security of AI itself.

A fourth group addresses intelligent and probabilistic support for cyber response and resilience. This includes self-healing control using autoencoders and reinforcement learning (RL), as well as security state estimation using hidden Markov models. These studies extend the role of AI-enabled and probabilistic methods from detection to response, recovery, and operational decision support (Lee et al., 2021; Yoo et al., 2025). They are therefore relevant to resilience, although they do not primarily analyse AI systems as attack targets.

4.2 Cyber threats, adversarial techniques and attack surfaces

With regard to RQ2, the reviewed literature shows that threats to AI systems in nuclear or nuclear-relevant environments should not be limited to attacks on the final model. The attack surface spans the whole AI lifecycle, including data collection, training and validation data, real-time process data, model artefacts, AI outputs, software dependencies, and the operational deployment environment.

The clearest attack surface is the data pipeline. Yockey et al. (2023) examine attacks against training data and real-time process data in ML-based autonomous control systems. This is particularly important because nuclear AI systems may rely on sensor data, process variables, historical operational data, and real-time communication between digital twins and control systems. In such settings, manipulation of data may be more feasible or more consequential than direct manipulation of the model architecture.

The second attack surface is the AI model and its software environment. This includes attacks against model architectures, model files, model parameters, model behaviour, software libraries, and dependencies. In their ACS case study, Yockey et al. (2023) assessed attacks against real-time data and training data as higher risk than direct attacks against ML model files or architectures, because data manipulation had more direct effects on model performance and ACS behaviour. Nevertheless, model-level attacks remain relevant because model substitution or manipulation could affect malfunction detection, forecasting, classification, or control-strategy selection.

The third attack surface is adversarial input manipulation. Several publications discuss adversarial examples, feature poisoning, evasion attacks, and false data injection (Al-Hawawreh et al., 2025; Khadka et al., 2024; Ndiaye et al., 2024; Yockey et al., 2023). In industrial and nuclear cyber-physical systems, such attacks differ from classical adversarial examples in image classification because the manipulated values must remain plausible within the physical process. Adversarial samples must therefore deceive the AI model while preserving the physical effect of the underlying attack.

The fourth attack surface is the interface between cyber events and physical process behaviour. False data injection (FDI), man-in-the-middle attacks (MITM), denial-of-service attacks, and Modbus function-code attacks are especially relevant in the reviewed literature (Thiyagarajan & Hammad, 2024; Yoo et al., 2025; Zebrowitz et al., 2026). These studies show that cyber attacks against AI-enabled systems in the nuclear sector cannot be understood only as information security events; they must also be evaluated according to their effects on physical process states, operator interpretation, and safety or security functions.

The fifth attack surface is deployment. In restricted or air-gapped nuclear ICS, concerns over third-party black-box AI tools favour white-box approaches, explainability, controlled deployment, and system-owner visibility into the model (Ndiaye et al., 2024; Park & Koo, 2026; Thiyagarajan & Hammad, 2024).

Table 2. Synthesis of attack surfaces and assurance mechanisms

Attack surface	Threat examples	Nuclear-specific concern	Assurance / resilience response
Training and validation data	Poisoning, label manipulation, biased or unrepresentative data	Real NPP data may be limited, sensitive, simulated or difficult to reproduce.	Data provenance, dataset validation, adversarial data testing, configuration control.
Real-time process data	FDI, sensor manipulation, MITM, replay	Manipulated values may remain physically plausible while misleading AI models or operators.	Process-aware validation, redundancy, plausibility checks, state estimation, testbed/HIL validation.
AI model and software artefacts	Model tampering, backdoors, model substitution, vulnerable libraries	AI components may support operationally significant or safety-relevant decisions.	Model integrity checks, dependency management, secure configuration, adversarial robustness testing.
AI outputs and decision support	False negatives, false positives, misleading confidence, opaque outputs	Incorrect or unexplained outputs may affect operator response and plant-state assessment.	XAI, confidence information, human oversight, alarm validation, independent confirmation.
Deployment and validation environment	Domain shift, cross-network degradation, black-box tools, testbed-to-plant transfer gaps	A model validated in one simulator, network or testbed may not generalise to another NPP environment.	Nuclear-specific testbeds, cross-environment evaluation, white-box deployment, lifecycle assurance.

Note. FDI = false data injection; HIL = hardware-in-the-loop; MITM = man-in-the-middle; XAI = explainable AI.

4.3 Resilience, robustness, verification, validation and secure lifecycle

With regard to RQ3, the reviewed literature addresses resilience and assurance through several mechanisms.

One prominent theme is adversarial robustness. Khadka et al. (2024) identify resilient machine learning in the nuclear sector as an emerging field and emphasise that both models and data must be tested against adversarial threats. Al-Hawawreh et al. (2025) similarly stress that ML-based IDS models in closed-loop CPS environments must be resilient to adversarial examples, especially where misclassification may affect physical processes.

A second theme is verification and validation. Park and Koo (2026) propose aligning AI model development with existing nuclear software V&V stages, while adding AI-specific verification activities. Their approach supports the view that model accuracy should be complemented by evidence on robustness, transparency, and development-process reliability.

A third theme is explainability. Ndiaye et al. (2024) discuss XAI in the context of adversarial testing and secure lifecycle considerations, while the FANCY-X testbed study uses explainability-oriented analysis to examine cross-environment degradation of ML-based attack detection models (Zebrowitz et al., 2026). Explainability is important because operators, assessors, and system owners need to understand why an AI model produced an alert and whether its behaviour remains meaningful when transferred across environments.

Sundaram and Abdel-Khalik (2021) add a complementary perspective by focusing on software cognizance and trustworthiness of predictive models. Their covert cognizance approach suggests that resilience may also depend on whether predictive software can maintain an internal record of its execution and detect manipulation without relying only on external logging, perimeter defences, or conventional intrusion detection mechanisms.

A fourth theme is resilience through response and recovery. Yoo et al. (2025) move beyond cyber attack detection by proposing a self-healing strategy using an autoencoder for detection and inferred value reconstruction, and reinforcement learning for system-level control. Lee et al. (2021) similarly support response by estimating the current security state and linking cyber alarm sequences to security and functional impact analysis. These studies suggest that future nuclear AI systems may need to support not only detection but also interpretation, response prioritisation, and recovery.

A fifth theme is nuclear-specific validation. Several studies rely on simulators, hardware-in-the-loop (HIL) testbeds, or nuclear cyber-physical testbeds rather than real NPP operational data. This is understandable because real NPP data are difficult to access, but it also limits generalisability. Studies using nuclear-specific environments, such as NPP network testbeds, nuclear safety shutdown testbeds, or nuclear I&C cyber-physical simulation environments, are particularly relevant because robust AI assurance requires validation environments that capture both cyber behaviour and physical process dynamics.

4.4 Cross-cutting observations

Across the ten included publications, three cross-cutting observations emerge. First, the literature increasingly recognises that AI systems in the nuclear sector create new attack surfaces. These include training datasets, real-time process data, model architectures, model files, AI libraries, deployment pipelines, and AI-enabled decision outputs.

Second, much of the literature still treats AI primarily as a cyber security tool rather than as a cyber-vulnerable asset. Anomaly detection, intrusion detection, and self-healing control are well represented, but fewer publications directly test the robustness of the AI system itself against adaptive adversaries. This distinction is important because an AI-based detector or controller that improves cyber security under normal assumptions may introduce additional risk if its own data, model, or outputs can be manipulated.

Third, nuclear-specific constraints substantially affect how AI cyber security and resilience should be evaluated. These constraints include safety relevance, digital I&C architecture, air-gapping, limited availability of representative datasets, regulatory expectations, V&V requirements, explainability, and the need to distinguish cyber attacks from equipment faults or normal transients. Therefore, cyber security and resilience of AI systems in the nuclear sector should not be evaluated only through conventional ML performance metrics. It should also consider lifecycle assurance, adversarial robustness, physical process validity, explainability, and the potential impact on nuclear safety and security functions.

5 Discussion

The reviewed literature shows that the cyber security and resilience of AI systems in the nuclear sector is an emerging but still fragmented research area. The selected studies approach the topic from different angles, including autonomous control, nuclear I&C assurance, resilient machine learning, adversarial testing, intrusion detection, security state estimation, and cyber-physical testbeds. This diversity reflects the early stage of the field, but it also shows that AI systems are becoming part of the digital ecosystem that must be considered in nuclear cyber security.

A central implication is the need to distinguish between AI used for cyber security and the cyber security of AI systems themselves. Several included studies use AI as a tool for detecting attacks, identifying anomalies, supporting response, or improving resilience. However, an AI system that supports cyber security is not necessarily secure, robust, explainable, or suitable for deployment in a nuclear environment. If an AI-based detector, digital twin, predictive model, or autonomous control component can be manipulated through training data, real-time data, model architecture, adversarial inputs, or deployment conditions, it may introduce additional risk instead of reducing it.

Building on the attack-surface synthesis presented in Table 2, the findings suggest that the cyber security and resilience of AI systems in the nuclear sector should be understood as an assurance problem spanning the AI lifecycle and the nuclear operational environment. As the main analytical contribution of this review, this paper proposes a synthesis-derived assurance agenda structured around five interrelated layers: data and lifecycle assurance, model robustness and integrity, OT and I&C deployment assurance, cyber-physical validation, and operational and regulatory assurance.

The first layer is data and lifecycle assurance. The reviewed literature shows that AI security risks are not limited to the trained model, but also include training data, validation data, real-time input data, data provenance, dataset representativeness, and lifecycle control. This is particularly important in the nuclear sector because real operational data may be limited, sensitive, simulated, or difficult to reproduce. Assurance activities should therefore address the integrity, traceability, representativeness, and protection of data across development, validation, deployment, operation, and maintenance.

The second layer is model robustness and integrity. AI components in nuclear I&C systems cannot be accepted only because they perform well on a test dataset. They must also be evaluated for adversarial robustness, model integrity, transparency, explainability, vulnerable software dependencies, and secure configuration management. This includes threats such as poisoning, evasion, adversarial perturbations, model tampering, backdoors, vulnerable open-source libraries, and model substitution. The implication is that AI models should be treated as cyber-relevant assets whose behaviour, parameters, dependencies, and updates require assurance.

The third layer is OT and I&C deployment assurance. Nuclear systems are characterised by restricted connectivity, long lifecycles, legacy systems, strict configuration control, limited tolerance for unverified change, and strong safety and security expectations. These characteristics may reduce some external attack paths, but they do not eliminate risks arising from portable media and mobile devices, internal networks, maintenance activities, supply chains, test environments, software dependencies, or poorly controlled updates. AI assurance must therefore be integrated with existing nuclear software assurance, cyber security, configuration management, and change-control practices.

The fourth layer is cyber-physical validation. Adversarial threats in nuclear cyber-physical systems differ from adversarial threats in conventional information technology because manipulated values must often remain physically plausible, respect process constraints, and preserve the intended physical effect of the underlying attack. Cyber security evaluation of AI systems in the nuclear sector must therefore account for process dynamics, sensor and actuator behaviour, control logic, safety functions, and the interaction between cyber events and physical plant states. Simulators, HIL environments, cyber-physical testbeds, and NPP network testbeds are therefore valuable and often necessary. However, they also create uncertainty regarding generalisability. The

FANCY-X study is particularly relevant in this respect, because it demonstrates that cyber attack detection performance may degrade when models are transferred between virtual and hybrid NPP network environments. This supports the need to evaluate AI cyber security methods not only for internal accuracy, but also for cross-environment robustness and deployment validity.

The fifth layer is operational and regulatory assurance. If AI is used to support detection, response, recovery, anomaly interpretation, operational decision-making, or autonomous control, then its reliability, failure modes, explainability, and potential for manipulation become relevant to operators, assessors, and regulators. AI outputs may affect operator understanding, response prioritisation, or plant-state assessment. Assurance should therefore include confidence information, XAI, human oversight, alarm validation, procedural integration, and evidence suitable for independent or regulatory review. In this sense, some AI systems may need to be assessed as cyber-relevant digital assets, and, depending on their function and potential consequences of compromise, some may also need to be treated as critical digital assets.

The proposed assurance agenda also clarifies the main gap in the current literature. There is comparatively more work on AI-enabled detection and response support than on the systematic protection of AI systems as cyber-vulnerable assets. Fewer studies directly evaluate poisoning, evasion, model manipulation, secure AI lifecycle, AI model supply chain, adaptive adversaries, or adversarial robustness in nuclear-specific settings. This indicates a gap between the expected future use of AI in advanced reactors, digital I&C, autonomous functions, and cyber-physical monitoring, and the maturity of methods for assuring the cyber security of such systems. Future studies should therefore place greater emphasis on reproducible nuclear-relevant testbeds, realistic process-aware attack scenarios, cross-environment validation, explainable and auditable AI, secure AI lifecycle management, and the relationship between AI model behaviour and nuclear safety and security functions.

These findings should also be interpreted in light of several limitations. The review is limited by its focus on Scopus and Web of Science, English-language publications, selected publication types, and the period from 2021 to 2026. As a result, relevant grey literature, regulatory guidance, technical reports, or earlier work may not have been included. The final corpus is also relatively small, because the review deliberately focused on the cyber security and resilience of AI systems in nuclear or nuclear-relevant environments rather than on the broader use of AI for cyber security. In addition, the classification of publications into core, conceptual core-supporting, and supporting studies involved interpretive judgement, particularly for borderline AI-enabled detection and response studies. The proposed assurance agenda should therefore be understood as an analytical structure derived from the reviewed literature, not as a validated maturity framework.

6 Conclusion

This literature review examined recent academic work on the cyber security and resilience of AI systems in the nuclear sector. Based on the final synthesis of ten publications, the review shows that the field is still emerging and fragmented, but increasingly relevant as AI is considered for cyber security, monitoring, decision support, digital twins, and control-related functions in nuclear environments. The main conclusion is that AI systems should not be treated only as analytical or cyber-defence tools. As they become part of nuclear digital systems, they may also become cyber-relevant assets with their own data, model, software, deployment, and operational attack surfaces.

The contribution of the review is twofold. First, it distinguishes the cyber security of AI systems from the broader literature on AI for cyber security in nuclear environments. Second, it synthesizes the reviewed literature into a nuclear-specific AI security assurance agenda structured around data and lifecycle assurance, model robustness and integrity, OT and I&C deployment assurance, cyber-physical validation, and operational and regulatory assurance. This synthesis indicates that conventional machine-learning performance metrics are not sufficient on their own. AI systems should also be assessed through assurance evidence related to integrity, robustness, explainability, deployment validity, human oversight, and regulatory review.

Important gaps remain in relation to poisoning, evasion, model manipulation, secure AI lifecycle management, AI model supply chains, adaptive adversaries, and validation in realistic nuclear I&C or cyber-physical environments. Future research should therefore focus on nuclear-specific adversarial testing, reproducible testbeds, cross-environment validation, explainable and auditable AI, and the relationship between AI model behaviour, cyber attack scenarios, and nuclear safety and security functions. As AI becomes more closely connected to nuclear digital systems, the cyber security and resilience of AI systems should become an explicit part of nuclear computer security programmes, V&V activities, and regulatory discussions.

References

Al-Hawawreh, M., Baig, Z., & Zeadally, S. (2025). Resilient Intrusion Detection Models for Closed Control-Loop in Cyber-Physical Systems: Combating Adversarial Examples. *IEEE Internet of*

Things Magazine, *8*(1), 73–80. https://doi.org/10.1109/IOTM.001.2400068

Arhouni, F. E., Abdo, M. A. S., Ouakkas, S., Bouhssa, M. L., & Boukhair, A. (2025). Artificial intelligence-driven advances in nuclear technology: Exploring innovations, applications, and future prospects. *Annals of Nuclear Energy*, *213*, 111151. https://doi.org/10.1016/j.anucene.2024.111151

Ayodeji, A., Mohamed, M., Li, L., Di Buono, A., Pierce, I., & Ahmed, H. (2023). Cyber security in the nuclear industry: A closer look at digital control systems, networks and human factors. *Progress in Nuclear Energy*, *161*, 104738. https://doi.org/10.1016/j.pnucene.2023.104738

Blakely, B., Li, Y., Dave, A., Kultgen, D., & Vilim, R. (2025). *AI-Driven Cybersecurity Testbed for Nuclear Infrastructure: Comprehensive Evaluation Using METL Operational Data* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2512.01727

Cancila, D., Daniel, G., Sirven, J.-B., Chihani, Z., Chersi, F., & Vinciguerra, R. (2024). Research Directions on AI and Nuclear. *EPJ Web of Conferences*, *302*, 17005. https://doi.org/10.1051/epjconf/202430217005

Eggers, S. (2021). A novel approach for analyzing the nuclear supply chain cyber-attack surface. *Nuclear Engineering and Technology*, *53*(3), 879–887. https://doi.org/10.1016/j.net.2020.08.021

Huang, Q., Peng, S., Deng, J., Zeng, H., Zhang, Z., Liu, Y., & Yuan, P. (2023). A review of the application of artificial intelligence to nuclear reactors: Where we are and what's next. *Heliyon*, *9*(3), e13883. https://doi.org/10.1016/j.heliyon.2023.e13883

IAEA. (2021a). Computer Security for Nuclear Security (IAEA Nuclear Security Series No. 42-G). International Atomic Energy Agency. https://www-pub.iaea.org/MTCD/Publications/PDF/PUB1918_web.pdf

IAEA. (2021b). Computer Security Techniques for Nuclear Facilities (IAEA Nuclear Security Series No. 17-T (Rev. 1)). International Atomic Energy Agency. https://www-pub.iaea.org/MTCD/Publications/PDF/PUB1921_web.pdf

Khadka, A., Sthapit, S., Epiphaniou, G., & Maple, C. (2024). Resilient Machine Learning: Advancement, Barriers, and Opportunities in the Nuclear Industry. *ACM Computing Surveys*, *56*(9), 1–29. https://doi.org/10.1145/3648608

Lee, C., Ho Chae, Y., & Hyun Seong, P. (2021). Development of a method for estimating security state: Supporting integrated response to cyber-attacks in NPPs. *Annals of Nuclear Energy*, *158*, 108287. https://doi.org/10.1016/j.anucene.2021.108287

Lezzi, M., Martino, L., Damiani, E., & Yeun, C. Y. (2025). A Systematic Literature Review on AI-Based Cybersecurity in Nuclear Power Plants. *Journal of Cybersecurity and Privacy*, *5*(4), 79. https://doi.org/10.3390/jcp5040079

Mendoza, H., Yadav, V., Alam, S. B., Eskins, D., & Iyengar, R. (2026). Advances in digital twins and AI/ML for condition monitoring in nuclear applications. *Frontiers in Energy Research*, *14*, 1716514. https://doi.org/10.3389/fenrg.2026.1716514

Ndiaye, N. G., Kirdan, E., & Waedt, K. (2024). *Cybersecurity Testing for Industry 4.0: Enhancing Deployments in operational I&C systems Through Adversarial Testing and Explainable AI*. https://doi.org/10.18420/INF2024_169

Park, J., & Koo, S. (2026). Reliability verification methods for artificial intelligence components in instrumentation and control systems at nuclear power plants. *Nuclear Engineering and Technology*, *58*(4), 104096. https://doi.org/10.1016/j.net.2025.104096

Sundaram, A., & Abdel-Khalik, H. (2021). Covert Cognizance: A Novel Predictive Modeling Paradigm. *Nuclear Technology*, *207*(8), 1163–1181. https://doi.org/10.1080/00295450.2020.1812349

Thiyagarajan, K., & Hammad, I. (2024). Anomaly Detection in Air-Gapped Industrial Control Systems of Nuclear Power Plants. *2024 Cyber Awareness and Research Symposium (CARS)*, 1–6. https://doi.org/10.1109/CARS61786.2024.10778886

Yockey, P., Erickson, A., & Spirito, C. (2023). Cyber threat assessment of machine learning driven autonomous control systems of nuclear power plants. *Progress in Nuclear Energy*, *166*, 104960. https://doi.org/10.1016/j.pnucene.2023.104960

Yoo, S., Mohler, G., & Zhang, F. (2025). Self-Healing Control of Nuclear Power Plants Under False Data Injection Attacks. *Nuclear Science and Engineering*, *199*(1), 162–175. https://doi.org/10.1080/00295639.2024.2372520

Zebrowitz, M., Doran, M. A., & Zhang, F. (2026). FANCY-X: A nuclear power plant network testbed for cross-network cybersecurity evaluation and generalizability. *Progress in Nuclear Energy*, *198*, 106444. https://doi.org/10.1016/j.pnucene.2026.106444

A Pulse-Based Semiconductor Reverse Breakdown Voltage Meter Using Inductive Flyback Energy Transfer

Andrej Brnušák, Peter Kolok, Marek Tebeľák, Ján Kapitulík
University of Žilina, Faculty of Management Science and Informatics
Department of Technical Cybernetics
Univerzitná 8215/1, 010 26 Žilina
brnusak@stud.uniza.sk, peter.kolok@fri.uniza.sk,
tebelak2@stud.uniza.sk, jan.kapitulik@fri.uniza.sk

Abstract. Reliable testing of semiconductor protection devices is important for embedded power electronics. This paper presents a compact pulse-based platform for measuring reverse breakdown and dynamic clamping behaviour using inductive flyback energy transfer. A controlled inductor current generates short high-voltage pulses through MOSFET switching, while an analog peak detector and embedded ADC capture the clamped voltage. The prototype was evaluated using SMAJ48CA, P6KE68CA, and 1.5KE170CA TVS diodes. Measurements agreed with expected datasheet ranges, and the highest verified voltage reached 177 V. The proposed method enables low-cost testing with reduced continuous thermal stress.

Keywords. reverse breakdown voltage, TVS diode, inductive flyback, pulse-based measurement

1 Introduction

Reliable verification of semiconductor protection devices is important in electronics manufacturing, diagnostics, and embedded power systems. Among their electrical parameters, the reverse breakdown voltage V_{BR} represents a key indicator of the functional state and operating limits of semiconductor junctions (Baliga, 2019). This parameter is particularly relevant for Transient Voltage Suppression (TVS) and Zener diodes, where the ability to enter reverse breakdown and clamp overvoltage transients directly determines the protection capability of the device (Littelfuse, 2020, 2025; Vishay General Semiconductor, 2026).

In practical diagnostics, catastrophic failures can often be detected as permanent short or open circuits. Partial degradation is more difficult to identify because a protection device may remain conductive and apparently functional while its breakdown or clamping characteristics have shifted outside the expected range. Experimental verification of reverse-bias behaviour is therefore useful for component screening, quality control, and evaluation of devices exposed to electrical overstress.

Conventional breakdown-voltage characterization is typically based on direct-current (DC) excitation, where a controlled current is applied to the device under test (DUT) and the corresponding voltage is measured. Although simple, continuous current injection causes power dissipation and local junction heating. Since semiconductor breakdown voltage is temperature-dependent, self-heating can introduce measurement drift and, under prolonged stress, may further degrade the tested component (Baliga, 2019).

Professional semiconductor characterization systems address this problem using short-pulse measurements. Source Measure Units (SMUs), for example, can limit the duration of electrical stress and thereby reduce junction self-heating during high-voltage or high-current characterization (Keithley Instruments, 2016). However, instruments combining accurate current control, fast pulse generation, and high-voltage measurement are relatively expensive and may therefore be impractical for educational laboratories, small-scale production environments, repair diagnostics, or portable embedded applications.

Existing approaches therefore exhibit a trade-off between simplicity, thermal stress, and instrumentation cost. Static DC methods are inexpensive but expose the DUT to continuous dissipation, whereas professional pulsed instruments provide improved control at significantly higher cost and complexity. Furthermore, conventional implementations generally rely on dedicated high-voltage sources and measurement equipment. This creates an opportunity for a compact system in which the excitation energy itself is inherently limited and generated only during a short measurement interval.

Low-cost embedded measurement platforms have been successfully applied in various sensing and monitoring applications, including road-traffic monitoring using MEMS sensors (Hodoň et al., 2021) and energy-efficient wireless data acquisition involving mains-voltage measurements (Karpiš et al., 2022), demonstrating the practical potential of compact

microcontroller-based measurement systems.

Inductive energy storage provides a suitable principle for such excitation. When current through an inductor is rapidly interrupted, the stored magnetic energy produces a flyback voltage transient whose amplitude is determined by the circuit conditions and connected load (Mohan et al., 2003). If a semiconductor protection device is connected to the switching node, the generated transient can force it into reverse breakdown while the delivered energy remains limited by the inductor current and switching duration. The resulting voltage can subsequently be captured using a peak detector, which converts the short transient into a stable voltage suitable for acquisition by a low-speed embedded Analog-to-Digital Converter (ADC) (Mancini, 2002).

Based on this principle, this paper presents a compact pulse-based semiconductor reverse-voltage measurement platform that combines inductive flyback excitation, Metal-Oxide-Semiconductor Field-Effect Transistor (MOSFET) switching, analog peak detection, and microcontroller-based acquisition. Before each measurement, a controlled current is established in the inductor. During a short measurement interval, the MOSFET periodically interrupts the current path and produces high-voltage transient pulses that are clamped by the connected DUT. The clamped peak voltage is attenuated, captured, and digitized by the embedded control system.

The main contribution of the proposed approach is a low-cost architecture for measuring pulse-induced reverse breakdown and dynamic clamping behaviour without requiring a continuously operating high-voltage source or laboratory-grade SMU. In comparison with static DC testing, the pulsed excitation reduces continuous thermal loading of the DUT, while the peak-hold front-end enables transient-voltage acquisition using standard embedded hardware. The developed prototype is experimentally evaluated using several TVS diode families with different nominal voltage ratings to assess measurement repeatability, agreement with datasheet breakdown ranges, and behaviour at increasing excitation currents.

The remainder of this paper is organized as follows. Section 2 describes the proposed measurement platform, including the system architecture, inductive pulse generation, analog signal conditioning, embedded control, and experimental verification. Section 3 presents and discusses the experimental results obtained for the tested TVS diodes. Finally, Section 4 summarizes the main findings and outlines directions for future development.

2 Materials and Methods

The proposed system was designed as a compact cyber-physical diagnostic platform for pulse-based reverse breakdown voltage measurement of semiconductor protection devices. The physical layer of the system generates controlled high-voltage transient pulses using inductive energy storage, while the embedded control layer configures the test conditions, drives the switching sequence, acquires the analog response, and presents the measured value to the operator. This architecture enables low-cost and repeatable verification of semiconductor devices without exposing the DUT to continuous thermal stress.

2.1 System Architecture

The measurement system consists of five main functional blocks: a digitally controlled excitation stage, an inductive energy storage loop, a MOSFET-based switching stage, an analog signal-conditioning circuit, and a microcontroller-based control unit. The general topology of the proposed system is shown in Fig. 1. The excitation stage establishes a predefined current through the inductor. This current is set indirectly by a controlled voltage source whose output is adjusted by an external Digital-to-Analog Converter (DAC). The DAC value is selected by the operator through the local user interface before the measurement sequence is initiated.

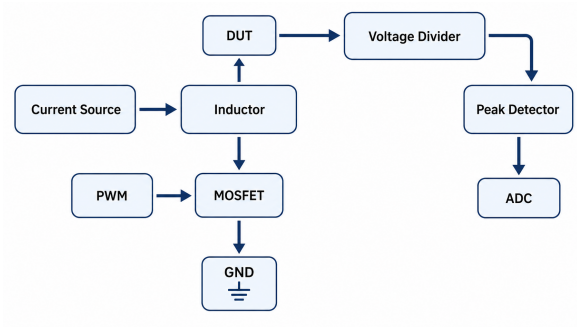


Figure 1. Functional block diagram of the proposed pulse-based reverse breakdown voltage measurement system.

The controlled voltage source is connected in series with a current-limiting resistor network and a power inductor. Under steady-state excitation, the target inductor current can be approximated as

$$I_{\mathrm{exc}} \approx \frac{V_{\mathrm{drv}}}{R_{\mathrm{eq}}}, \tag{1}$$

where I_{exc} is the excitation current, V_{drv} is the drive voltage generated by the controlled source, and R_{eq} is the equivalent resistance of the current-limiting resistor network. As expressed in Eq. (1), the excitation current is determined primarily by the applied drive voltage and the equivalent resistance of the excitation path. Therefore, the measurement energy can be controlled by adjusting the DAC-controlled drive voltage. In the implemented prototype, the operator selects the target current through the embedded interface, and the microcontroller configures the corresponding DAC output

value.

This configuration provides a simple means of controlling the initial energy stored in the inductor before the switching transient is generated. Since the excitation current directly influences the magnetic energy available in the inductor, adjusting the drive voltage also changes the energy released during the flyback event. In this way, the system can adapt the excitation level to semiconductor devices with different reverse-breakdown voltage ranges while avoiding unnecessarily high current settings. The digitally controlled excitation stage therefore provides both repeatability and flexibility during the measurement process, as the same hardware can be used for multiple DUT types without manual modification of the resistor-inductor path.

The physical implementation of the developed prototype is shown in Fig. 2.

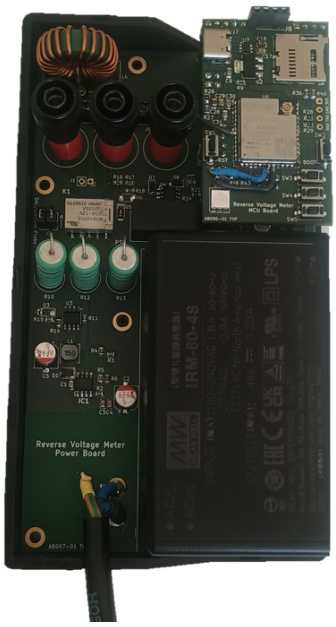


Figure 2. Implemented hardware prototype of the proposed pulse-based semiconductor reverse breakdown voltage meter.

2.2 Inductive Pulse Generation

The core measurement principle is based on inductive flyback energy transfer. During the charging phase, the MOSFET switch is turned on by a Pulse-Width Modulation (PWM) signal generated by the microcontroller. The inductor current flows through the MOSFET to ground, and magnetic energy is stored in the inductor. The stored energy is given by

$$E_L = \frac{1}{2} L I_{\text{exc}}^2, \tag{2}$$

where E_L is the magnetic energy stored in the inductor and L is the inductance. As shown in Eq. (2), the stored magnetic energy depends quadratically on the excitation current. Therefore, the selected test current directly determines the amount of energy available during the transient event and consequently the electrical stress applied to the DUT.

When the MOSFET is rapidly switched off, the inductor current cannot change instantaneously. As a result, a high-voltage transient is generated at the switching node. The induced voltage follows the basic inductor relation

$$v_L(t) = L \frac{di_L(t)}{dt}, \tag{3}$$

where $v_L(t)$ is the instantaneous inductor voltage and $di_L(t)/dt$ is the rate of change of the inductor current. According to Eq. (3), a rapid change in the inductor current produces a corresponding transient voltage across the inductor. The DUT, represented by a TVS diode, Zener diode, or another semiconductor junction, is connected to the switching node. When the generated flyback voltage exceeds the reverse breakdown threshold of the DUT, the semiconductor enters its breakdown region and clamps the transient voltage. The measured peak voltage therefore represents the pulse-induced reverse breakdown and dynamic clamping response of the tested component, depending on the selected excitation current and stored inductive energy.

2.3 Analog Signal Conditioning and Voltage Measurement

Because the generated transient voltage can significantly exceed the safe input range of the microcontroller, the signal is first attenuated using a high-impedance voltage divider. The relationship between the peak reverse voltage at the DUT and the voltage at the ADC input is expressed as

$$V_{\text{ADC}} = V_{\text{DUT,peak}} \frac{R_2}{R_1 + R_2}, \tag{4}$$

where V_{ADC} is the attenuated voltage applied to the analog input, $V_{\text{DUT,peak}}$ is the peak reverse voltage across the DUT, and R_1 and R_2 are the upper and lower resistors of the voltage divider. Equation (4) describes the attenuation introduced by the voltage divider before the transient signal reaches the analog acquisition stage.

The peak reverse voltage at the DUT can therefore be reconstructed as

$$V_{\text{DUT,peak}} \approx V_{\text{ADC}} \frac{R_1 + R_2}{R_2}. \tag{5}$$

Using Eq. (5), the original peak voltage across the DUT can be estimated from the measured ADC input voltage and the known voltage-divider ratio.

To allow the microcontroller to acquire the transient peak reliably, the attenuated pulse is processed by an analog peak detector. The peak detector captures the maximum amplitude of the pulse train and stores it on

a capacitor, creating a quasi-static voltage suitable for ADC conversion. This approach avoids the need for high-speed digitization of the transient waveform and enables the use of a low-cost embedded acquisition platform.

The ADC output code is converted to voltage according to

$$V_{\mathrm{ADC}} = \frac{N_{\mathrm{ADC}}}{2^n - 1} V_{\mathrm{ref}}, \tag{6}$$

where N_{ADC} is the digital ADC value, n is the ADC resolution, and V_{ref} is the ADC reference voltage. According to Eq. (6), the measured ADC code is converted into the corresponding analog voltage using the ADC resolution and reference voltage.

In the practical implementation, the voltage-divider ratio and ADC conversion characteristics are calibrated experimentally to compensate for component tolerances and analog front-end deviations. The calibrated peak reverse voltage can be expressed as

$$V_{\mathrm{DUT,peak,cal}} = aV_{\mathrm{ADC}} + b, \tag{7}$$

where a and b are calibration coefficients obtained from reference measurements. Equation (7) represents the experimentally determined calibration relationship used to convert the acquired ADC voltage into the final peak reverse-voltage estimate.

2.4 Embedded Control and Measurement Sequence

The embedded control unit coordinates the entire measurement process. The system is built around an ESP32-S3 microcontroller, which provides the required digital control signals, ADC acquisition, user-interface handling, and potential wireless communication capabilities.

The local user interface used for selecting the excitation current and initiating the measurement sequence is shown in Fig. 3.

After the user confirms the selected test configuration, the microcontroller performs the following sequence:

1. the selected current setting is converted into a DAC output value;
2. the excitation stage establishes the target current through the resistor-inductor network;
3. the PWM signal activates the MOSFET switching stage for a predefined short measurement window;
4. inductive flyback pulses are generated and clamped by the DUT;
5. the peak detector captures the maximum transient voltage;

Figure 3. Local user interface used for excitation current selection and manual initiation of the measurement sequence.

6. the ADC samples the stabilized peak-detector output;
7. the calibrated peak reverse voltage is calculated and displayed to the operator.

The measurement is intentionally restricted to a short active time window in order to reduce the accumulated energy delivered to the DUT and limit junction self-heating. The energy dissipated in the DUT during a pulse can be generally described as

$$E_{\mathrm{DUT}} = \int_0^{t_p} v_{\mathrm{DUT}}(t) i_{\mathrm{DUT}}(t)\, dt, \tag{8}$$

where t_p is the pulse duration, $v_{\mathrm{DUT}}(t)$ is the instantaneous voltage across the DUT, and $i_{\mathrm{DUT}}(t)$ is the instantaneous current through the DUT. As expressed in Eq. (8), the total energy delivered to the DUT depends on the instantaneous voltage and current over the duration of the pulse. Limiting the active measurement interval therefore reduces the accumulated energy compared with continuous DC excitation and helps suppress junction self-heating and associated thermal distortion during breakdown-voltage measurement.

2.5 Experimental Verification

The prototype was experimentally evaluated using selected semiconductor protection devices with different nominal breakdown voltage ranges. The tested components included TVS diodes from the SMAJ, P6KE, and 1.5KE series. These devices were selected because they represent common overvoltage protection components used in embedded electronics, power interfaces, and energy-related systems.

Each DUT was connected to the measurement terminals and tested under a selected excitation current. The excitation-current values were determined experimentally during preliminary measurements. For each

device, the lowest practical excitation current capable of storing sufficient energy in the inductor to generate a stable and measurable reverse-breakdown response was preferred. Preliminary tests showed that substantially higher excitation currents, approximately 500 mA to 1 A, produced stronger voltage transients but also increased the energy delivered during the switching event and could introduce additional junction heating. Therefore, excessive excitation currents were avoided unless a higher stored energy was required to reach the breakdown region of a higher-voltage device.

It should be emphasized that the excitation current represents the current established in the inductor before the switching transient and does not correspond directly to the instantaneous current flowing through the DUT during breakdown. During the flyback event, the DUT current is transient and depends on the stored magnetic energy, switching dynamics, inductance, and the nonlinear current-voltage characteristic of the tested semiconductor. Consequently, the selected excitation currents should be interpreted as operating parameters of the pulse-generation stage rather than as conventional DC test currents specified in semiconductor datasheets.

The tested excitation-current levels were configured through the local interface and reached up to 300 mA. The measured peak reverse voltage was obtained from the ADC value after voltage-divider scaling, peak detection, and calibration. The experimental verification focused on three main aspects:

- whether the prototype can safely force the DUT into the reverse breakdown region;
- whether the low-energy measurement produces a voltage consistent with the expected datasheet breakdown range;
- whether increasing the excitation current produces a physically consistent increase in the measured clamping voltage.

The excitation-current sweep was performed for the SMAJ48CA as a representative device. Its purpose was not to obtain a complete current-voltage characteristic for every tested component, but to verify that the proposed measurement platform responds consistently to changes in the energy stored in the inductor. Increasing the excitation current increases the stored magnetic energy and shifts the transient operating point of the TVS diode further into its breakdown region. Since the clamping voltage of a TVS diode depends on the transient current, a corresponding increase in the measured peak voltage is expected. The remaining devices were primarily used to verify the ability of the prototype to identify the reverse-breakdown region across different nominal voltage classes.

The measurement chain was calibrated and verified using a precision laboratory digital multimeter as a reference instrument. For each tested component, the prototype measurement was compared with the expected breakdown voltage range specified in the corresponding manufacturer documentation. The resulting measurements were summarized in a comparative table and used to evaluate the functional behaviour and basic consistency of the developed system.

The presented experimental verification should therefore be interpreted as a functional validation of the developed prototype and its measurement chain rather than as a complete metrological characterization. Detailed uncertainty analysis, long-term repeatability testing, and waveform-level validation will be addressed in future work.

2.6 Relevance to Cyber-Physical and Intelligent Energy Systems

From a cyber-physical systems perspective, the developed prototype integrates physical transient generation, embedded control, analog measurement, and human-machine interaction into a single diagnostic node. The physical process is represented by inductive energy storage and semiconductor breakdown behaviour, while the cyber layer performs configuration, timing control, signal acquisition, calibration, and result interpretation.

The proposed system is particularly relevant for intelligent energy and embedded power applications, where semiconductor protection devices are used to protect power converters, renewable energy interfaces, battery-powered nodes, and sensor systems against transient overvoltage events. By enabling low-cost pulse-based testing of these protection devices, the developed platform supports preventive diagnostics, component screening, and reliability assessment in energy-aware electronic systems.

3 Results

The experimental evaluation was performed to verify the capability of the proposed pulse-based measurement platform to induce and measure the reverse breakdown and dynamic clamping response of selected semiconductor protection devices. The tested components covered nominal breakdown voltage ranges from approximately 48 V up to 170 V. The obtained results are summarized in Table 1.

For the SMAJ48CA device tested at an excitation current of 50 mA, the measured voltage reached 56 V, which lies within the nominal breakdown range specified by the manufacturer (Littelfuse, 2025). The P6KE68CA device produced a measured voltage of 67 V, which is also consistent with its specified range of 64.60–71.40 V (Littelfuse, 2020). Similarly, the 1.5KE170CA device reached 177 V, remaining within the expected range of 162–179 V (Vishay General Semiconductor, 2026). These results indicate that the

Table 1. Experimental results obtained with the proposed reverse breakdown voltage meter.

Component type	Nominal V_{BR} range (V)	Excitation current (mA)	Measured voltage (V)
SMAJ48CA	53.30–58.90	50	56
SMAJ48CA	53.30–58.90	150	65
SMAJ48CA	53.30–58.90	250	73
P6KE68CA	64.60–71.40	50	67
1.5KE170CA	162–179	300	177

proposed platform can identify the reverse breakdown region of TVS protection devices across different voltage classes.

The current-sweep measurements of the SMAJ48CA device show a clear dependence of the measured clamping voltage on the selected excitation current. Increasing the excitation current from 50 mA to 150 mA increased the measured voltage from 56 V to 65 V, while a further increase to 250 mA resulted in 73 V. This monotonic behaviour is consistent with the dynamic clamping characteristic of TVS diodes, where higher transient energy shifts the operating point further into the breakdown region and increases the observed clamping voltage.

It should be noted that the current values reported in Table 1 represent the target excitation current established in the inductor path before the switching transient. They do not correspond to a static DC current continuously forced through the DUT. During the flyback event, the actual DUT current is transient and depends on the stored magnetic energy, switching dynamics, inductance, and semiconductor clamping behaviour.

The highest measured voltage was 177 V for the 1.5KE170CA device, demonstrating that the analog front-end, voltage divider, peak detector, and ADC acquisition chain can process high-voltage transient events within the tested operating range. For the SMAJ48CA device, increasing the excitation current from 50 mA to 250 mA increased the measured voltage by 17 V, corresponding to approximately 30.4% relative to the 50 mA measurement. This result confirms that the measurement chain responds consistently to changes in excitation energy.

Overall, the results demonstrate that the proposed pulse-based reverse breakdown voltage meter provides physically consistent measurements for different TVS diode families. The system enables rapid reverse breakdown verification without continuous DC stress and supports its intended use as a low-cost embedded diagnostic platform for cyber-physical measurement systems, power electronics, and intelligent energy applications.

4 Conclusion

This paper presented a compact pulse-based semiconductor reverse-voltage measurement platform designed as an embedded cyber-physical diagnostic system for characterizing pulse-induced breakdown and clamping behaviour. The proposed system combines inductive flyback energy transfer, MOSFET-based switching, analog peak detection, voltage scaling, and microcontroller-based control into a standalone measurement device. Unlike conventional continuous DC testing methods, the presented approach applies short transient pulses to the DUT, thereby reducing the accumulated thermal stress and limiting the risk of measurement distortion caused by junction self-heating.

The experimental verification demonstrated that the developed prototype is capable of characterizing the pulse-induced reverse breakdown and dynamic clamping response of different TVS protection devices across a broad voltage range. The measured values for the SMAJ48CA, P6KE68CA, and 1.5KE170CA components were consistent with the expected breakdown and clamping behaviour specified by the corresponding manufacturer documentation. The highest measured voltage reached 177 V, confirming that the proposed analog front-end and voltage-scaling stage can process high-voltage transient events while maintaining compatibility with embedded ADC acquisition.

The results also showed that increasing the excitation current leads to a physically consistent increase in the measured clamping voltage. This behaviour was observed particularly for the SMAJ48CA device, where the measured voltage increased from 56 V to 73 V as the excitation current was raised from 50 mA to 250 mA. This confirms that the proposed measurement method does not only detect a static breakdown threshold, but also enables the evaluation of dynamic clamping behaviour under controlled pulse-energy conditions.

From the perspective of cyber-physical systems and intelligent energy applications, the developed prototype demonstrates the feasibility of integrating physical high-voltage transient generation with embedded control, signal acquisition, and local diagnostic interpretation. Such a solution is relevant for testing protection components used in power electronics, renewable energy interfaces, battery-powered sensor nodes, and other energy-related embedded systems where overvoltage protection reliability is critical.

Future work will focus on extending the voltage range of the system, improving the calibration procedure, implementing automatic current sweeping, and adding wireless telemetry for remote monitoring and data logging. These enhancements would allow the proposed platform to evolve from a standalone diagnostic prototype into a more flexible intelligent measurement node suitable for laboratory, educational, and field-deployment scenarios.

References

Baliga, B. J. (2019). *Fundamentals of power semiconductor devices*. Springer International Publishing. https://doi.org/10.1007/978-3-319-93988-9

Hodoň, M., Karpiš, O., Ševčík, P., & Kociánová, A. (2021). Which digital-output MEMS magnetometer meets the requirements of modern road traffic survey? *Sensors*, *21*(1), 266. https://doi.org/10.3390/s21010266

Karpiš, O., Olešnaníková, V., Formanek, L., Hodoň, M., Čechovič, L., Šarafín, P., & Ševčík, P. (2022). Compressed sensing – a way to spare energy in WSN for UAV. *IFAC-PapersOnLine*, *55*(4), 170–176. https://doi.org/10.1016/j.ifacol.2022.06.028

Keithley Instruments. (2016). *Optimizing pulsed semiconductor measurements with the model 2461 high current SourceMeter SMU instrument* [Application Note]. Tektronix. Beaverton, OR, USA.

Littelfuse. (2020). *P6KE series TVS diodes datasheet* [Product Data Sheet]. Littelfuse. Chicago, IL, USA.

Littelfuse. (2025). *SMAJ series surface mount TVS diodes specifications* [Product Data Sheet]. Littelfuse. Chicago, IL, USA.

Mancini, R. (Ed.). (2002). *Op amps for everyone* [Design Reference Document No. SLOD006B]. Texas Instruments. Dallas, TX, USA.

Mohan, N., Undeland, T. M., & Robbins, W. P. (2003). *Power electronics: Converters, applications, and design* (3rd ed.). John Wiley & Sons.

Vishay General Semiconductor. (2026). *1.5KE series TransZorb transient voltage suppressors datasheet* [Product Data Sheet]. Vishay Intertechnology. Malvern, PA, USA.

Predictive Control Algorithm for Pneumatic Positioning Cyber-Physical Systems Using Standard ON/OFF Valves

Marek Tebeľák, Lukáš Čechovič, Ondrej Karpiš, Ján Kapitulík
Department of Technical Cybernetics
Faculty of Management Science and Informatics, University of Žilina
Univerzitná 8215/1, 010 26 Žilina, Slovakia
{marek.tebelak,lukas.cechovic,ondrej.karpis,jan.kapitulik}@fri.uniza.sk

Abstract. *Precise positioning of pneumatic linear actuators is challenging due to nonlinear pneumatic dynamics and the discrete nature of standard ON/OFF valves. This paper presents a low-cost positioning approach in which an ESP32 controller uses an empirically identified pulse-duration–displacement model to estimate the initial valve actuation pulse, followed by position-feedback corrections. Experimental validation at 2.5 bar using identical sets of 50 target positions yielded mean absolute endpoint errors of 3.44 mm for the Hafner configuration and 1.50 mm for the Pneumax configuration.*

Keywords. Pneumatic Positioning, ON/OFF Valves, Linear Regression, ESP32, Cyber-Physical Systems

1 Introduction

In the field of industrial automation, pneumatic linear actuators are widely implemented due to their ability to provide high power-to-weight ratios, fast response, and relatively low cost, making them suitable for a broad range of industrial applications (Tootchi & Chaibakhsh, 2025; Xavier et al., 2022). However, their application in precise positioning tasks is significantly complicated by the inherent compressibility of air and nonlinear flow dynamics, which introduce nonlinear, delayed, and state-dependent responses in the system (Qi, Bone, & Zhang, 2019; Xavier et al., 2022). These effects, combined with friction and uncertain load conditions, make accurate trajectory tracking and positioning a challenging control problem (Tootchi & Chaibakhsh, 2025). High-accuracy pneumatic positioning is therefore commonly addressed using proportional or servo valves capable of continuous flow modulation and fine control of the pneumatic actuation process (Šitum & Ćorić, 2022; Zhang et al., 2022).

In cost-sensitive industrial applications, such as the development of an automated bow testing machine, the cost and implementation complexity of proportional valves may represent a significant disadvantage. Consequently, standard 5/3-way discrete ON/OFF valves provide an attractive alternative. These valves are used in digital fluid power systems but are inherently constrained by their discrete switching nature and lack of continuous control authority (Gao et al., 2023). Their operation is characterized by switching delays, finite response times, and the delivery of compressed air through discrete valve states rather than continuously adjustable flow, which can affect positioning performance (Gao et al., 2023). These characteristics make precise endpoint positioning with conventional ON/OFF actuation challenging.

To address such limitations, research has investigated control strategies that account for the nonlinear and hybrid behaviour of pneumatic systems. In particular, model-based and predictive control approaches have demonstrated the potential to improve positioning performance when discrete pneumatic actuation is used (Qi, Bone, & Zhang, 2019; Tootchi & Chaibakhsh, 2025). These approaches incorporate information about system behaviour into the control process to reduce the influence of pneumatic nonlinearities, valve dynamics, and disturbances.

In this context, this paper investigates a computationally lightweight positioning approach for a pneumatic linear actuator controlled by standard ON/OFF valves. The proposed cyber-physical system combines an ESP32-based controller with an empirically identified pulse-duration–displacement model and position-feedback corrections. Separate models are identified for both directions of motion and for the tested valve configurations. The approach is experimentally evaluated under controlled operating conditions with emphasis on calibration-model behaviour and endpoint positioning accuracy.

2 Related Work

The control of pneumatic positioning systems has been investigated using several control architectures. One common approach relies on servo-pneumatic systems equipped with proportional or servo valves. These systems provide continuous airflow regulation and are therefore suitable for applications requiring smooth motion and accurate positioning (Gao et al., 2023;

Mandali & Dong, 2022). Their use, however, is generally associated with higher component cost and increased implementation complexity compared with conventional discrete pneumatic hardware (Sciatti et al., 2024).

An alternative research direction uses discrete ON/OFF valves as part of digital fluid power systems. In such systems, high-frequency switching techniques such as Pulse Width Modulation (PWM) can be used to approximate continuous airflow regulation (De Carvalho et al., 2023). The achievable performance nevertheless depends on valve characteristics including hysteresis, finite switching times, and nonlinear dynamic behaviour (Gao et al., 2023; Li et al., 2023). These non-idealities constrain the useful switching frequency and can introduce deviations between the commanded and resulting pneumatic response.

More advanced control strategies have also been investigated to compensate for nonlinearities and uncertainties in pneumatic systems, including sliding mode control, feedback linearization, and model-based approaches (Duong et al., 2023; Xavier, Fleming, & Yong, 2022). Sliding mode control can provide robustness against parameter variations and external disturbances, although its implementation requires appropriate system modelling and controller tuning (Xavier, Fleming, & Yong, 2022). Model-based approaches may additionally require detailed knowledge of the pneumatic process, increasing modelling and computational requirements for embedded implementations.

These studies demonstrate that improved pneumatic positioning can be achieved either by increasing the capabilities of the pneumatic hardware or by introducing more sophisticated control strategies. For low-cost embedded applications using conventional industrial ON/OFF valves, however, there remains motivation for control methods that require only limited model complexity while explicitly accounting for discrete valve actuation.

The approach investigated in this paper uses an empirically identified pulse-duration–displacement relationship rather than a detailed thermodynamic model or continuous-flow approximation. Separate models are identified for both directions of motion and for the tested valve configurations. The model provides the pulse-duration estimate, while position feedback is used to evaluate the remaining displacement error and perform corrective movements. This structure is intended to maintain low computational complexity while adapting the pulse calculation to the characteristics of the tested pneumatic configuration.

3 Materials and Methods

This section describes the experimental setup, actuator force analysis, embedded control architecture, calibration procedure, pulse-duration–displacement model, and feedback-based positioning method.

3.1 Hardware Architecture

To emphasize the cost-effectiveness of the design, the cyber-physical system was constructed entirely using standard, off-the-shelf industrial parts. The primary actuator driving the mechanical guide is a standard double-acting pneumatic cylinder manufactured by Hafner (type HIF 80/800), featuring a piston diameter of 80 mm and a piston rod diameter of 25 mm.

The central control unit of the system is an ESP32 microcontroller, selected for its 32-bit architecture and hardware timers. However, because the ESP32 operates at a 3.3 V logic level and cannot directly drive the power-hungry 24 V DC industrial solenoid valves, an intermediary power switching interface is required. For this purpose, a dual N-channel MOSFET switching module (utilizing AOD4184A transistors) was integrated. This module safely handles the higher current demands and acts as a robust voltage-level translator between the microcontroller and the pneumatic hardware.

Through this switching module, the ESP32 discrete digital outputs govern the airflow and directional control via standard 5/3-way solenoid valves with closed center positions (such as the Hafner MH 531 121 or Pneumax alternatives). Finally, real-time positional feedback is captured and transmitted back to the microcontroller via a high-resolution (0.005 mm) optical linear encoder.

3.2 Actuator Force Analysis

For a double-acting pneumatic cylinder, the extension and retraction forces depend on the applied pressure and the corresponding effective piston areas. The practical actuator forces are estimated as

$$F_{\text{push}} = p \cdot A_{\text{push}} \cdot \eta \tag{1}$$

$$F_{\text{pull}} = p \cdot (A_{\text{push}} - A_{\text{rod}}) \cdot \eta \tag{2}$$

where p is the supply pressure, A_{push} is the full piston area, A_{rod} is the piston-rod cross-sectional area, and η is a practical force correction coefficient. According to the cylinder manufacturer's guidance, approximately 5% of the theoretical cylinder force may be deducted as a rule-of-thumb estimate of frictional losses (Hafner-Pneumatik Krämer GmbH & Co. KG, 2024). Therefore, $\eta = 0.95$ was adopted in this study. This coefficient was not identified experimentally and is used only for estimating the practical actuator force; it does not represent a dynamic model of the pneumatic system.

The parameters of the experimental testbed used for the force calculation are summarized in Table 1.

Table 1. Main testbed parameters used for the actuator force calculation.

Parameter	Symbol	Value
Cylinder bore diameter	D	80 mm
Piston-rod diameter	d_{rod}	25 mm
Nominal cylinder stroke	–	800 mm
Usable experimental stroke	–	≈610 mm
Supply pressure	p	2.5 bar
Piston area	A_{push}	5026.5 mm^2
Piston-rod area	A_{rod}	490.9 mm^2
Force correction coefficient	η	0.95
Estimated extension force	F_{push}	1193.8 N
Estimated retraction force	F_{pull}	1077.2 N

At the experimental supply pressure of 2.5 bar, the estimated practical extension and retraction forces are therefore approximately 1194 N and 1077 N, respectively. The difference results from the reduced effective piston area during retraction and provides additional motivation for identifying separate pulse-duration–displacement models for the two directions of motion.

3.3 Software Architecture and Model-Based Pulse Calculation

The custom control software, implemented in C++ on the ESP32, utilizes an asynchronous state machine to manage the operational flow without blocking primary background processes. The positioning method uses an empirically identified pulse-duration–displacement relationship to estimate the required valve actuation. The software execution is divided into two primary phases: an automated calibration sequence and the active positioning calculation.

Auto-Calibration Sequence

When calibration is initiated, the cylinder repeatedly returns to a predefined home position and dispatches test pulses of varying durations, as illustrated in Fig. 1. After each pulse, the resulting traveled distance is recorded using the linear encoder.

In the experimental calibration procedure, pulse durations from 40 to 520 ms were evaluated in 40 ms increments. Each pulse duration was repeated five times, resulting in 65 displacement measurements for each direction of motion and each tested valve configuration.

Once the calibration dataset is collected, the method of least squares is applied to establish a linear regression model relating the pulse duration t to the traveled distance d:

$$d = k \cdot t + q \tag{3}$$

where k represents the experimentally identified displacement gain and q represents the model offset. Separate models are identified for forward and backward motion because the pneumatic actuator exhibits direction-dependent behaviour resulting from the different effective piston areas and other direction-dependent pneumatic and mechanical effects.

The calibration is intended as a commissioning or re-identification procedure rather than as a step required before every positioning command. After successful identification, the model parameters are stored in the Non-Volatile Storage (NVS) of the ESP32 and can be reused while the pneumatic and mechanical configuration remains unchanged. Recalibration may therefore be performed when relevant system parameters change, for example after replacing the valve, modifying the operating pressure, or changing the mechanical configuration.

The calibration procedure must be executed within predefined actuator limits. In the present experiments, calibration pulses were restricted to the 40–520 ms range and the cylinder was referenced to a predefined home position before subsequent test pulses. These constraints reduce the risk of excessive actuator travel during system identification. In practical deployment, the admissible calibration range should be selected according to the physical stroke and safety constraints of the specific machine and supplemented by appropriate hardware or software interlocks.

Linear Model Evaluation

Equation 3 is not intended to estimate the internal position of the valve spool or to represent the complete pneumatic dynamics. Instead, it provides a first-order empirical relationship between the commanded valve pulse duration and the resulting piston displacement under a fixed pneumatic and mechanical configuration. Its purpose is to provide a computationally lightweight estimate of the pulse duration required for a requested displacement.

The suitability of the linear approximation was evaluated using the coefficient of determination (R^2), the root mean square error (RMSE), and an analysis of the regression residuals. For each calibration measurement, the residual was calculated as

$$e_i = d_i - \hat{d}_i, \tag{4}$$

where d_i is the measured displacement and $\hat{d}_i$ is the displacement predicted by the fitted linear model.

Residual normality was assessed using the Shapiro–Wilk test at a significance level of $\alpha = 0.05$. This analysis was used as a diagnostic of whether the remaining approximation error could reasonably be described as normally distributed measurement variability. A rejection of residual normality was not treated as sufficient evidence to discard the linear model, since systematic pneumatic and mechanical effects may remain outside the first-order approximation. The linear model is therefore used as a lightweight pulse-duration estimator, while position feedback is used to evaluate the

remaining displacement and perform subsequent corrective movements when required.

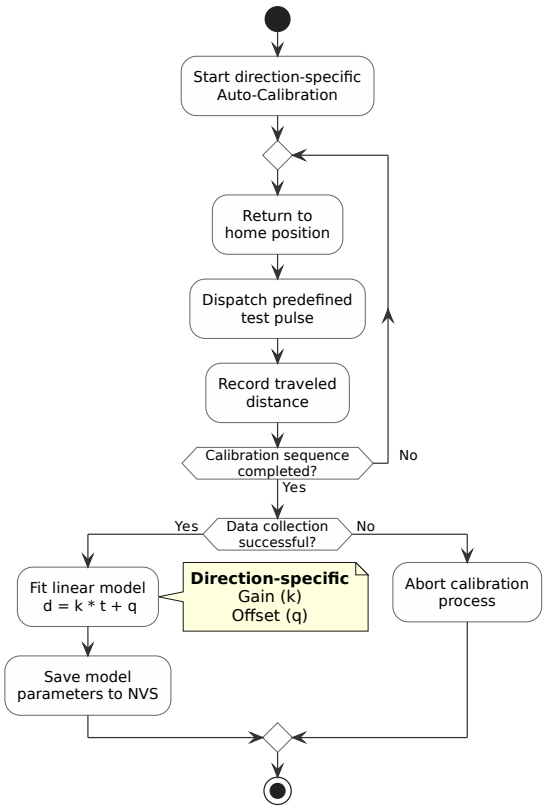


Figure 1. Auto-calibration workflow for direction-specific identification of the pulse-duration–displacement model.

Active Positioning and Pulse Calculation

During active positioning, the controller determines the required displacement from the current piston position and the requested target position. The corresponding forward or backward calibration model is selected according to the direction of motion. Using the inverse of the identified regression model in Equation 3, the required base pulse duration (t_{base}) is calculated as

$$t_{\text{base}} = \frac{d - q}{k}, \tag{5}$$

where d is the required displacement and k and q are the parameters of the corresponding directional calibration model.

The complete logical execution of the pulse calculation is illustrated in Fig. 2. The control sequence consists of the following steps:

1. **Target and Model Validation:** The system receives the target position, determines the required displacement, and verifies the validity of the corresponding regression model. If the stored model is invalid, pulse calculation is aborted and recalibration is required.

2. **Pulse Calculation and Heuristic Adjustment:** If the model is valid, t_{base} is calculated using Equation 5. A fixed pulse-duration factor of 0.95 is subsequently applied:

$$t_{\text{final}} = 0.95 \cdot t_{\text{base}}. \tag{6}$$

   The factor was selected heuristically during controller development to reduce the commanded pulse duration and decrease the likelihood of target overshoot. It was not determined through formal parameter optimization. Furthermore, this pulse-duration factor is independent of the force correction coefficient $\eta = 0.95$ used in the actuator force analysis.

3. **Pulse-Duration Constraint Filtering:** Before execution, t_{final} is evaluated against the configured pulse-duration limits. The lower pulse-duration limit was set to 38 ms for the Hafner configuration and 25 ms for the Pneumax configuration. Pulse requests at or below the corresponding limit are not executed. These values represent controller configuration parameters used in the experiments and should not be interpreted as universal physical switching thresholds of the valves.

   For both configurations, the maximum commanded pulse duration was limited to 1100 ms. Pulse requests exceeding this value are saturated at the upper limit.

4. **Execution and Feedback Correction:** If the calculated pulse satisfies the execution conditions, the corresponding ON/OFF valve is actuated for the commanded duration. After the pulse, a stabilization interval is applied before the piston position is evaluated using the linear encoder. The remaining displacement error is then used to determine whether an additional corrective movement is required.

The calibration data used to identify the linear models covered pulse durations from 40 to 520 ms. Therefore, pulse-duration estimates above 520 ms represent an extrapolation of the fitted first-order model. For longer movements, the resulting position is re-evaluated through feedback and any remaining displacement error is handled by subsequent corrective movements.

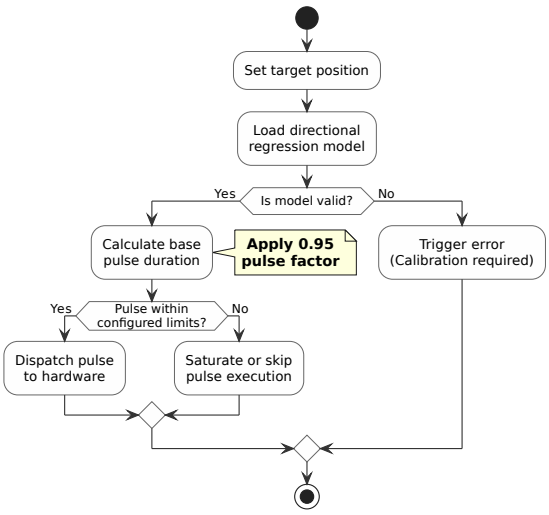


Figure 2. Pulse-duration calculation and execution workflow including the heuristic pulse factor and configured pulse-duration limits.

4 Results

This section presents the experimental results obtained for the two tested ON/OFF valve configurations. The evaluation is divided into calibration-model performance, residual analysis, and endpoint positioning validation.

4.1 Calibration Results

The calibration procedure produced separate pulse-duration–displacement models for both directions of motion and for both tested valve configurations. Each model was identified from 65 displacement measurements covering pulse durations from 40 to 520 ms.

The identified first-order models are summarized in Table 2, while the corresponding model parameters and goodness-of-fit metrics are reported in Table 3.

All four models achieved high coefficients of determination over the evaluated calibration range. The identified displacement gains also differed substantially between the two valve configurations. For example, the forward gain was approximately 0.4333 mm/ms for the Hafner configuration and 0.2017 mm/ms for the Pneumax configuration. These differences support the use of configuration-specific calibration rather than a single fixed pulse-duration model.

Table 2. Identified first-order pulse-duration–displacement models.

Configuration	Direction	Identified model
Hafner	FWD	$d = 0.4333t - 8.3791$
Hafner	BWD	$d = 0.4001t - 5.8931$
Pneumax	FWD	$d = 0.2017t - 5.5962$
Pneumax	BWD	$d = 0.1693t - 1.8552$

Table 3. Parameters and goodness-of-fit metrics of the identified linear calibration models.

Config.	Dir.	k	q	R^2	RMSE [mm]
Hafner	FWD	0.4333	-8.3791	0.9985	2.509
Hafner	BWD	0.4001	-5.8931	0.9975	2.980
Pneumax	FWD	0.2017	-5.5962	0.9970	1.655
Pneumax	BWD	0.1693	-1.8552	0.9943	1.913

4.2 Residual Analysis

The residuals of the four calibration models were analysed using the Shapiro–Wilk test at a significance level of $\alpha = 0.05$. The results are summarized in Table 4.

Table 4. Shapiro–Wilk test results for the calibration-model residuals.

Configuration	Direction	p-value
Hafner	FWD	0.093
Hafner	BWD	< 0.001
Pneumax	FWD	0.060
Pneumax	BWD	0.004

Residual normality was not rejected for the Hafner forward and Pneumax forward models. In contrast, normality was rejected for both backward models. Thus, although the first-order model captures the dominant pulse-duration–displacement relationship, the remaining approximation error cannot in all cases be described as normally distributed measurement variability.

The remaining residual structure may include pneumatic and mechanical effects that are not explicitly represented by the linear model, such as transient pressure behaviour, friction, valve dynamics, or state-dependent actuator response. The residual analysis alone does not allow the individual contribution of these effects to be identified. A higher-order or piecewise model could potentially reduce systematic approximation errors; however, the present implementation retains the linear model because of its low computational complexity and its use together with subsequent position-feedback corrections.

4.3 Positioning Validation

Fig. 3 compares the distributions of absolute endpoint positioning errors for the two tested valve configura-

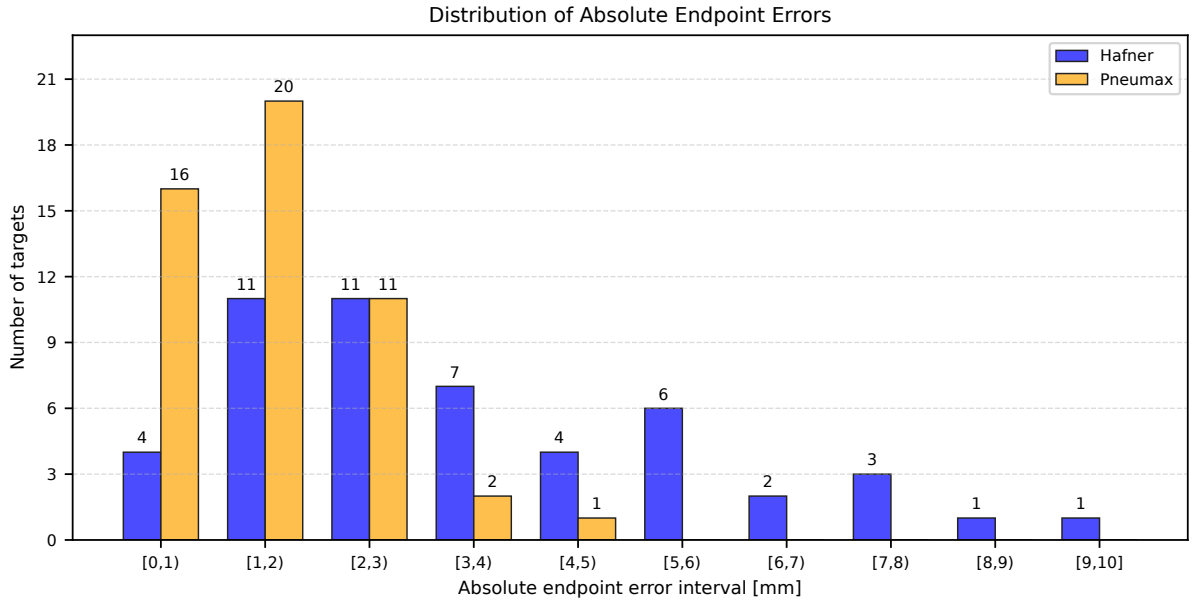


Figure 3. Comparison of the absolute endpoint-positioning error distributions for the Hafner and Pneumax valve configurations over the identical 50-target validation sequence.

tions.

The final positioning performance was evaluated using the same deterministic sequence of 50 target positions for both valve configurations. The experiments were performed using the same pneumatic cylinder, mechanical arrangement, position encoder, operating pressure of 2.5 bar, stabilization logic, and pulse-duration factor of 0.95. The target generator was reset to the same seed before each validation run, resulting in an identical target sequence for both configurations. The configured lower pulse-duration limits were 38 ms for the Hafner configuration and 25 ms for the Pneumax configuration.

The resulting endpoint-positioning metrics are summarized in Table 5.

Table 5. Endpoint positioning results for the identical 50-target validation sequence.

Metric	Hafner	Pneumax
Number of targets	50	50
MAE [mm]	3.441	1.495
SD of $\lvert e \rvert$ [mm]	2.262	0.929
Maximum $\lvert e \rvert$ [mm]	9.350	4.200

Under the evaluated experimental conditions, the Pneumax configuration produced lower endpoint positioning errors than the Hafner configuration. Its mean absolute error was approximately 56.5% lower. Because both configurations were evaluated using the identical sequence of target positions, a direct target-by-target comparison was also possible. The Pneumax configuration produced the lower absolute endpoint error for 38 of the 50 targets, whereas the Hafner configuration produced the lower error for 12 targets.

The two configurations also differed in the number of commanded movements required during the validation procedure. The Hafner configuration required approximately 1.08 commanded movements per target on average, whereas the Pneumax configuration required approximately 2.16. This indicates that the lower endpoint error obtained with the Pneumax configuration was accompanied by a greater number of corrective movements.

The observed performance difference should not be attributed solely to the configured lower pulse-duration limits. The two valve configurations also exhibited substantially different pulse-duration gains, residual behaviour, and short-pulse responses. The results therefore characterize the complete tested valve-controller configurations rather than isolating a single physical cause of the positioning difference.

The calibration models were identified using pulse durations up to 520 ms, whereas the positioning controller allowed commanded pulse durations up to 1100 ms. Pulse-duration estimates above the calibration range therefore represent an extrapolation of the fitted first-order models. In such cases, the position-feedback correction procedure remains an essential part of the complete positioning method.

5 Future Work

Future work will include validation using additional target sequences, operating pressures, and loading conditions. Further development will investigate adaptive pressure control, where a higher supply pressure could be used for rapid motion over larger distances and a lower pressure near the target for finer position-

ing. Pressure feedback from both cylinder chambers, velocity-aware pulse termination, and online adaptation of the pulse-duration factor will also be investigated to reduce positioning time and the number of corrective movements.

In addition, automatic model re-identification and over-the-air (OTA) firmware updates will be implemented to enable remote deployment of control algorithm improvements, calibration updates, and maintenance fixes without physical access to the system, thereby reducing downtime and simplifying lifecycle management of field-deployed cyber-physical systems (Kubaščík et al., 2024).

6 Conclusion

This study experimentally evaluated a lightweight positioning approach for a pneumatic linear actuator controlled by standard discrete ON/OFF valves. The method combines configuration- and direction-specific pulse-duration–displacement models with position-feedback corrections and was implemented on an ESP32-based controller.

The calibration experiments showed that a first-order linear model provides a close approximation of the dominant pulse-duration–displacement relationship over the evaluated 40–520 ms calibration range. The coefficients of determination exceeded 0.994 for all four identified models. Residual analysis, however, showed that the remaining errors were not normally distributed in all cases, indicating that the linear model does not capture all pneumatic and mechanical effects. The model should therefore be interpreted as a computationally lightweight pulse-duration estimator rather than as a complete physical model of the pneumatic system.

Using an identical sequence of 50 target positions at an operating pressure of 2.5 bar, the Hafner configuration achieved a mean absolute endpoint error of 3.441 mm, whereas the Pneumax configuration achieved 1.495 mm. The corresponding maximum absolute errors were 9.350 mm and 4.200 mm, respectively. Under the evaluated experimental conditions, the Pneumax configuration therefore provided lower endpoint positioning errors, although it also required a greater number of commanded corrective movements.

The observed differences should not be attributed solely to the configured lower pulse-duration limits, since the tested configurations also exhibited different pulse-duration gains, residual behaviour, and short-pulse responses. Furthermore, the results are specific to the tested valve specimens and experimental configuration and should not be interpreted as a general comparison of all Hafner and Pneumax valves.

References

De Carvalho, G., Bideaux, E., Massioni, P., Sesmat, S., & Bristiel, F. (2023). Pneumatic pressure control using on-off valves with pulse width modulation. *Chemical Engineering & Technology*, *46*(1), 21–28. doi:10.1002/ceat.202200410.

Duong, M.-D., Pham, Q.-T., Vu, T.-C., Bui, N.-T., & Dao, Q.-T. (2023). Adaptive fuzzy sliding mode control of an actuator powered by two opposing pneumatic artificial muscles. *Scientific Reports*, *13*(1), 8242. doi:10.1038/s41598-023-34491-3.

Gao, Q., Wang, J., Zhu, Y., Wang, J., & Wang, J. (2023). Research status and prospects of control strategies for high speed on/off valves. *Processes*, *11*(1), 160. doi:10.3390/pr11010160.

Hafner-Pneumatik Krämer GmbH & Co. KG. (2024, March 22). *How-to: Cylinder forces*. Retrieved August 25, 2026, from https://hafner-pneumatik.com/blog/post/how-to-cylinder-forces

Kubaščík, M., Tupý, A., Šumský, J., & Bača, T. (2024). Ota firmware updates on esp32 based microcontrolers, 185–189. doi:10.1109/Informatics62280.2024.10900824.

Li, Q., Hao, P., Wang, J., & Deng, H. (2023). Pulse-width-modulation-based time-delay compensation control for high-speed on/off valves. *Electronics*, *12*(17), 3627. doi:10.3390/electronics12173627.

Mandali, A., & Dong, L. (2022). Modeling and cascade control of a pneumatic positioning system. *Journal of Dynamic Systems, Measurement, and Control*, *144*(6), 061004. doi:10.1115/1.4053966.

Qi, H., Bone, G. M., & Zhang, Y. (2019). Position control of pneumatic actuators using three-mode discrete-valued model predictive control. *Actuators*, *8*(3), 56. doi:10.3390/act8030056.

Sciatti, F., Tamburrano, P., Distaso, E., & Amirante, R. (2024). Digital hydraulic valves: Advancements in research. *Heliyon*, *10*(5), e27264. doi:10.1016/j.heliyon.2024.e27264.

Šitum, Ž., & Ćorić, D. (2022). Position control of a pneumatic drive using a fuzzy controller with an analytic activation function. *Sensors*, *22*(3), 1004. doi:10.3390/s22031004.

Tootchi, A., & Chaibakhsh, A. (2025). Model-based and model-free position control of a servo pneumatic system in the presence of uncertainties and disturbances. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, *239*(4), 1081–1101. doi:10.1177/09544062241281558.

Xavier, M. S., Fleming, A. J., & Yong, Y. K. (2022). Model-based nonlinear feedback controllers for

pressure control of soft pneumatic actuators using on/off valves. *Frontiers in Robotics and AI*, *9*, 818187. doi:10.3389/frobt.2022.818187.

Xavier, M. S., Tawk, C. D., Zolfagharian, A., Pinskier, J., Howard, D., Young, T., Lai, J., Harrison, S. M., Yong, Y. K., Bodaghi, M., & Fleming, A. J. (2022). Soft pneumatic actuators: A review of design, fabrication, modeling, sensing, control and applications. *IEEE Access*, *10*, 59442–59485. doi:10.1109/ACCESS.2022.3179589.

Zhang, B., Jiang, A., Jiang, J., Qi, Y., Xue, L., & Wang, Y. (2022). A new positioning strategy based on parameter tuning and optimal control technique for pneumatic control valve. *Actuators*, *11*(10), 279. doi:10.3390/act11100279.

Towards Memory-Efficient Topological Feature Extraction for Biomedical Signals on Resource-Constrained Platforms

Prema memorijski učinkovitom izdvajanju topoloških značajki za biomedicinske signale na platformama s ograničenim resursima

Thi Kim-Anh Vo [1,2]

[1]Technical University of Ostrava
Faculty of Electrical Engineering and Computer Science
17. listopadu 2172/15 708 00 Ostrava-Poruba Czech Republic
thi.kim.anh.vo.st@vsb.cz

[2]Ton Duc Thang University
Faculty of Information Technology
19 Nguyen Huu Tho Street, Tan Hung Ward, Ho Chi Minh City, Vietnam
vothikimanh@tdtu.edu.vn

Abstract. *Edge-intelligence systems face a computational trilemma of precision, latency, and energy consumption. This paper introduces a hardware-software co-design framework utilizing Persistent Sheaf Laplacians for geometry-aware feature extraction. We present an optimized C/C++ Vietoris-Rips implementation that flattens symmetric distance matrices to reduce SRAM usage by 50%, alongside an Early-Terminated Union-Find algorithm halting topological reductions at the sheaf global section threshold. Simulations using MIT-BIH and PPG-DaLiA datasets demonstrate execution latencies under 8 ms and a 31.75 KB peak memory footprint, establishing real-time feasibility for embedded medical wearables*

Keywords. Topološka analiza podataka, perzistentni snopovni Laplaceov operator, kibernetic˘ko-fizic˘ki sustavi, zajednic˘ki dizajn sklopovlja i programske opreme, detekcija anomalija

Sažetak. *Sustavi rubne inteligencije suočavaju se s trilemom preciznosti, latencije i energije, dok je klasična perzistentna homologija ograničena DRAM pristupu. Ovaj rad predlaže okvir zajedničkog dizajna sklopovlja i programske opreme s perzistentnim snopovnim Laplaceovim operatorima za lokalizirano izdvajanje značajki. Uvodimo optimiziranu C/C++ implementaciju Vietoris-Ripsovog kompleksa s izravnavanjem matrice u 1D niz, smanjujući SRAM za 50%. Rano prekinuti Union-Find algoritam zaustavlja redukcije kod β_0. Profiliranje u ARM Cortex-M okruženju uz MIT-BIH i PPG-DaLiA pokazuje latencije ispod 8 ms i memorijski otisak od 31,75 KB za 128 uzoraka, potvrdivši izvedivost detekcije anomalija u nosivim uređajima.*

Ključne riječi. Topological Data Analysis, Persistent Sheaf Laplacian, Cyber-Physical Systems, Hardware-Software Co-Design, Anomaly Detection

1 Introduction

Raw-data streaming in edge-intelligence systems - such as wearable physiological monitors - faces a severe "computational trilemma" involving precision, latency, and energy consumption. The primary bottleneck for edge devices is not arithmetic computation, but rather the energy cost of off-chip memory access and external data transmission. Continuously streaming high-dimensional raw time-

1 Uvod

Prijenos sirovih podataka u sustavima rubne inteligencije - kao što su nosivi fiziološki monitori - suočava se s ozbiljnom „računalnom trilemom” koja uključuje preciznost, latenciju i potrošnju energije. Primarno usko grlo za rubne uređaje nije aritmetičko računanje, već energetski trošak pristupa izvančipnoj memoriji i vanjskog prijenosa podataka. Kontinuirani prijenos visokodimenzionalnih sirovih podataka vremenskih nizova u oblak troši za nekoliko redova

series data to the cloud consumes orders of magnitude more energy than processing data locally on- chip, rendering cloud-dependent streaming unsustainable for the stringent power envelopes of embedded medical wearables (Warden and Situnayake, 2019; Ray, 2022).

Deploying topological data analysis (TDA) (Carlsson, 2009; Edelsbrunner and Harer, 2008) at the extreme edge introduces significant engineering trade-offs. For a time series of length N, constructing a classical Vietoris-Rips complex requires evaluating all pairwise metric distances, scaling at a minimum quadratic complexity of O (N^2). Furthermore, conventional persistent homology suffers from combinatorial explosion during boundary matrix reduction, which typically requires worst-case cubic time complexity, O (M^3) (Otter et al., 2017; Bois et al., 2025). Even state-of-the-art computational optimizations (Bauer, 2021) clash directly with the constraints of microcontroller units (MCUs), which typically feature less than one megabyte of on-chip Static Random-Access Memory (SRAM).

To resolve this computational trilemma, this paper utilizes a cellular sheaf-theoretic representation to shift the paradigm from raw data transmission to highly localized, geometry-aware feature extraction.

Figure 1. The paradigm shift: Local topological feature extraction avoids the energy-intensive cloud streaming of raw physiological data.

Summary of Contributions:

- **Algorithmic Compression:** We propose a 1D symmetric distance matrix flattening technique that reduces the memory requirements of topological graph construction by over 50%, preventing dynamic SRAM overflow.
- **Computational Acceleration:** We introduce an Early-Terminated Union-Find algorithm that halts boundary reduction once the sheaf global section is reached, enabling execution latencies securely under 8 ms.
- **Preliminary Signal-Level Evaluation:** We validate the framework on the MIT-BIH and PPG-DaLiA datasets, suggests computational feasibility that localized sheaf features successfully isolate physiological anomalies from heterophilic motion artifacts.

veličine više energije nego lokalna obrada podataka na čipu, što prijenos ovisan o oblaku čini neodrživim za stroge energetske okvire ugrađenih medicinskih nosivih uređaja (Warden i Situnayake, 2019.; Ray, 2022.).

Implementacija topološke analize podataka (TDA) (Carlsson, 2009.; Edelsbrunner i Harer, 2008.) na samom rubu donosi značajne inženjerske kompromise. Za vremenski niz duljine N, konstrukcija klasičnog Vietoris-Ripsovog kompleksa zahtijeva evaluaciju svih parnih metričkih udaljenosti, skalirajući se s minimalno kvadratnom složenošću od O (N^2). Nadalje, konvencionalna perzistentna homologija pati od kombinatorne eksplozije tijekom redukcije granične matrice, što obično zahtijeva kubičnu vremensku složenost u najgorem slučaju, O(M^3) (Otter et al., 2017.; Bois et al., 2025.). Čak i najsuvremenije računalne optimizacije (Bauer, 2021.) izravno se sukobljavaju s ograničenjima mikrokontrolerskih jedinica (MCU), koje obično imaju manje od jednog megabajta statičke memorije s nasumičnim pristupom na čipu (SRAM).

Kako bi se riješila ova računalna trilema, ovaj rad koristi teorijski prikaz celularnih snopova kako bi se paradigma pomaknula s prijenosa sirovih podataka na visoko lokalizirano izdvajanje značajki svjesno geometrije.

Slika 1. Promjena paradigme: Lokalno izdvajanje topoloških značajki izbjegava energetski intenzivno strujanje sirovih fizioloških podataka u oblak.

Sažetak doprinosa:

- Algoritamska kompresija: Predlažemo tehniku izravnavanja simetrične matrice udaljenosti u 1D niz koja smanjuje memorijske zahtjeve konstrukcije topološkog grafa za više od 50\%, sprječavajući dinamičko prelijevanje SRAM-a.
- Računalno ubrzanje: Uvodimo rano prekinuti Union-Find algoritam koji zaustavlja redukciju granica kada se dosegne globalni presjek snopa, omogućujući latencije izvršavanja koje su sigurno ispod 8 ms.
- Preliminarna evaluacija na razini signala: Validiramo okvir na skupovima podataka MIT-BIH i PPG-DaLiA, što ukazuje na računalnu izvedivost toga da lokalizirane značajke snopa uspješno izoliraju fiziološke anomalije od heterofilnih artefakata kretanja.

2 Related work

2.1 Limitations of Standard Graph Networks

The development of Graph Neural Networks (GNNs) has indicateed a standard paradigm for modeling relational data. However, conventional GNNs impose severe performance bottlenecks when applied to heterophilic networks and non-stationary physiological time-series (Ju et al., 2024). Mainstream GNN architectures operate under the assumption of homophily - the premise that connected nodes share identical features. Spatial neighborhood aggregation recursively averages feature representations, behaving as a low-pass filter that smooths out local high-frequency variations. When applied to physiological systems characterized by sudden anomalies, this low-pass filter destroys critical localized signatures. Recent works have introduced Cellular Sheaves and Neural Sheaf Diffusion (Bodnar et al., 2021; Hansen et al., 2020) as a geometric mechanism to overcome this oversmoothing, allowing models to act as geometry-aware high-pass filters.

Table 1. Comparison of Relational Modeling Paradigms

Paradigm	Heterophily Handling	Edge Hardware Feasibility
Standard GNNs	Poor (Oversmoothing)	Moderate
Classical TDA	Excellent	Low (SRAM Overflow)
Persistent Sheaf (Ours)	Excellent (Restriction Maps)	High (1D Compression)

2.2 Topological Data Analysis on Edge Hardware

TDA has proven highly effective in extracting structural features from complex molecular and physical systems (Larraz, 2026; Wang et al., 2025). Recent advances in edge computing have introduced energy-aware machine learning models, such as MCUNet (Lin et al., 2020), which optimize neural architectures for microcontrollers. Concurrently, efforts to optimize data transmission on resource-scarce hardware have explored advanced edge-computing mechanisms, such as real-time D-PMU data compression for digital distribution networks (Yang et al., 2023). However, deploying Persistent

2 Srodni radovi

2.1 Ograničenja standardnih grafovskih mreža

Razvoj grafovskih neuronskih mreža (GNN) ukazao je na standardnu paradigmu za modeliranje relacijskih podataka. Međutim, konvencionalni GNN-ovi nameću ozbiljna uska grla u performansama kada se primijene na heterofilne mreže i nestacionarne fiziološke vremenske nizove (Ju et al., 2024.). Vodeće GNN arhitekture djeluju pod pretpostavkom \textit{homofilije} — premise da povezani čvorovi dijele identične značajke. Prostorna agregacija susjedstva rekurzivno usrednjava prikaze značajki, ponašajući se kao niskopropusni filtar koji izglađuje lokalne visokofrekventne varijacije. Kada se primijeni na fiziološke sustave koje karakteriziraju iznenadne anomalije, ovaj niskopropusni filtar uništava ključne lokalizirane potpise. Nedavni radovi uveli su celularne snopove i neuronsku difuziju snopova (Bodnar et al., 2021.; Hansen et al., 2020.) kao geometrijski mehanizam za prevladavanje ovog prekomjernog izglađivanja, omogućujući modelima da djeluju kao visokopropusni filtri svjesni geometrije.

Tablica 1. Usporedba paradigma relacijskog modeliranja

Paradigm	Heterophily Handling	Edge Hardware Feasibility
Standardne GNN mreže	Slabo (Prekomjerno izglađivanje)	Umjereno
Klasična TDA	Izvrsno	Nisko (Prelijevanje SRAM-a)
Perzistentni snop (Naš)	Izvrsno (Mape restrikcije)	Visoko (1D kompresija)

2.2 Topološka analiza podataka na rubnom sklopovlju

TDA se pokazala iznimno učinkovitom u izdvajanju strukturnih značajki iz složenih molekularnih i fizičkih sustava (Larraz, 2026.; Wang et al., 2025.). Nedavni napredak u rubnom računarstvu uveo je modele strojnog učenja svjesne potrošnje energije, kao što je MCUNet (Lin et al., 2020.), koji optimiziraju neuronske arhitekture za mikrokontrolere. Istodobno, napori za optimizaciju prijenosa podataka na sklopovlju s ograničenim resursima istražili su napredne mehanizme rubnog računarstva, kao što je kompresija D-PMU podataka u stvarnom vremenu za digitalne distribucijske mreže

topological Laplacians (Wei et al., 2025) on resource-scarce hardware still requires radical algorithmic compression to avoid SRAM overflow, a gap this paper directly addresses (Table 1).

(Yang et al., 2023.). Međutim, implementacija perzistentnih topoloških Laplaceovih operatora (Wei et al., 2025.) na sklopovlju s ograničenim resursima i dalje zahtijeva radikalnu algoritamsku kompresiju kako bi se izbjeglo prelijevanje SRAM-a, što je jaz koji ovaj rad izravno rješava (Tablica 1).

3 Theoretical Framework

Sheaf theory facilitates distributed signal fusion by capturing multi-modal data streams directly on the device. Persistent sheaf Laplacians encode both geometrical and non-geometrical information (Wei et al., 2025), capturing multiscale flexibility in biomolecular systems (Hayes et al, 2025). We extend this property to non-stationary cardiovascular signals.

- Local Opinions and Restriction Maps: Cellular sheaf theory equips a sensor network with a geometric structure. The raw signal captured by a sensor node acts as its "private opinion”
- The Coboundary Operator: The sheaf Laplacian explicitly measures the aggregated "disagreement of opinions" at each sensor node. It is formulated using block-matrix coboundary operators δ_h.

$$L_h = \delta_h^T \delta_h + \delta_{h-1} \delta_{h-1}^T \quad (1)$$

Normal physiological states are modeled to reside within the Harmonic Space ($H^0(G,F) = \ker(L_0)$), representing global consensus. Conversely, sudden anomalies (arrhythmias or motion artifacts) are projected outside this kernel as topological disagreements. Because this computation is driven by restriction maps linking adjacent nodes, the operations are transformed into sparse linear algebra routines that fit comfortably within constrained on-chip SRAM.

3 Teorijski okvir

Teorija snopova olakšava distribuiranu fuziju signala hvatanjem multimodalnih tokova podataka izravno na uređaju. Perzistentni snopovni Laplaceovi operatori kodiraju i geometrijske i negeometrijske informacije (Wei et al., 2025.), hvatajući višeskalnu fleksibilnost u biomolekularnim sustavima (Hayes et al., 2025.). Ovo svojstvo proširujemo na nestacionarne kardiovaskularne signale.

- Lokalna mišljenja i mape restrikcije: Teorija celularnih snopova oprema senzorsku mrežu geometrijskom strukturom. Sirovi signal koji uhvati senzorski čvor djeluje kao njegovo "privatno mišljenje".
- Operator kogranice: Snopovni Laplaceov operator eksplicitno mjeri agregirano "neslaganje mišljenja" na svakom senzorskom čvoru. Formuliran je korištenjem blok-matričnih operatora kogranice δ_h.

$$L_h = \delta_h^T \delta_h + \delta_{h-1} \delta_{h-1}^T \quad (1)$$

Normalna fiziološka stanja modelirana su tako da borave unutar Harmonijskog prostora ($H^0(G,F) = \ker(L_0)$), predstavljajući globalni konsenzus. Obrnuto, iznenadne anomalije (aritmije ili artefakti kretanja) projiciraju se izvan ove jezgre kao topološka neslaganja. Budući da je ovo izračunavanje vođeno mapama restrikcije koje povezuju susjedne čvorove, operacije se pretvaraju u rijetke rutine linearne algebre koje se udobno smještaju unutar ograničenog SRAM-a na čipu.

4 Hardware-Software Co-Design

To bypass the memory overhead of Python libraries, we implement a bare-metal C/C++ architecture tailored for Cortex-M processors. The proposed architecture is illustrated in Fig. 2.

4 Zajednički dizajn sklopovlja i programske opreme

Kako bismo izbjegli memorijsko preopterećenje koje uzrokuju Python biblioteke, implementiramo „bare-metal” C/C++ arhitekturu prilagođenu Cortex-M procesorima. Predložena arhitektura prikazana je na Slici 2.

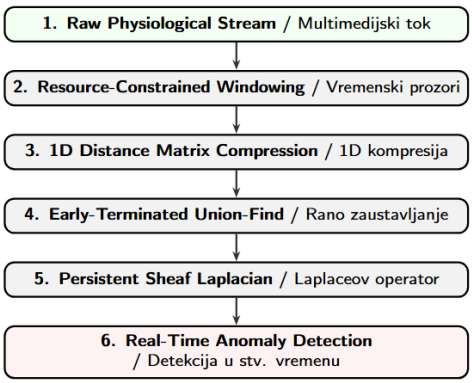
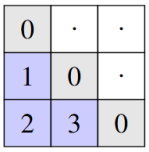


Figure 2. The proposed memory-optimized pipeline executing entirely on-chip.



Discarded (Upper + Diagonal)

Figure 3. Symmetric matrix compression: We index only the lower triangular portion (blue) for a 50% SRAM reduction.

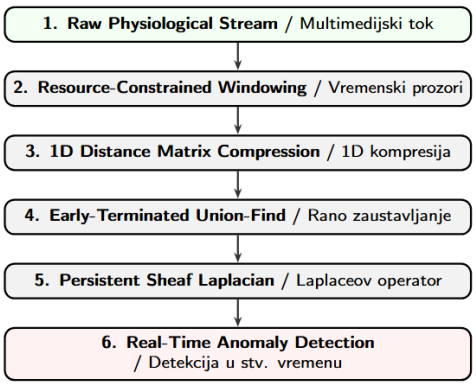
4.1 1D Distance Matrix Compression (Fig. 3)

The allocation for an N * N distance matrix demands substantial contiguous memory. Since the distance matrix of a real-valued physiological signal is symmetric with a zero diagonal, we flatten the 2D structure into a contiguous 1D array requiring strictly N*(N-1)/2 floating-point elements. For N=64, the required combinations drop from 4,096 to 2,016, saving precisely 50% of required SRAM.

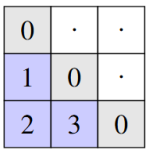
4.2 Early-Terminated Union-Find Algorithm

To identify the persistence of 0-dimensional topological features (β_0), we deploy a Kruskal-style Union-Find algorithm optimized with path compression. The algorithm executes an early stoppage the moment $\beta_0 = 1$. This threshold effectively caps the underlying simplicial complex and prevents dynamic memory overflow. Stopping at the sheaf global section threshold is mathematically viable for our specific use case because we assume a simplified, constant cellular sheaf, where global graph connectivity directly aligns with the harmonic consensus.

4.3 C++ Architectural Implementation



Slika 2. Predloženi cjevovod optimiziran za memoriju koji se izvršava na čipu.



Discarded (Upper + Diagonal)

Slika 3. Kompresija simetric˘ne matrice: Indeksiramo samo donji trokutasti dio (plavo) za 50%-tno smanjenje SRAM-a.

4.11D kompresija matrice udaljenosti (Slika 3)

Alokacija za matricu udaljenosti N * N zahtijeva znatnu kontinuiranu memoriju. Budući da je matrica udaljenosti fiziološkog signala realnih vrijednosti simetrična s nultom dijagonalom, izravnavamo 2D strukturu u kontinuirani 1D niz koji zahtijeva strogo N*(N-1)/2 elemenata s pomičnim zarezom. Za N=64, potrebne kombinacije padaju s 4096 na 2016, štedeći točno 50\% potrebnog SRAM-a.

4.2 Rano prekinuti Union-Find algoritam

Za identifikaciju perzistencije 0-dimenzionalnih topoloških značajki (β_0), primjenjujemo Kruskalov Union-Find algoritam optimiziran kompresijom puta. Algoritam izvršava rano zaustavljanje u trenutku kada je $\beta_0 = 1$. Ovaj prag učinkovito ograničava temeljni simplicijalni kompleks i sprječava dinamičko prelijevanje memorije. Zaustavljanje na pragu globalnog presjeka snopa matematički je održivo za naš specifični slučaj upotrebe jer pretpostavljamo pojednostavljeni, konstantni celularni snop, gdje se globalna povezanost grafa izravno podudara s harmonijskim konsenzusom.

To guarantee adherence to stringent SRAM limits, the framework avoids dynamic memory allocation during the continuous monitoring loop. To comply with modern embedded engineering standards, the bare-metal C++ implementation (Listing 1) utilizes memory-safe std::span for buffer views and constexpr for compile-time evaluation. The symmetric 2D distance matrix is mapped to a contiguous 1D array, and the Union-Find algorithm forcefully breaks the moment $\beta_0 = 1$, locking the memory state.

```cpp
#include <span>
#include <algorithm>

// 1. 1D Compressed Symmetric Matrix Indexing (Compile-time evaluated)
constexpr uint16_t get1DIndex(uint8_t i, uint8_t j) noexcept {
    if (i < j) std::swap(i, j); // Ensure lower-triangular
    return (i * (i - 1)) / 2 + j;
}

// 2. Early-Terminated Union-Find Execution utilizing memory-safe views
void computePersistentSheaf(uint8_t N, std::span<const Edge> sorted_edges) {
    uint8_t betti_0 = N; // Initialize components

    for (const auto& edge : sorted_edges) {
        uint8_t root_u = findRoot(edge.u);
        uint8_t root_v = findRoot(edge.v);

        if (root_u != root_v) {
            parent[root_u] = root_v; // Execute Union
            betti_0--;               // Decrement components

            // Sheaf Global Section Threshold Reached
            if (betti_0 == 1) {
                capSimplicialComplex();
                break; // HALT execution: Prevents memory overflow
            }
        }
    }
}
```

Listing 1. Modern C++ logic for 1D matrix compression and early-terminated topological reduction utilizing std::span and constexpr.

5. Results and Evaluation

5.1 Experimental Setup and Real-World Datasets

To validate the real-time feasibility of the proposed framework, physical hardware profiling was simulated using two distinct datasets: The MIT-BIH Arrhythmia Database provided a clinical baseline, while the PPG-DaLiA dataset introduced severe ambulatory motion artifacts (MAs) spanning 0.1 to 10 Hz (Reiss et al., 2019). Execution latencies and SRAM footprints were measured across dynamic analytical windows (N = 32, 64, 128).

5.2 SRAM Optimization and The 32KB Threshold

By implementing a 1D symmetric distance matrix compression, our algorithm precisely bounds the required floating-point elements. As demonstrated in Fig. 4, the memory footprint scales predictably: 1.94 KB for N=32, and 7.88 KB for N=64. Critically, even at the extended window size of N=128, the peak memory consumption is strictly capped at 31.75 KB. It is important to acknowledge that this reported 31.75 KB primarily represents the compressed distance matrix. While this proves that the core topological pipeline avoids SRAM overflow within the constrained 32 KB cache limits of ARM Cortex-M

4.3 Arhitektonska implementacija u jeziku C++

Kako bi se zajamčilo poštivanje strogih ograničenja SRAM-a, okvir izbjegava dinamičku alokaciju memorije tijekom petlje kontinuiranog praćenja. U skladu s modernim inženjerskim standardima za ugrađene sustave, “bare-metal” C++ implementacija (Ispis 1) koristi memorijski siguran std::span za preglede međuspremnika i constexpr za evaluaciju tijekom prevođenja. Simetrična 2D matrica udaljenosti preslikava se u kontinuirani 1D niz, a Union-Find algoritam se prisilno prekida u trenutku kada je $\beta_0 = 1$, zaključavajući stanje memorije.

```cpp
#include <span>
#include <algorithm>

// 1. 1D Compressed Symmetric Matrix Indexing (Compile-time evaluated)
constexpr uint16_t get1DIndex(uint8_t i, uint8_t j) noexcept {
    if (i < j) std::swap(i, j); // Ensure lower-triangular
    return (i * (i - 1)) / 2 + j;
}

// 2. Early-Terminated Union-Find Execution utilizing memory-safe views
void computePersistentSheaf(uint8_t N, std::span<const Edge> sorted_edges) {
    uint8_t betti_0 = N; // Initialize components

    for (const auto& edge : sorted_edges) {
        uint8_t root_u = findRoot(edge.u);
        uint8_t root_v = findRoot(edge.v);

        if (root_u != root_v) {
            parent[root_u] = root_v; // Execute Union
            betti_0--;               // Decrement components

            // Sheaf Global Section Threshold Reached
            if (betti_0 == 1) {
                capSimplicialComplex();
                break; // HALT execution: Prevents memory overflow
            }
        }
    }
}
```

Ispis 1. Moderna C++ logika za kompresiju 1D matrice i rano prekinutu topološku redukciju korištenjem std::span i constexpr.

5. Rezultati i evaluacija

5.1 Eksperimentalno postavljanje i stvarni skupovi podataka

Kako bi se potvrdila izvedivost predloženog okvira u stvarnom vremenu, profiliranje fizičkog sklopovlja simulirano je pomoću dva različita skupa podataka: baza podataka MIT-BIH Arrhythmia pružila je kliničku osnovu, dok je skup podataka PPG-DaLiA uveo teške artefakte kretanja (MA) u rasponu od 0,1 do 10 Hz (Reiss et al., 2019). Latencije izvršavanja i SRAM otisci mjereni su kroz dinamičke analitičke prozore (N = 32, 64, 128).

5.2 Optimizacija SRAM-a i prag od 32 KB

Implementacijom 1D kompresije simetrične matrice udaljenosti, naš algoritam precizno ograničava potrebne elemente s pomičnim zarezom. Kao što je prikazano na Slici 4, memorijski otisak skalira se predvidljivo: 1,94 KB za N=32 i 7,88 KB za N=64. Kritično, čak i pri proširenoj veličini prozora od $N=128$, vršna potrošnja memorije strogo je

microcontrollers, future work will comprehensively profile the total memory state of the application, including signal buffers, sorted edge lists, and sheaf-related matrices.

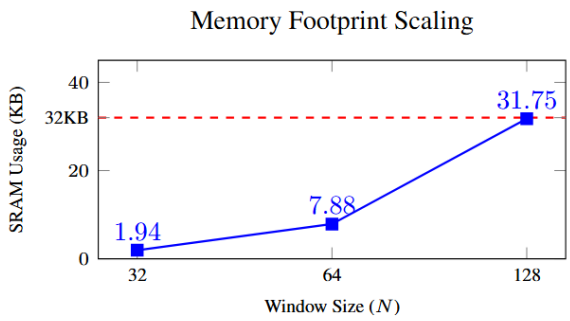


Figure 4. Quadratic scaling of the 1D compressed matrix under the 32 KB threshold

5.3 Execution Latency and Cache Anomalies

The early-terminated Union-Find algorithm yielded exceptional computational latency. As visualized in Fig. 5, execution required only 2.026 ms for MIT-BIH data (N=64) and an average of 2.290 ms for the highly noisy PPG-DaLiA dataset. Notably, the profiling revealed a distinct architectural anomaly: the N=32 window registered a higher execution time (7.889 ms) than the N=64 window (2.026 ms). This inversion is attributed to cold-start execution penalties and processor cache-misses during initial array allocations on the bare-metal architecture. Once the cache pipeline is primed, the framework processes 128-sample multi-channel streams in just 5.06 ms. These results suggest feasibility, while full hardware validation and energy measurements remain future work.

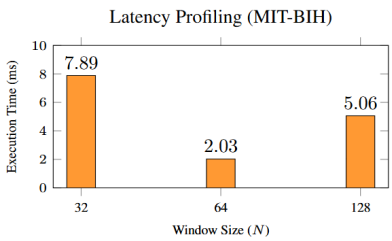
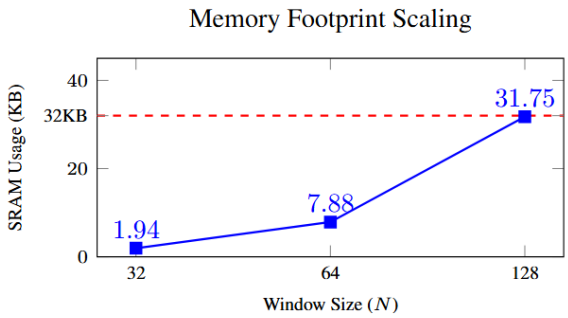


Figure 5. Execution latency across dynamic windows with cold-start anomaly at N = 32.

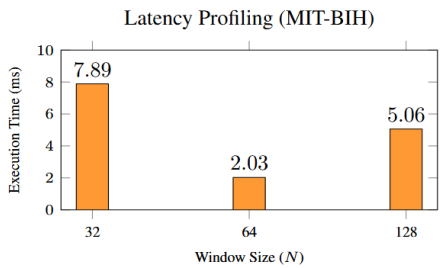
ograničena na 31,75 KB. Važno je napomenuti da ova prijavljena vrijednost od 31,75 KB prvenstveno predstavlja komprimiranu matricu udaljenosti. Iako ovo dokazuje da glavni topološki cjevovod izbjegava prelijevanje SRAM memorije unutar strogih ograničenja predmemorije (cache) od 32 KB kod ARM Cortex-M mikrokontrolera, budući rad će iscrpno profilirati cjelokupno stanje memorije aplikacije, uključujući međuspremnike signala, sortirane liste rubova i matrice povezane sa snopovima.



Slika 4. Kvadratno skaliranje komprimirane 1D matrice ispod praga od 32 KB.

5.3 Latencija izvršavanja i anomalije predmemorije

Rano prekinuti Union-Find algoritam pružio je iznimnu računalnu latenciju. Kao što je prikazano na Slici 5, izvršavanje je zahtijevalo samo 2,026 ms za podatke MIT-BIH (N=64) i prosječno 2,290 ms za iznimno šumom opterećen skup podataka PPG-DaLiA. Primjetno, profiliranje je otkrilo jasnu arhitektonsku anomaliju: prozor s N=32 zabilježio je dulje vrijeme izvršavanja (7,889 ms) od prozora s N=64 (2,026 ms). Ova inverzija pripisuje se penalizacijama izvršavanja zbog "hladnog starta" (cold-start) i promašajima predmemorije (cache-misses) procesora tijekom početnih alokacija nizova na bare-metal arhitekturi. Jednom kada je cjevovod predmemorije pripremljen, okvir obrađuje višekanalne tokove od 128 uzoraka u samo 5,06 ms. Ovi rezultati ukazuju na izvedivost, dok potpuna validacija sklopovlja i mjerenja energije ostaju ciljevi za budući rad.



Slika 5. Latencija izvršavanja kroz dinamičke prozore s anomalijom hladnog starta pri N = 32.

5.4 Heterophilic Artifact Isolation in PPG

A deeper clinical analysis of the PPG-DaLiA outputs reveals why the persistent sheaf framework is mathematically necessary. In standard Euclidean networks or traditional Short-Time Fourier Transforms (STFT), high-frequency motion artifacts overlap with cardiac frequencies. This results in oversmoothing, where the network erroneously averages the signal and locks onto the dominant heterophilic noise. Our early-terminated Hodge filtration effectively acts as a geometry-aware high-pass filter, isolating these rapid motion anomalies into localized gradient flows.

While traditional deep learning approaches often rely on heavy feature extraction to report classification metrics like F1-scores, our goal is to validate the mathematical viability of topological features themselves under strict hardware limits. Under multiple linear regression, the isolated harmonic energy fraction (H%) of the cardiovascular wave retains a statistically significant correlation with the ground-truth heart rate ($p = 0.015$). This robust statistical significance provides that the early-terminated sheaf global section successfully captures the true physiological state, providing a rigorous foundation for anomaly detection without the computational overhead of a secondary neural network.

5.4 Izolacija heterofilnih artefakata u PPG-u

Dublja klinička analiza izlaza PPG-DaLiA otkriva zašto je okvir perzistentnih snopova matematički nužan. U standardnim euklidskim mrežama ili tradicionalnim kratkovremenim Fourierovim transformacijama (STFT), visokofrekventni artefakti kretanja preklapaju se s kardiološkim frekvencijama. To rezultira prekomjernim izglađivanjem, pri čemu mreža pogrešno usrednjava signal i fiksira se na dominantni heterofilni šum. Naša rano prekinuta Hodgeova filtracija učinkovito djeluje kao visokopropusni filtar svjestan geometrije, izolirajući ove brze anomalije kretanja u lokalizirane gradijentne tokove.

Dok se tradicionalni pristupi dubokog učenja često oslanjaju na intenzivno izdvajanje značajki kako bi izvijestili o klasifikacijskim metrikama poput F1-mjera, naš je cilj potvrditi matematičku održivost samih topoloških značajki pod strogim ograničenjima sklopovlja. Pod višestrukom linearnom regresijom, izolirani udio harmonijske energije (H%) kardiovaskularnog vala zadržava statistički značajnu korelaciju sa stvarnom frekvencijom srca ($p = 0.015$). Ova robusna statistička značajnost potvrđuje da rano prekinuti globalni presjek snopa uspješno zahvaća stvarno fiziološko stanje, pružajući rigorozan temelj za detekciju anomalija bez dodatnog računalnog opterećenja sekundarnom neuronskom mrežom.

5.5 Comparative Baseline Analysis

To better contextualize the efficiency of the Persistent Sheaf Laplacian, Table 2 compares our approach against standard lightweight deep learning and signal processing techniques under strict microcontroller limits. Unlike Convolutional Neural Networks (CNNs) which demand heavy SRAM for feature maps, or Short-Time Fourier Transforms (STFT) which lack geometric awareness, our pipeline achieves ultra-low memory requirements while preserving heterophilic anomaly isolation.

Table 2. Comparative Baseline: Edge Intelligence Methods for Biomedical Signals

Methodology	SRAM Footprint	Execution Latency	Heterophily Handling
Lightweight CNN (Lin et al., 2020)	High (~ 217 KB)	> 100 ms	Poor (Oversmoothing)

5.5 Usporedna bazna analiza

Kako bi se bolje kontekstualizirala učinkovitost perzistentnog snopovnog Laplaceovog operatora, Tablica 2 uspoređuje naš pristup sa standardnim laganim tehnikama dubokog učenja i obrade signala pod strogim ograničenjima mikrokontrolera. Za razliku od konvolucijskih neuronskih mreža (CNN) koje zahtijevaju velik SRAM za mape značajki, ili kratkovremenih Fourierovih transformacija (STFT) kojima nedostaje geometrijska svjesnost, naš cjevovod postiže iznimno niske memorijske zahtjeve uz istovremeno očuvanje izolacije heterofilnih anomalija.

Tablica 2: Usporedna osnovica: Metode rubne inteligencije za biomedicinske signale

Metodologija	Otisak SRAM-a	Latencija izvršavanja	Upravljanje heterofilijom
Lagani CNN (Lin et al., 2020)	Visoko (~ 217 KB)	> 100 ms	Slabo (Prekomjerno izglađivanje)

ICA / Adaptive Filtering (Martinek et al., 2018)	Medium (> 50 KB)	Iterative	Moderate
Persistent Sheaf (Ours)	Ultra-Low (31.75 KB)	< 8 ms	Excellent (Isolated)

Note: The CNN baseline reflects MobileNetV2 performance via CMSIS-NN (Lin et al., 2020). Traditional ICA reflects standard matrix-heavy filtering (Martinek et al., 2018). 'Persistent Sheaf' metrics are derived from our experimental profiling.

5.6 Interactive Edge Hardware Simulator

To facilitate reproducibility and allow the research community to visually explore the dynamic scaling limits of the proposed hardware-software co-design, an interactive web-based simulator was developed (Fig. 6). The dashboard mathematically models the exact static RAM footprint scaling of the 1D compressed distance matrix against the 32 KB physical threshold of Cortex-M architectures. Furthermore, it allows users to dynamically inject heterophilic motion artifacts (mimicking the PPG-DaLiA dataset) into stable rhythmic signals to observe the real-time early-terminated Hodge filtration catching the topological void. The interactive dashboard and associated simulation source code have been temporarily hosted at an anonymized repository: https://anonymous.4open.science/w/Memory-Optimized-Persistent-Sheaf-LaplaciansJune-D690/

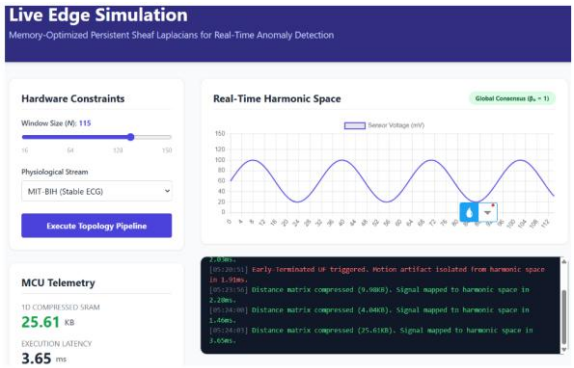


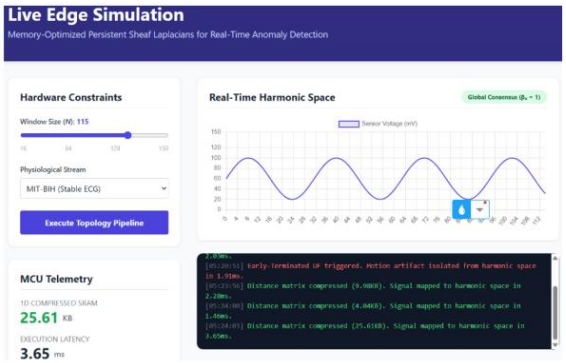
Figure 6. The Interactive Edge CPS Simulator. The dashboard visualizes the mathematical thresholds where unoptimized simplicial complexes trigger 32 KB SRAM memory overflows.

ICA / Adaptivno filtriranje (Martinek et al., 2018)	Srednje (> 50 KB)	Iterativno	Umjereno
Perzistentni snop (Naš)	Izuzetno nisko (31.75 KB)	< 8 ms	Izvrsno (Izolirano)

Napomena: CNN osnovica odražava performanse modela MobileNetV2 putem CMSIS-NN knjižnice (Lin et al., 2020). Tradicionalna ICA odražava standardno filtriranje s teškim matričnim operacijama (Martinek et al., 2018). Metrike za „Perzistentni snop" izvedene su iz našeg eksperimentalnog profiliranja.

5.6 Interaktivni simulator rubnog sklopovlja

Kako bi se olakšala ponovljivost i omogućilo istraživačkoj zajednici da vizualno istraži dinamičke granice skaliranja predloženog zajedničkog dizajna sklopovlja i programske opreme, razvijen je interaktivni simulator temeljen na webu (Slika 6). Nadzorna ploča matematički modelira točno skaliranje otiska statičkog RAM-a 1D komprimirane matrice udaljenosti u odnosu na fizički prag od 32 KB Cortex-M arhitektura. Nadalje, omogućuje korisnicima da u stabilne ritmičke signale dinamički unesu heterofilne artefakte kretanja (koji oponašaju skup podataka PPG-DaLiA) kako bi promatrali kako rano prekinuta Hodgeova filtracija u stvarnom vremenu hvata topološku prazninu. Interaktivna nadzorna ploča i pripadajući izvorni kod simulacije privremeno su smješteni u anonimiziranom repozitoriju: https://anonymous.4open.science/w/Memory-Optimized-Persistent-Sheaf-LaplaciansJune-D690/



Slika 6. Interaktivni simulator rubnog CPS-a. Nadzorna ploˇca vizualizira matemati˘cke pragove na kojima neoptimizirani simplicijalni kompleksi uzrokuju prelijevanje SRAM memorije od 32 KB.

6 Limitations and Hardware Trade-offs

While this exploratory work resolves the immediate memory bottleneck for pairwise-distance calculations, the framework has inherent limitations. First, the boundary condition of the early-terminated Union-Find algorithm acts as a strict safety mechanism against combinatorial explosion. However, stopping at $\beta_0=1$ discards higher-dimensional topological features (such as 1D cycles or 2D voids). Future work must mathematically quantify this lost analytical resolution using a localized Wasserstein distance metric to rigorously evaluate the trade-off between memory safety and topological fidelity.

Finally, the current evaluation relies on simulated ARM Cortex-M environments. True physical deployment will introduce variable power dynamics and memory bus contentions that simulation cannot perfectly replicate. Consequently, future physical deployments will target specific physical microcontroller boards - such as the STM32L4 and Nordic nRF52840 families - to validate actual power consumption and hardware-in-the-loop performance.

7 Future Directions and Strategic Applications

The successful hardware-software co-design of edge-based sheaf topologies opens several highly strategic vectors for future Cyber-Physical Systems (CPS) research, specifically in advanced biomedical and industrial applications.

7.1 Non-Invasive Fetal ECG and Complex Signal Separation

A critical future application of this framework is the real-time extraction of Non-Invasive Fetal Electrocardiograms (NI-FECG) from abdominal sensors. Current state-of-the-art methods rely heavily on Independent Component Analysis (ICA) and adaptive filtering to separate fetal and maternal signals (Martinek et al., 2018; Jaros et al., 2020). However, these statistical methods often struggle with the overlapping spectral frequencies of maternal and fetal cardiac cycles in highly noisy ambulatory environments. Because the persistent sheaf Laplacian acts as a non-parametric, geometry-aware filter, future iterations will explore using restriction maps to topologically isolate the high-frequency fetal QRS complexes (treated as topological voids) from the dominant maternal cardiac cycle (the harmonic consensus). Achieving this on-edge without the

6 Ograničenja i kompromisi sklopovlja

Iako ovaj istraživački rad rješava trenutačno memorijsko usko grlo za izračune udaljenosti po parovima, okvir ima inherentna ograničenja. Prvo, granični uvjet rano prekinutog Union-Find algoritma djeluje kao strogi sigurnosni mehanizam protiv kombinatorne eksplozije. Međutim, zaustavljanje na $\beta_0=1$ odbacuje višedimenzionalne topološke značajke (kao što su 1D ciklusi ili 2D praznine). Budući rad mora matematički kvantificirati ovu izgubljenu analitičku rezoluciju pomoću lokalizirane metrike Wassersteinove udaljenosti kako bi se rigorozno procijenio kompromis između memorijske sigurnosti i topološke vjernosti.

Naposljetku, trenutačna evaluacija oslanja se na simulirana okruženja ARM Cortex-M. Stvarna fizička implementacija uvest će varijabilnu dinamiku potrošnje energije i nadmetanja na memorijskoj sabirnici koje simulacija ne može savršeno replicirati. Stoga će buduće fizičke implementacije ciljati na specifične fizičke mikrokontrolerske ploče - poput obitelji STM32L4 i Nordic nRF52840 - kako bi se potvrdila stvarna potrošnja energije i performanse u petlji sa sklopovljem (hardware-in-the-loop).

7 Budući smjerovi i strateške primjene

Uspješan zajednički dizajn sklopovlja i programske opreme za rubne topologije snopova otvara nekoliko iznimno strateških vektora za buduća istraživanja kibernetičko-fizičkih sustava (CPS), specifično u naprednim biomedicinskim i industrijskim primjenama.

7.1 Neinvazivni fetalni EKG i razdvajanje složenih signala

Ključna buduća primjena ovog okvira je izdvajanje neinvazivnih fetalnih elektrokardiograma (NI-FECG) u stvarnom vremenu iz abdominalnih senzora. Trenutačne najsuvremenije metode uvelike se oslanjaju na analizu nezavisnih komponenti (ICA) i adaptivno filtriranje kako bi se razdvojili fetalni i majčini signali (Martinek et al., 2018.; Jaros et al., 2020.). Međutim, ove se statističke metode često bore s preklapajućim spektralnim frekvencijama majčinih i fetalnih srčanih ciklusa u iznimno šumom opterećenim ambulantnim okruženjima. Budući da perzistentni snopovni Laplaceov operator djeluje kao neparametarski filtar svjestan geometrije, buduće iteracije istraživat će korištenje mapa restrikcije za

computational overhead of iterative ICA matrices represents a major leap for continuous, low-power maternal-fetal monitoring.

7.2 Smart Mobility and Automotive CPS

Beyond wearable medical monitors, this hardware-optimized framework has immediate applications in Smart Mobility. We plan to adapt the algorithm for automotive cyber-physical systems, specifically for continuous driver fatigue and anomaly monitoring using steering-wheel-embedded PPG and ECG sensors. Processing these multi-channel signals locally on automotive microcontrollers ensures zero-latency safety interventions independent of 5G cellular network stability.

7.3 Industrial IoT and the Edge-to-Cloud Continuum

Because the sheaf Laplacian natively compresses distributed sensor signals into lightweight topological summaries (Betti numbers and harmonic fractions), it is an ideal candidate for large-scale Industrial IoT (IIoT) and predictive maintenance. Future research will explore deploying these algorithms on ultra-low-power industrial sensors to detect acoustic and vibrational micro-anomalies in heavy machinery. Furthermore, instead of transmitting raw telemetry, edge nodes will push these minimal topological summaries to centralized supercomputing infrastructures for integration into global Digital Twins, optimizing the edge-to-cloud continuum.

8 Conclusion

This paper introduces a memory-optimized Persistent Sheaf Laplacian algorithm designed specifically for resource-scarce embedded platforms. By compressing spatial distance computations into a 1D matrix and deploying an early-terminated Union-Find execution, we restrict dynamic SRAM consumption to under 32 KB while processing multi-channel signals in under 8 ms. The preliminary results suggest that this hardware-software co-design makes selected topological feature extraction operations computationally feasible on memory-constrained embedded platforms. This lays the groundwork for future physical hardware validation, total memory profiling, energy measurements, and task-level anomaly detection in cyber-physical systems.

References

Bauer, U. (2021). Ripser: Efficient Computation of Vietoris-Rips Persistence Barcodes. *Journal of*

topološku izolaciju visokofrekventnih fetalnih QRS kompleksa (tretiranih kao topološke praznine) od dominantnog majčinog srčanog ciklusa (harmonijskog konsenzusa). Postizanje ovoga na samom rubu bez računalnog opterećenja iterativnih ICA matrica predstavlja golem iskorak za kontinuirano praćenje majke i fetusa uz nisku potrošnju energije.

7.2 Pametna mobilnost i automobilski CPS

Osim nosivih medicinskih monitora, ovaj sklopovski optimizirani okvir ima izravne primjene u pametnoj mobilnosti. Planiramo prilagoditi algoritam za automobilske kibernetičko-fizičke sustave, specifično za kontinuirano praćenje umora vozača i anomalija pomoću PPG i EKG senzora ugrađenih u upravljač. Lokalna obrada ovih višekanalnih signala na automobilskim mikrokontrolerima osigurava sigurnosne intervencije s nultom latencijom, neovisno o stabilnosti 5G mobilne mreže.

7.3 Industrijski IoT i kontinuum od ruba do oblaka

Budući da snopovni Laplaceov operator izvorno komprimira distribuirane senzorske signale u lagane topološke sažetke (Bettijeve brojeve i harmonijske udjele), on je idealan kandidat za industrijski IoT (IIoT) velikih razmjera i prediktivno održavanje. Buduća istraživanja istraživat će primjenu ovih algoritama na industrijskim senzorima iznimno niske potrošnje energije radi otkrivanja akustičkih i vibracijskih mikroanomalija u teškim strojevima. Nadalje, umjesto prijenosa sirove telemetrije, rubni će čvorovi slati ove minimalne topološke sažetke u centralizirane superračunalne infrastrukture radi integracije u globalne digitalne blizance, čime se optimizira kontinuum od ruba do oblaka.

8 Zaključak

Ovaj rad predstavlja memorijski optimiziran algoritam perzistentnog snopovnog Laplaceovog operatora, dizajniran specifično za ugrađene platforme s ograničenim resursima. Komprimiranjem izračuna prostornih udaljenosti u 1D matricu i primjenom rano prekinutog Union-Find izvršavanja, ograničavamo dinamičku potrošnju SRAM-a na manje od 32 KB, uz obradu višekanalnih signala za manje od 8 ms. Preliminarni rezultati ukazuju na to da ovaj zajednički dizajn sklopovlja i programske opreme čini odabrane operacije izdvajanja topoloških značajki računalno izvedivima na memorijski ograničenim ugrađenim platformama. Time se postavljaju temelji za buduću validaciju na fizičkom sklopovlju, cjelovito profiliranje memorije, mjerenja

energije i detekciju anomalija na razini zadataka u kibernetičko-fizičkim sustavima.

Reference

Bauer, U. (2021). Ripser: Efficient Computation of Vietoris-Rips Persistence Barcodes. *Journal of Applied and Computational Topology*, 5(3), 391–423.

Bodnar, C., Di Giovanni, F., Chamberlain, B. P., Lio, P., & Bronstein, M. M. (2021). Neural Sheaf Diffusion: A Topological Perspective on Heterophily and Oversmoothing in GNNs. *Advances in Neural Information Processing Systems (NeurIPS), 34.*

Bois, A., Henneuse, H., Tervil, B., & Oudre, L. (2025). Persistent Homology for Time Series: A Selective Review. *OpenReview*.

Carlsson, G. (2009). Topology and Data. *Bulletin of the American Mathematical Society*, 46(2), 255–308.

Edelsbrunner, H., & Harer, J. (2008). Persistent Homology - a Survey. *Contemporary Mathematics*, 453, 257–282.

Hansen, J., & Gebhart, T. (2020). Sheaf Neural Networks. *NeurIPS 2020 Workshop on Topological Data Analysis and Beyond*.

Hayes, N., Wei, X., Feng, H., Merkurjev, E., & Wei, G.-W. (2025). Persistent Sheaf Laplacian Analysis of Protein Flexibility. *The Journal of Physical Chemistry B*, 129(17), 4169–4178.

Jaros, R., Martinek, R., & Danys, L. (2020). Fetal ECG Signal Processing by Different ICA-based Algorithms. *2020 21st International Scientific Conference on Electric Power Engineering* (EPE), 1–6.

Ju, W., et al. (2024). A survey of graph neural networks in real world: Imbalance, noise, privacy and OOD challenges. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *48*(3), 3036–3055.

Larraz, R. (2026). Unsupervised Diagnosis of Industrial Alarm Patterns Using Topological Data Analysis. *ChemRxiv*.

Lin, J., Chen, W.-M., Lin, Y., Cohn, J., Gan, C., & Han, S. (2020). MCUNet: Tiny Deep Learning on IoT Devices. *Advances in Neural Information Processing Systems (NeurIPS)*, 33, 11711–11722.

Martinek, R., Konecny, J., Bilik, P., et al. (2018). Comparative Effectiveness of ICA and PCA in Extraction of Fetal ECG From Abdominal Signals: Toward Non-invasive Fetal Monitoring. *Frontiers in Physiology*, 9, 648.

Otter, N., Porter, M. A., Townsend, U., Grindrod, P., & Harrington, H. A. (2017). A Roadmap for the Computation of Persistent Homology. *EPJ Data Science*, 6(1), 17.

Ray, P. P. (2022). A Review on TinyML: State-of-the-art and Prospects. *Journal of King Saud University*

Bauer, U. (2021). Ripser: Efficient Computation of Vietoris-Rips Persistence Barcodes. *Journal of Applied and Computational Topology*, 5(3), 391–423.

Bodnar, C., Di Giovanni, F., Chamberlain, B. P., Lio, P., & Bronstein, M. M. (2021). Neural Sheaf Diffusion: A Topological Perspective on Heterophily and Oversmoothing in GNNs. *Advances in Neural Information Processing Systems (NeurIPS), 34.*

Bois, A., Henneuse, H., Tervil, B., & Oudre, L. (2025). Persistent Homology for Time Series: A Selective Review. *OpenReview*.

Carlsson, G. (2009). Topology and Data. *Bulletin of the American Mathematical Society*, 46(2), 255–308.

Edelsbrunner, H., & Harer, J. (2008). Persistent Homology - a Survey. *Contemporary Mathematics*, 453, 257–282.

Hansen, J., & Gebhart, T. (2020). Sheaf Neural Networks. *NeurIPS 2020 Workshop on Topological Data Analysis and Beyond*.

Hayes, N., Wei, X., Feng, H., Merkurjev, E., & Wei, G.-W. (2025). Persistent Sheaf Laplacian Analysis of Protein Flexibility. *The Journal of Physical Chemistry B*, 129(17), 4169–4178.

Jaros, R., Martinek, R., & Danys, L. (2020). Fetal ECG Signal Processing by Different ICA-based Algorithms. *2020 21st International Scientific Conference on Electric Power Engineering* (EPE), 1–6.

Ju, W., et al. (2024). A survey of graph neural networks in real world: Imbalance, noise, privacy and OOD challenges. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *48*(3), 3036–3055.

Larraz, R. (2026). Unsupervised Diagnosis of Industrial Alarm Patterns Using Topological Data Analysis. *ChemRxiv*.

Lin, J., Chen, W.-M., Lin, Y., Cohn, J., Gan, C., & Han, S. (2020). MCUNet: Tiny Deep Learning on IoT Devices. *Advances in Neural Information Processing Systems (NeurIPS)*, 33, 11711–11722.

Martinek, R., Konecny, J., Bilik, P., et al. (2018). Comparative Effectiveness of ICA and PCA in Extraction of Fetal ECG From Abdominal Signals: Toward Non-invasive Fetal Monitoring. *Frontiers in Physiology*, 9, 648.

– *Computer and Information Sciences*, 34(4), 1595–1623.

Reiss, A., Indlekofer, I., Schmidt, P., & Van Laerhoven, K. (2019). Deep PPG: Large-Scale Heart Rate Estimation with Convolutional Neural Networks. *Sensors*, 19(14), 3079.

Wang, B., Feng, B., Lv, L., Li, S., & Pan, F. (2025). Structural Feature Extraction via Topological Data Analysis. *The Journal of Physical Chemistry Letters*, 16(32), 8056–8067.

Warden, P., & Situnayake, D. (2019). TinyML: Machine Learning with TensorFlow Lite on Arduino and Ultra-Low-Power Microcontrollers. *O'Reilly Media*.

Wei, X., & Wei, G.-W. (2025). Persistent Sheaf Laplacians. *Foundations of Data Science*, 7(2), 446.

Yang, J., Yu, H., Ji, H., Xi, W., Wu, J., & Wang, C. (2023). Real-Time D-PMU Data Compression for Edge Computing Devices in Digital Distribution Networks. *IEEE Transactions on Power Systems*, *39*(4), 5712–5725.

Otter, N., Porter, M. A., Townsend, U., Grindrod, P., & Harrington, H. A. (2017). A Roadmap for the Computation of Persistent Homology. *EPJ Data Science*, 6(1), 17.

Ray, P. P. (2022). A Review on TinyML: State-of-the-art and Prospects. *Journal of King Saud University – Computer and Information Sciences*, 34(4), 1595–1623.

Reiss, A., Indlekofer, I., Schmidt, P., & Van Laerhoven, K. (2019). Deep PPG: Large-Scale Heart Rate Estimation with Convolutional Neural Networks. *Sensors*, 19(14), 3079.

Wang, B., Feng, B., Lv, L., Li, S., & Pan, F. (2025). Structural Feature Extraction via Topological Data Analysis. *The Journal of Physical Chemistry Letters*, 16(32), 8056–8067.

Warden, P., & Situnayake, D. (2019). TinyML: Machine Learning with TensorFlow Lite on Arduino and Ultra-Low-Power Microcontrollers. *O'Reilly Media*.

Wei, X., & Wei, G.-W. (2025). Persistent Sheaf Laplacians. *Foundations of Data Science*, 7(2), 446.

Yang, J., Yu, H., Ji, H., Xi, W., Wu, J., & Wang, C. (2023). Real-Time D-PMU Data Compression for Edge Computing Devices in Digital Distribution Networks. *IEEE Transactions on Power Systems*, *39*(4), 5712–5725.

Securing user data in mobile application development: review and case study

Andrej Pavešić, Renata Mekovec
University of Zagreb Faculty of Organization and Informatics
Pavlinska 2, 42000 Varaždin, Croatia
apavesic22@student.foi.hr, renata.mekovec@foi.unizg.hr

***Abstract**. This paper examines recent advances in mobile app development research and analyzes practices employed by developers to ensure user data security. Utilizing bibliometric analysis—specifically performance analysis and science mapping—we identify emerging trends and examine the domain's intellectual structure. Additionally, the paper outlines critical security threats confronting developers and users. A qualitative case study examines BetterHelp, MyFitnessPal, and Noom—three popular Android health apps handling sensitive medical and personal data that raise privacy concerns. The investigation employs qualitative comparative content analysis alongside platform-specific technical examinations.*

Keywords. Mobile application, self-regulation, data protection

1 Introduction

Mobile applications serve as a primary channel for accessing digital content and services. Since smartphones are personal devices, they fall under the category of intimacy and privacy for their owners. Therefore, it is imperative for individuals to manage the data accessible to mobile applications. Sensor Tower states that the average person spent 3.6 hours a day on their mobile phone last year and uses 10 apps each day (SensorTower, 2026). While users have over 80 apps on average (roughly 62% of those are unused) the majority of a user's time (up to 90%) is spent in a select few key categories: social media, communication, and entertainment (Raas Cloud, n.d.).

Mobile apps are increasingly used by users for daily tasks, however processing personal data through them raises security and privacy problems. The variety of data and sensors in mobile devices, the use of different identifiers and the possibility of users being tracked, the complex mobile app ecosystem and app developers' limitations, and the extensive use of third-party software and services pose such risks. Mobile apps confront significant problems in implementing the GDPR's key data protection standards. This may prevent mobile app developers and suppliers from complying with GDPR standards on transparency and consent, data protection by design and default, and processing security (ENISA, 2017). Stakeholders engaged in the development of mobile applications must guarantee adherence to data protection obligations and the rights of users (CNIL, 2025). Identifying the privacy and data protection risks for individuals associated with mobile app usage necessitates an understanding of the processing context to identify risk sources, assess the potential impact on individuals, and recognize the constraints in implementing measures to mitigate those risks (World Intellectual Property Organization, 2021). Bian et al. reported that on average, an app collects 24 data items across 16 types, with 70% used for purposes unrelated to core functionality—most often analytics and developer advertising or marketing (Bian et al., 2022).

Organizations that create mobile applications must comply with requirements such as the GDPR and corresponding national legislative frameworks. However, within the application development process, enterprises have some flexibility in establishing internal rules and standards to improve personal data security and privacy. Such sets of ethical principles and operational rules are known as Codes of Conduct, and they are encouraged by Article 40 of the GDPR since they help to ensure its proper implementation.

The aim of this paper is to investigate current trends in research related with mobile app development, as well as to analyse practice used by mobile apps developers to secure the process of users' data maintenance. This paper addresses two research questions:

R1: What key security challenges are currently explored in mobile application development research?

R2: What mechanisms do mobile applications employ to secure user data?

The paper is structured as follows. The second section presents state of research in the domain of mobile apps using a bibliometric analysis. Third section presents codes of conduct as a self-regulation instrument that can be used in process of development of mobile apps. Fourth section introduces the most common security risks and vulnerabilities of mobile applications. Fifth section presents results of comparative analysis of three mobile application

regarding their users' data collection and usage practice.

2 State of research in domain

In order to identify the most important studies on security risks connected with mobile application we employed a publication-related metrics and co-word analysis following the steps of the bibliometric analysis procedure (Donthu et al., 2021). The articles included in the analysis were retrieved from the Scopus database. This database was chosen because it contains the most recent full-text peer-reviewed publications and database research resulted with the maximum number of articles. The following keywords were used in a keywords-based search method: ("security" AND "risk*" OR "vulnerabilities") AND ("mobile" AND "applications" OR "app*"). Articles were selected if the terms defined in the search query appeared in the titles, abstracts, or keywords. Furthermore, no time frame was specified, allowing for a comprehensive coverage of the entire discourse on topic. The literature search was conducted on 1 June 2026. The result was 546 articles presented in Fig. 1.

From 1999 to 2026, the graph displays a long-term increasing tendency in the quantity of papers, which may be broken down into three major stages. Less than five documents are produced annually during the first phase, which runs from 1999 to 2010. Thereafter, there is an increase that happens in waves, peaking in 2014 and again in 2019 with more than 40 documents. This is followed by a brief fall in 2020. The final stage is characterized by rapid, nearly exponential expansion, reaching a peak of about 85 documents in 2025. These two waves can be connected with the amendment of two important regulations, Regulation (EU) 2016/679 (GDPR) (European Parliament and the Council of the European Union, 2016) and the NIS Directive Network and Information Systems Directive) as foundational of cybersecurity legislation. The original NIS Directive was adopted in 2016. However, it was significantly updated by NIS2 (Directive (EU) 2022/2555) (European Parliament and the Council of the European Union, 2022).

In the second step, a science mapping technique is utilized to locate the most significant authors in the domain. The quantity of citations of the publications was taken into account while selecting important papers, and some of the most widely referenced articles in our study (Table 1). Based on the bibliometric data, the analyzed papers exhibit a high citation impact, with citations ranging from 465 to 847, indicating their significant influence on the mobile application research domain. The study by Chin et al. (2021) on inter-application communication in Android stands out as the most prominent work in the dataset, accumulating the highest number of citations (847).

The data reveals a strong research focus on mobile security and user privacy, which comprises four out of the five top-cited papers, while health-related mobile applications (m-Health) also represent a highly impactful niche.

Fig. 2 displays a keyword co-occurrence map of selected articles derived by bibliometric analysis in VOSViewer. In the network visualization, items are represented by their label and by default also by a circle The size of the label and the circle of an item is determined by the weight of the item. The higher the weight of an item, the larger the label and the circle of the item. The colour of an item is determined by time overlay. The focus of vulnerability analysis has matured over time. Traditional themes like 'risk assessment' and 'authentication' (blue/teal) have evolved into domains such as 'cyber-physical systems' and 'app-in-app ecosystem' security vulnerabilities, which represent the latest research frontiers heading into 2026.

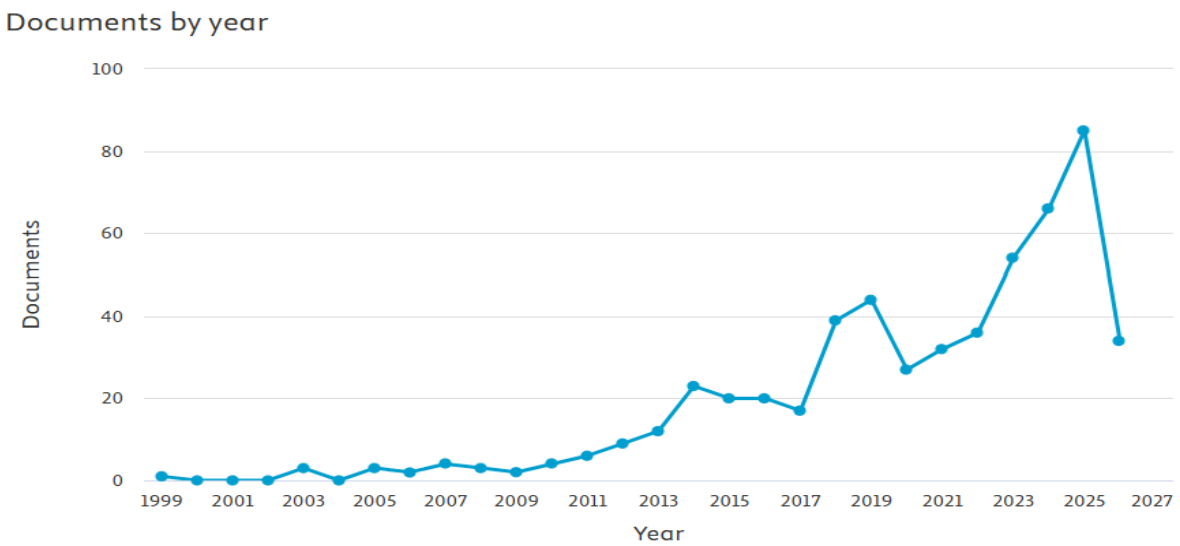


Figure 1. Number of articles by year

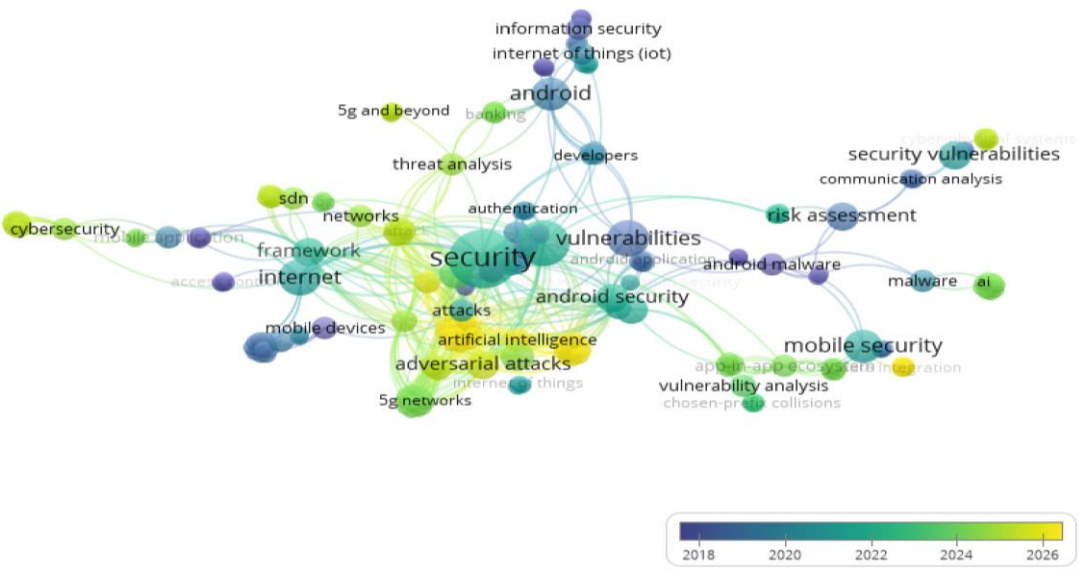


Figure 2. Keyword co-occurance network map with time overlay

Table 1. Most cited articles

Titled	Cited
Analyzing inter-application communication in Android…. (Chin et al., 2021)	847
m-Health adoption by healthcare professionals: a systematic review (Gagnon, et al. 2016)	667
RiskRanker: scalable and accurate zero-day android malware detection: …(Grace et al., 2012)	597
Android permissions: User attention, comprehension, and behavior Felt et al., 2012)	553
Device Fingerprinting in Wireless Networks: Challenges and Opportunities…(Xu et al., 2016)	465

It is evident that security issues now are connected with artificial intelligence, cybersecurity and networks (presented with yellow colour).

3 Codes of conduct and self-regulation in the development of mobile applications

Entities engaged in mobile application development must align their activities with regulations such as the GDPR and corresponding national legal frameworks. Nonetheless, during the application development process, businesses possess a degree of liberty in establishing internal procedures and standards to enhance the security and protection of collected (personal) data.

The collections of ethical principles and operational directives are referred to as Codes of conduct and are endorsed by Article 40 of the GDPR. Codes of conduct established by the profession itself are the last step in converting law into code, while national and European standards offer broad guidance. The engineering community can operationalize abstract legal principles (such as those of the GDPR) into tangible design patterns, security procedures, and programming standards on a particular platform.

An explicit example is EU Cloud Code of Conduct (EDPB, 2020), which aims to address the issue of insufficient management of personal data by cloud service providers. A significant number of Android applications depend on cloud services (such as Google Cloud) for data storage and processing, constituting a vital third-party component. If a service is inadequately reliable or improperly structured, there is a significant danger of information leakage.

Code of Conduct on Privacy for mHealth Applications (Marot, n.d.) is intended for development engineers, offering standards for secure software development while enhancing end-user trust. The code delineates the fundamental principles of privacy tailored for the mobile environment. This encompasses the requirement to secure explicit agreement from the user prior to any processing, the execution of the privacy by design principle, and the establishment of the most stringent privacy settings as default. Moreover, stringent transparency criteria and the principle of data minimization are underscored, urging developers to gather just the health data essential for the application's primary function.

ICO also defines Code of practice for app store operators and app developers (Information Commissioner's Office, 2023). This voluntary Code of Practice specifies actionable measures for App Store Operators and App Developers to safeguard users. The eight principles outlined in the Code pertain to internationally acknowledged security and privacy standards. The Joint Guidelines on the Interplay between the Digital Markets Act and the General Data Protection Regulation (EDPB, 2025) is a document collaboratively published by the European Commission and the European Data Protection Board (EDPB) in late 2025. The primary aim of these

guidelines is to describe how major technology firms (referred to as gatekeepers or access controllers, including Google, Meta, and Apple) can fulfil their competition-related obligations mandated by the DMA while remaining compliant to the stringent privacy and data protection regulations established by the GDPR.

4 Security risks and vulnerabilities of mobile applications

Security vulnerabilities most often do not arise from structural errors within the operating system itself, but rather from developers' carelessness in implementing security protocols or sophisticated methods used by attackers to deceive the end user. Threats are constantly evolving, and those who create them are continually finding new ways to cause damage or collect unauthorized data. Kaspersky Lab conducted a study on the prevalence of mobile malware in the second quarter of 2023 and the second quarter of 2025 by type of malware (Statista, 2025.). At the very beginning of the observed period, RiskTool was identified as the most prevalent threat. Although it is not a malicious program in the classical sense, it is a legitimate application whose functionality can be misused by attackers to compromise the system. AdWare (advertising malware) recorded a sharp increase in prevalence from the second to the third quarter of 2023. AdWare is most often installed with seemingly free applications, after which it displays ads in an invasive manner and secretly collects metadata about the user, such as interests and location, without their consent. The decline in the prevalence of AdWare began a trend of growth in the prevalence of banking Trojans (Trojan-Banker), its primary goal is to steal users' financial data. It most often arrives on a device through the installation of unofficial and compromised copies of popular software or as attachments in phishing emails. The main goals of this malware are to intercept sensitive information by monitoring keystrokes (keylogging), finding passwords stored within the file system, or displaying fake interfaces (overlays) for entering banking information. In addition to direct financial damage, these programs also enable the theft of the victim's identity (Point, 2026.)

A Trojan-Spy represents a specific type of malicious software whose primary goal is unauthorized monitoring of users and data collection without their consent. While the user believes that he is using a standard and harmless application, this program secretly collects a wide range of information in the background. Unlike bank Trojans that are narrowly focused on finance, the spy variant indiscriminately collects all available data – from private messages and call history to sound recording via microphone or location tracking (Threat, 2026.) In the first quarter of 2025, the share of this threat was a high 20.24%, which made it the second most represented type of malicious software. However, the latest data from the second quarter of 2025 indicate a dramatic reversal; the prevalence of this malware dropped to only 2.61%. Other malicious programs are represented to a much lesser extent, but it is significant that most of them are based on some variant of a Trojan, whether it is a Trojan-Dropper, Trojan-Downloader or Trojan-SMS distribution.

4.1 Context of security risks

The most important security threats that developers and users must deal with are identified and ranked by the OWASP Mobile Top 10 project (2026). First published in 2016, the study was updated in 2024 to take into account contemporary technical issues. Data from these two eras can be compared to gain insight into how security threats have changed and how mobile application protection has evolved. This evolution is presented directly in Fig. 3. The modern threat landscape introduces a strong emphasis on data governance and integration ecosystem vulnerabilities. Some new threats emerge such as inadequate supply chain security (M2) and inadequate privacy controls (M6) which demonstrates that contemporary mobile applications face risks far beyond local code execution. Besides that, the most important risk which emerged as completely new is the risk of improper credential usage (M1). This threat by itself can result in automated attacks if the credentials are hardcoded and identified by the attackers as well as other weaknesses caused by improper credential usage.

Furthermore, the structural reordering reveals that traditional vulnerabilities have matured, altering their critical prioritization.

Comparison Between 2016-2024		
OWASP-2016	OWASP-2024-Release	Comparison Between 2016-2024
M1: Improper Platform Usage	M1: Improper Credential Usage	New
M2: Insecure Data Storage	M2: Inadequate Supply Chain Security	New
M3: Insecure Communication	M3: Insecure Authentication / Authorization	Merged M4&M6 to M3
M4: Insecure Authentication	M4: Insufficient Input/Output Validation	New
M5: Insufficient Cryptography	M5: Insecure Communication	Moved from M3 to M5
M6: Insecure Authorization	M6: Inadequate Privacy Controls	New
M7: Client Code Quality	M7: Insufficient Binary Protections	Merged M8&M9 to M7
M8: Code Tampering	M8: Security Misconfiguration	Rewording [M10]
M9: Reverse Engineering	M9: Insecure Data Storage	Moved from M2 to M9
M10: Extraneous Functionality	M10: Insufficient Cryptography	Moved from M5 to M10

Figure 3. OWASP comparison of top 10 mobile risks in 2016 and 2024 (OWASP 2026

For instance, fundamental structural issues like insecure data storage (M2 in 2016) and insecure communication (M3 in 2016) have been downgraded to lower priority slots (M9 and M5 in 2024). This shift doesn't imply that these threats are obsolete but rather that security of operating systems have improved which changed the developer's focus primarily on authentication mechanics, credential lifecycle management and validation layers.

Additionally, the unification of older separate categories - such as merging 2016's distinct authentication and authorization threats into a singular 2024 entry (M3) which reflects a more unified approach to defending digital endpoints.

5 Case study

This section presents a qualitative case study analyzing three widely used mobile health applications on Android: BetterHelp (v3.05), MyFitnessPal (v26.27.0) and Noom (v14.26.0). These specific platforms were selected because they process highly sensitive medical and personal data which raises major privacy and security concerns. To ensure a standardized assessment, the case study utilizes a qualitative comparative content analysis combined with a platform-specific technical review. The data is gathered from two primary sources: the official Privacy Policies of the applications and their permission manifests available on the Google Play store. To ensure the empirical reproducibility of the study, all technical evaluations were conducted under a standardized protocol between June 26, 2026 and June 28, 2026. The primary testing environment was deployed in Croatia. Testing was carried out on physical reference hardware: a Xiaomi 15T Pro running an unrooted, clean instance of Android 16.

5.1 Methodoloy and evaluation criteria

The testing procedure followed a four step protocol: inspecting requested Android permissions and Google Play store disclosures, executing initial launch to log mandatory registration walls, cookie banners and profiling questionnaires prior to account setup, systematically denying requested dangerous permissions and finally navigating settings menus to test the actual availability of data control. Furthermore, this research is structured around four distinct data privacy and security criteria which are all based on previously explained testing procedure: dangerous permissions, granular permission control, third-party SDK disclosure and in-app data privacy controls. These specific criteria were established to bridge the gap between developer self-regulation and the practical security risks inherent to the Android operating system and because codes of conduct often provide theoretical frameworks for data privacy while they often lack the technical enforcement mechanisms.

The dangerous permissions criterion examines whether the application requests access to sensitive Android system resources which could reveal sensitive personal data and compromise user privacy if mismanaged. It will be graded on a three-tier qualitative scale: high, medium and low. "High" indicates that the application requests access to 2 or more highly sensitive system vectors (e.g. precise location, biometric sensors, system storage etc.). "Medium" strictly represents transactional, session-based and user-initiated dangerous permissions (e.g. runtime hardware access to camera and recording audio during active user interactions). "Low" is assigned if the application requests minimal or no system permissions beyond basic internet access.

Table 2. Results of comparative analysis

Criteria	**BetterHelp**	**MyFitnessPal**	**Noom**
Dangerous permissions	Medium	High	High
Granular permission control	Yes	Yes	Yes
Third-party SDK disclosure	Disclosed	Disclosed	Disclosed
In-app data privacy control	Yes	Yes	Partial

Granular permission control is a second criterion which examines whether the application remains fully functional after the user's denial of mentioned dangerous permissions. his criterion will be graded with a simple "yes" if core features remain stable without access to restricted resources or "no" statement if it is the opposite.

Third-party SDK disclosure criterion investigates if the application's official documentation transparently lists the integration of external SDK-s used for analytics or marketing. It will be graded either as a "disclosed" if the specific partners for telemetry and marketing were named or a "undisclosed" if they weren't named.

Finally, the in-app data privacy controls criterion evaluates whether the user interface provides an explicit option allowing users to manage their optional data sharing. This criterion is measured across three levels: yes, partial, or limited. A "yes" grade requires a dedicated, easily accessible privacy dashboard or toggle within the application interface. "Partial" indicates that options exist but are split across complex sub-menus, while "limited" means the application offers no native toggles, forcing users to rely entirely on general Android operating system settings to restrict tracking.

6 Discussion of results

After applying the established methodology and criteria to the selected applications the results were structured and summarized in Table 2. By applying a case study and comparative analysis approach, the findings reveal substantial variations in how mHealth applications use and process user data. Applications mostly behave in accordance with industry standards while some of the results show that certain things can and should be changed in order to improve user autonomy and privacy.

Upon initial execution, the BetterHelp application captures immediate cookie consent before any other user interaction. However, before it allows the user to provide an e-mail address or login credentials it gathers user data by presenting questions regarding biological age, sexual orientation, religious affiliations and relationship status among other variables. This data is extracted
in order to construct a user profile prior to the formal creation of the account. Regarding the dangerous permissions criterion the platform is graded as medium because it avoids tracking background biometric data. Only dangerous permissions that can be requested are access to camera and microphone for live video and audio therapy sessions. Granular permission control is graded as yes because users are not obliged to communicate with therapists through video calls as they can be communicating with simple text messages. Therefore, the core functionality of the application stays functional even after denying the access to camera and microphone. Furthermore, BetterHelp's third-party SDK disclosure is marked as disclosed as it is explicitly mentioned in its privacy policy which third parties can get a share of user data while some of them were even mentioned specifically with their name. Application provides in-app data privacy control and therefore that criterion is marked as yes.

MyFitnessPal application on the other hand upon its initial execution forces the creation of an account before any other activity making its features completely inaccessible without that. During the e-mail registration phase users are required to explicitly consent to the processing of sensitive personal data and their cross-border transfer to servers located in the United States. Data protection laws may differ to those in the European Union what makes this transfer a mandatory prerequisite if the user wants to create an account. Despite this rigid requirement, the platform provides an interface through which the user can choose which data can be used for the personalisation and user profiling right after the registration. Those can be set manually or simply fully accepted or rejected. Testing of "deny all" option showed no issues that compromise software performance. Consequently, granular permission control is rated as yes. Furthermore, the dangerous permissions criterion is rated as high due to its extensive integration with critical system layers and external health repositories as well as its use and process of user's biometrical data. Third-party SDK disclosure is marked as disclosed as the external entities were mentioned repeatedly and transparently in privacy policy as well as in application itself. Finally, the platform receives a yes for in-app

data privacy control as it contains a dedicated privacy center which helps users to manage their data processing consents even allowing them to view and restrict access for individual third-party vendors by specific name.

Noom application presents an intrusive user experience upon its first execution as it immediately requests runtime permission to send push notifications before any primary features can be accessed. Following this prompt, the interface provides a lengthy questionnaire as it was the case with BetterHelp but these questions are more related to the user's body weight. These questions as well serve the purpose of extracting behavioral and personal data before account creation. Noom received a high grade for dangerous permissions criterion as can request permissions to access physical activity monitoring, biometric sensors and location data. However, users can block the platform from having these permissions and still use its functionalities normally which consequently resulted with grade yes in granular permission control criterion. Third-party SDK disclosure is marked as disclosed because all of the external entities that can process data gathered by this application are named. In contrast the application fails to provide a universal in-app data privacy control framework. Testing the application on this criterion revealed its complete absence although it provides users with an option to fill out the form within Noom's documentation which serves as an opt-out form. This form appeared to be non-existent in some geographic regions and that happened in this case. Consequently, not all users can opt-out of their data being processed which resulted in a "partial" grade for the last criterion.

7 Conclusion

This study reviews current advances in mobile app development research and assesses the strategies used by developers to protect user data. We use bibliometric analysis, specifically performance analysis and scientific mapping, to uncover new trends and investigate the domain's intellectual structure.

Furthermore, the paper discusses major security threats to developers and users. A qualitative case study examines at BetterHelp, MyFitnessPal, and Noom, three popular Android health applications that handle sensitive medical and personal data, which raises privacy concerns. The findings show that, while these applications preserve fundamental functionality without requiring excessive system rights, gaps in universal in-app privacy settings persist. Ultimately, safeguarding sensitive user data requires a combination of proactive developer responibility and informed user.

Acknowledgments

This work is supported by the project "HeyAIde - Health Empowerment for You through AI Direct Engagement" co-funded by the European Union from the European Regional Development Fund.

References

Chin, E., Felt, A. P., & Wagner, D. (2021). Analyzing Inter-Application Communication in Android. *MobiSys 2021 - Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services,* (pp. 239–252.)

CNIL. (2025). *Recommendation on mobile applications*., Published on 8 April 2025, Retrieved from https://www.cnil.fr/sites/cnil/files/2025-05/recommendation-mobiles-app.pdf

Donthu, N., Kumar, S., Mukherjee, D., Pandey, N. and Lim W.M. (2021). How to conduct a bibliometric analysis: An overview and guidelines, *Journal of Business Research,* 133, 285-296.

EDPB. (2025). *Joint Guidelines on the Interplay between the Digital Markets Act and the General Data Protection Regulation*. Retrieved from https://www.edpb.europa.eu/public-consultations/joint-guidelines-on-the-interplay-between-the-digital-markets-act-and-the_en

EDPB. (2020). *EU Data Protection Code of Conduct for Cloud Service Providers*. Retrieved from https://eucoc.cloud/en/

ENISA. (2017). *Privacy and data protection in mobile applications A study on the app development ecosystem and the technical implementation of GDPR*. Retrieved from https://www.enisa.europa.eu/sites/default/files/publications/WP2017%20O-2-2-4%20GDPR%20Mobile.pdf

European Parliament and the Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 (General Data Protection Regulation). Official Journal of the European Union, L 119, pp. 1–88. Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/ bHTML/?uri=CELEX :32016R0679

European Parliament and the Council of the European Union. (2022). Directive (EU) 2022/2555 on measures for a high common level of cybersecurity across the Union (NIS 2 Directive), Official Journal of the European Union, L 333, pp. 80–152, Retrieved from https://eur-lex.europa.eu/eli/dir/2022/2555/oj

Felt, A. P., Ha, E., Egelman, S., Haney, A., Chin, E., & Wagner, D. (2012). Android permissions: User

attention, comprehension, and behavior. *SOUPS 2012 - Proceedings of the 8th Symposium on Usable Privacy and Security*. https://doi.org/10.1145/2335356.2335360

Grace, M., Zhou, Y., Zhang, Q., Zou, S., & Jiang, X. (2012). RiskRanker: Scalable and accurate zero-day android malware detection. *MobiSys'12 - Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services*, (pp. 281–293). https://doi.org/10.1145/2307636.2307663

Information Commissioner's Office. (2023). *Code of practice for app store operators and app developers (updated)*. Retrieved from https://www.gov.uk/government/publications/code-of-practice-for-app-store-operators-and-app-developers/code-of-practice-for-app-store-operators-and-app-developers-new-updated-version

Marie-Pierre Gagnon, Patrice Ngangue, Julie Payne-Gagnon, M. D. (2016). m-Health adoption by healthcare professionals: a systematic review. *Journal of the American Medical Informatics Association*, 23(1), 212–220.

Marot, N. (2016). *European Commission_ 2016_Draft Code of Conduct on privacy for mobile health applications*. 1–23. Retrieved from file://134.60.40.17/h-klips$/!Austausch/Eva/CitaviAttachments/European Commission_2016_Draft Code of Conduct on privacy for mobile health applications.pdf

OWASP. (2026). *OWASP Mobile Top 10 | OWASP Foundation*. Retrieved from https://owasp.org/www-project-mobile-top-10/

Point, C. (2026). *What is a Banking Trojan? - Check Point Software*. Retrieved from https://www.checkpoint.com/cyber-hub/cyber-security/what-is-trojan/what-is-a-banking-trojan/

Raas Cloud. (2026). *50+ Surprising Mobile App Development Statistics 2026*. 2026. Retrieved from https://raascloud.io/mobile-app-develo pment - statistics/

SensorTower. (2026). *State of Mobile 2026 The industry's leading report*. Retrieved from https://sensortower.com/report/state-of-mobile-2026/download?mkt_tok=MzUxLVJXSC0zMTUAAAGiZeg-F7h3yr_dQ4TxMaUIFVv7lBdMX0VAAtUVObcpjnMzs7SUBrITuuSpzkch7SR-ZorhTlMsNN3Dd0LyZWX9F_KLFlHwtajL4CBYRF79CBU2I3M

Statista. (2025.). *Distribution of global mobile malware by type Q2 2025| Statista*. Retrieved from https://www.statista.com/statistics/653688/distribution-of-mobile-malware-type/

Threat, K. (2026). *Trojan-Spy*. Retrieved from https://threats.kaspersky.com/en/class/Trojan-Spy/

World Intellectual Property Organization. (2021). *A Guide to Data Protection in Mobile Applications 2021*. Retrieved from https://www.wipo.int/export/sites/www/ip-development/en/agenda/pdf/wipo-guide-data-protection-mobile-apps.pdf

Xu, Q., Zheng, R., Walid, S., & Han, Z. (2016). Device Fingerprinting in Wireless Networks: Challenges and Opportunities. *IEEE Communications Surveys & Tutorials*.

Digital Transformation and Corporate Culture

Katarina Kostelić, Cristina Fleșeriu, Vlad Bocăneț
From Artificial Intelligence as Automation to Generative AI as a Work Tool: A Text Network Analysis of Public Discourse Before and After ChatGPT

Antun Biloš, Berislav Andrlić, Vladimir Ristanović
Booking Platform, Generative AI, or Open Web? An Experimental Comparison of User Experience in Tourism Accommodation Search

Anamarija Mladinic
Designing Olivia AI: User Expectations for a Trustworthy GDPR Compliance Assistant

Antonija Srok, Rikardo Protrka
Cybersecurity and Patient Data Protection in Healthcare: Research Trends and Organisational Challenges in the Context of NIS2

Matija Šajn
Generative Artificial Intelligence in Business Process Modeling: A Systematic Literature Review of Applications, Opportunities, and Limitations

From Artificial Intelligence as Automation to Generative AI as a Work Tool: A Text Network Analysis of Public Discourse Before and After ChatGPT

Katarina Kostelić
Juraj Dobrila University of Pula
Faculty of Informatics
Ulica Alda Negrija 6
52100 Pula, Croatia
katarina.kostelic@unipu.hr

Cristina Fleșeriu
Babeș-Bolyai University
Faculty of Business
Horea Street 7
400000 Cluj-Napoca, Romania
cristina.fleseriu@ubbcluj.ro

Vlad Bocăneț
Technical University of Cluj-Napoca
Faculty of Industrial Engineering, Robotics and Production Management
B-dul Muncii nr. 103-105
400000 Cluj-Napoca, Romania
vlad.bocanet@tcm.utcluj.ro

Abstract. *This paper examines how public-facing discourse on AI and work was reframed after the public release of ChatGPT in November 2022. Using text network analysis of articles from The Conversation, it shows that the debate shifted from automation-centred narratives toward a more differentiated view of generative AI as a practical work tool. The post-ChatGPT discourse connects work with education, writing, organisational adoption, productivity, copyright, language models and governance concerns. Sentiment patterns indicate thematic ambivalence. The paper contributes by demonstrating how text network analysis captures changing public meanings of AI and how work transformation is discussed across topic areas.*

Keywords. generative AI, ChatGPT, intelligent systems, text network analysis, public discourse

1 Introduction

Artificial intelligence has long been part of public debate, but the public launch of ChatGPT in November 2022 (OpenAI, 2022) created a new point of reference for discussions of generative AI. Although large generative language models had already been examined before ChatGPT, particularly in relation to the automatic production of textual and semantic artefacts (Floridi & Chiriatti, 2020), ChatGPT contributed to the wider public visibility of this class of systems by offering a conversational interface that could be used by non-specialist users. Since then, generative AI has become increasingly visible in discussions about work, education, productivity, creativity, regulation and professional knowledge (Dwivedi et al., 2023). However, it remains less clear how this visibility changed the semantic organisation of public-facing AI-and-work discourse.

The discussion is relevant because AI plays an important role in transforming work, learning, communication, and organisational processes. It also extends earlier AI-and-work debates, which often focused on automation and job displacement, toward questions of task change, augmentation and human–AI interaction (Huang & Rust, 2018; Warner & Wäger, 2019). These developments raise economic, organisational and educational questions, but also discursive ones. Public understanding of AI is shaped through media articles, expert commentary, educational explainers and research communication, which influence how intelligent technologies are interpreted in relation to work transformation across different domains.

This paper examines how public-facing discourse represents the role of artificial intelligence in transforming work processes across different domains before and after the public launch of ChatGPT. The study asks:

RQ1: How has the central vocabulary of public discourse on artificial intelligence changed in the context of transforming work processes across different domains before and after the public launch of ChatGPT?

RQ2: Which themes became structurally prominent in the post-ChatGPT discourse on AI and work transformation?

RQ3: How is sentiment distributed across thematic communities in the later stage of the discourse?

Rather than measuring the actual AI-related transformation of work processes across different domains, the study analyses how these changes are communicated, interpreted and semantically organised in expert-oriented public discourse. The study applies text network analysis, modelling words as nodes and

co-occurrences as ties to identify structurally important terms. In this sense, the study is exploratory: it maps changes in the vocabulary, thematic structure and sentiment of AI-and-work discourse.

2 Literature review

Research on AI and work has examined the implications of intelligent technologies for jobs, tasks, skills, decision-making and organisational processes. A recurring theme in AI-and-work research concerns automation, technological unemployment and the substitution of human labour, particularly in relation to robotics and machine intelligence. However, AI-and-work research also includes more differentiated perspectives. Huang and Rust (2018) argue that AI job replacement proceeds at the task level before it reaches whole occupations, while Raisch and Krakowski (2021) conceptualise automation and augmentation as interdependent dimensions of AI use in organisations. Zirar et al. (2023) further highlight worker–AI coexistence, including issues of distrust, skills, reskilling and augmentation.

Large language models and generative AI add a further dimension to this debate because they are directly relevant to cognitive, communicative and knowledge-intensive work. Studies of occupational exposure suggest that LLMs may affect a substantial share of work tasks, especially when integrated into software systems, although exposure should not be equated with direct replacement (Eloundou et al., 2024; Gmyrek et al., 2023, 2025). Empirical studies also point to heterogeneous effects. Noy and Zhang (2023) found productivity and quality gains in professional writing tasks, while Dell'Acqua et al. (2026) show that generative AI can improve performance on some tasks but reduce it on others, depending on task characteristics and the limits of the technology.

Generative AI has also become relevant beyond workplace productivity narrowly defined. ChatGPT and related systems have been discussed in relation to research, education, organisations, policy, ethics, privacy, misinformation, bias and security (Budhwar et al., 2023; Dwivedi et al., 2023). In education, studies have identified both possible benefits, such as feedback, learning support and enhanced interaction, and concerns related to academic integrity, overreliance, privacy, truthfulness and assessment security (Tlili et al., 2023; Dempere et al., 2023). Broader discussions of generative AI also point to issues of bias, stereotypes and trust in AI-generated content (Ferrara, 2023; Vaccari & Chadwick, 2020).

This shows that AI and work can be examined through several overlapping lenses: automation, task exposure, augmentation, organisational adoption, education, knowledge work, ethics and governance. However, less is known about how these issues are connected in public-facing, research-informed discourse. Much existing research examines occupational exposure, productivity effects, organisational implications, educational uses or public attitudes. Fewer studies examine how AI-and-work discourse is semantically organised, or how its structure differs before and after the public release of ChatGPT. This is the gap addressed in the present paper. Rather than measuring public opinion or the actual labour-market effects of AI, the study examines how public-facing discourse organises the meaning of AI and work before and after ChatGPT.

3 Methods

3.1 Research Design and Corpus

This study applies a text-as-data approach to examine public-facing discourse on artificial intelligence and work before and after the public launch of ChatGPT. Such methods are useful for detecting patterns across a corpus, but their results require substantive interpretation rather than fully automated reading (Grimmer & Stewart, 2013; Welbers et al., 2017).

The corpus consists of articles retrieved from The Conversation using the search query *artificial intelligence work*. This source was selected because it publishes research-based commentary written by academics and researchers for wider audiences, making it an intermediate discourse space between academic literature and general public debate. Since The Conversation allows redistribution under a Creative Commons Attribution–NoDerivatives licence, article texts were used only as input for computational analysis and results are reported in aggregate form.

The scraping procedure collected article links, titles, authors, publication dates, and article bodies. Records were retained if they contained a URL, title and non-empty article body; texts of 500 characters or fewer were excluded as too short for meaningful analysis. The final corpus contained 200 articles. Documents were divided by the public launch date of ChatGPT, 30 November 2022, into a pre-ChatGPT period (78 articles) and a post-ChatGPT period (122 articles, ending in April 2026). Texts were cleaned and tokenised into individual words. Standard English stop words were removed, while negation terms were retained. Stemming and lemmatisation were not applied in order to preserve interpretability of terms in tables, figures and community labels. The analysis was conducted in R version 4.4.2.

3.2 Co-occurrence Network Construction

Word co-occurrence networks were constructed to analyse how terms are structurally connected in the corpus. In such networks, words are represented as nodes, while ties indicate co-occurrence within a defined textual context. This approach captures relationships between terms rather than only their

individual frequencies (Danowski, 1993; Diesner & Carley, 2005; Drieger, 2013).

Each token was assigned a sequential position within its document, and word pairs were created when two terms appeared within a ten-word window. Edge weights represent the number of times two words co-occurred within this window. Separate undirected weighted networks were constructed for the pre- and post-ChatGPT periods. Only word pairs with at least ten co-occurrences were retained, and the largest connected component was extracted from each network.

The analysis reports word frequencies, normalized degree, and normalized betweenness. Normalized degree indicates the share of possible direct connections a term has, while normalized betweenness indicates the extent to which a term bridges other parts of the network (Freeman, 1978; Wasserman & Faust, 1994). In addition, thematic communities were extracted with Louvain clustering applied to the 2025 network (Blondel et al., 2008). For community and sentiment analyses, a 2025 network was used as the most recent complete post-ChatGPT calendar year. Communities were interpreted as thematic clusters, and labels were assigned by inspecting the most central and substantively meaningful words in each community.

3.3 Sentiment Analysis

Lexicon-based sentiment analysis was used to examine whether thematic communities in the 2025 network differed in affective vocabulary. The Bing lexicon was used to classify sentiment-bearing words as positive or negative (Hu & Liu, 2004; Silge & Robinson, 2017). Sentiment is treated as an indicative measure of affective vocabulary, not as a definitive measure of article tone, because of the well-known limitations (not capturing irony, negation scope, domain-specific meanings, argument structure or author stance; Liu, 2012).

4 Results

4.1 Most Frequent Terms Before and After ChatGPT

The comparison of the most frequent terms indicates a shift in the vocabulary of AI-and-work discourse before and after ChatGPT (Table 1). The pre-ChatGPT period is marked by a vocabulary associated with general AI, human intelligence, jobs, data, technology companies and future-oriented technological change. In contrast, the post-ChatGPT period gives greater visibility to terms associated with practical use, including tools, students, research, workers and tasks, alongside the emergence of ChatGPT-specific and generative AI vocabulary.

Table 1. Most frequent terms before and after ChatGPT

Rank	Before ChatGPT: word	Before: n	After ChatGPT: word	After: n
1	ai	447	ai	1654
2	not	298	not	437
3	human	254	people	321
4	data	246	data	267
5	people	232	human	267
6	intelligence	230	technology	230
7	jobs	210	research	216
8	google	202	workers	206
9	artificial	172	time	203
10	world	168	tools	182

At the level of frequent terms, this suggests a movement from AI as a technology associated with transformation and possible replacement of work toward AI as a tool discussed in relation to workflows, classrooms, writing practices, research activities and organisational routines.

4.2 Structural Changes in the Co-occurrence Networks

The co-occurrence networks provide additional insight into how AI-and-work discourse is structurally organised. As shown in Table 2, the pre-ChatGPT network is organised around a relatively concentrated semantic core. The most central terms include *ai*, *jobs*, *data*, *intelligence*, *human*, *people*, *google* and *workers*, suggesting a discourse structured around AI as a technological and labour-market issue. The presence of *jobs*, *workers*, *machine*, *robots*, *skills* and *future* indicates an emphasis on automation, employment change and technological transformation.

Table 2. Most central terms in the pre- and post-ChatGPT co-occurrence networks

Metric	Pre-ChatGPT	Post-ChatGPT
Highest normalized degree terms	ai (0.48)	ai (0.9)
	jobs (0.23)	not (0.09)
	data (0.19)	people (0.06)
	intelligence (0.18)	human (0.05)
	not (0.18)	research (0.05)
	human (0.18)	data (0.05)
	people (0.15)	time (0.04)
	google (0.12)	chatgpt (0.04)
	workers (0.12)	students (0.04)
	artificial (0.11)	workers (0.04)
Highest norm	ai (0.5)	ai (0.92)
	data (0.18)	not (0.08)
	people (0.15)	simply (0.06)

	workers (0.14)	data (0.04)
	google (0.13)	algorithms (0.04)
	jobs (0.13)	chatgpt (0.04)
	intelligence (0.11)	academic (0.04)
	job (0.09)	business (0.04)
	time (0.09)	digital (0.04)
	computer (0.07)	trained (0.04)

The post-ChatGPT network is broader and more differentiated. The degree of *ai* increases strongly, indicating that the term becomes connected to a wider range of semantic contexts. Newer structurally visible terms, including *chatgpt*, *students*, *tools*, *language*, *productivity*, *generative*, *models*, *companies*, *employees* and *decision*, suggest a discourse more closely connected to practical use, education, organisational adoption, workplace processes and decision-making.

This supports the interpretation that public-facing discourse shifted from AI as a future-oriented automation technology toward generative AI as a practical work and learning tool. The post-ChatGPT network does not replace the automation narrative entirely; instead, it embeds it within a broader set of debates about institutional adoption, human-AI collaboration, education, productivity and governance.

4.3 Thematic Communities in the 2025 Network

The 2025 co-occurrence network reveals a differentiated post-ChatGPT discourse. Louvain community detection identified 19 communities (Table 3, Fig. 1). These clusters of words are more densely connected to one another than to the rest of the network, so their interpretation combines the algorithmic grouping of terms with qualitative reading of the representative words.

Table 3. Louvain communities in the 2025 co-occurrence network

Community	Representative words	Interpretive label
1	ai, systems, employees, literacy, training, generative, job, creative	Workplace adoption and AI literacy
2	future, time, intelligence, artificial, participants, world	Future-oriented societal framing
3	students, tasks, writing, education, university, public, thinking	Education and academic writing
4	people, not, workers, college, technology, knowledge	Workers, knowledge and technological change
5	research, review, peer, consciousness, social, academic	Research, peer review and academic discourse
6	learning, bio, machine, signal, learn, skills, algorithm	Machine learning and technical systems
7	tools, found, design, chatgpt, digital, process, business, software	Tools, business processes and digital transformation
8	data, human, quantum, collaboration, video, scale, chips, judgement	Data, human judgement and technical infrastructure
9	jobs, roles, entry, level, low, positions, replace, robots, income	Labour-market automation and job restructuring
10	language, models, text, produce, word, native	Language models and text generation
11	information, google, company, million, prompt, users, active	Platform use, prompting and user adoption
12	companies, copyright, publishers, authors, industry, tech	Copyright, publishing and industry conflict
13	calls, complex, chickadees, mountain, chick, call, simple	Non-human communication theme
14	health, monitoring, care	Health monitoring and care
15	prompting, rhetorical, method	Prompting methods and interaction design
16	ability, belief, person	Human ability and belief
17	capacity, fitness	Capacity and performance
18	loss, mourned	Loss and mourning
19	reform, week, roundtable	Policy discussion

Note: Representative words are shown as returned by the community-level summary. Interpretive labels were assigned by the authors based on the most central and substantively meaningful words in each Louvain community. Smaller communities should be

interpreted cautiously because they may reflect specific article-level topics rather than broad discourse themes.

The largest and most substantively coherent communities show that AI-and-work discourse had become multidimensional by 2025. Several communities concern workplace adoption, AI literacy, business processes and digital transformation. These clusters frame generative AI as a practical tool entering organisational systems, employee training, software processes and workplace routines. This supports the interpretation that post-ChatGPT discourse increasingly treats AI as work infrastructure rather than only as a future automation technology.

A second group of communities links AI to education, academic writing and knowledge production. The presence of communities centred on students, writing, universities, research and peer review shows that the discourse extends beyond current workplaces. It also concerns the preparation of future knowledge workers and the transformation of academic practices. This is consistent with the broader visibility of ChatGPT in debates about learning, authorship, academic integrity and scholarly evaluation.

The network also retains technological and labour-market themes. Communities related to machine learning, algorithms, data, language models and text generation show that public-facing discourse still includes explanations of how AI systems work. At the same time, the labour-market automation community shows that job replacement, entry-level roles, income and robots remain part of the discussion. However, this theme appears as one community among many, rather than as the dominant organising frame.

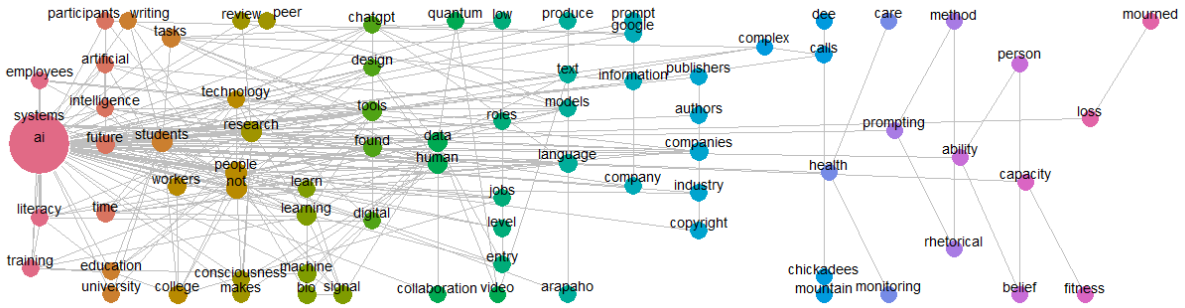


Figure 1. Thematic communities in the giant component of 2025 network

Notes: The graph shows the largest connected component of the 2025 co-occurrence network, using up to five top words per community ranked by normalized degree. For readability, the force-directed Fruchterman-Reingold layout was horizontally constrained by Louvain community; node proximity should therefore be interpreted visually rather than metrically.

Finally, several communities capture newer conflicts and practices associated with generative AI, including prompting, platform use, copyright, publishing and industry conflict. Smaller communities, such as those related to health monitoring, non-human communication or mourning, appear to reflect specific article-level topics and are therefore not considered as major discourse-level themes.

The 2025 communities show that the post-ChatGPT discourse is differentiated and multidimensional. AI is no longer discussed primarily as a general technological force or job-replacement threat. Instead, it is embedded in several interconnected debates about organisational adoption, education, knowledge work, language models, labour-market restructuring, copyright, prompting and digital transformation. This supports the broader interpretation of a shift from AI as future-oriented automation toward generative AI as an everyday work and learning tool.

4.4 Sentiment as Thematic Ambivalence

The community-level sentiment analysis for 2025 shows that AI-and-work discourse is affectively uneven across thematic areas (Fig. 2; community numbers correspond to the communities reported in Table 3). Communities related to future-oriented societal change, tools and business processes, education and academic writing, workplace adoption, AI literacy, and technical infrastructure are ambivalent, but predominantly positive.

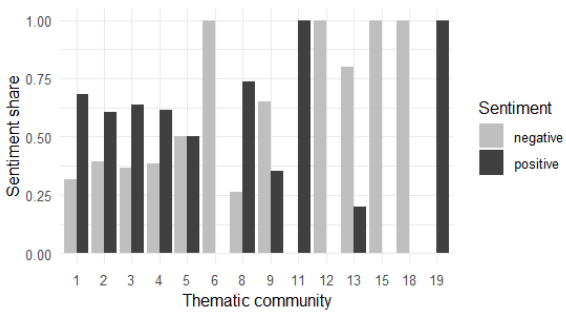


Figure 2. Sentiment distribution by 2025 thematic community

In contrast, communities related to research and peer review, workers and technological change, and smaller risk-oriented themes show more negative or more ambivalent sentiment patterns. These differences

suggest that post-ChatGPT discourse does not form a single positive or negative narrative. AI is framed more positively when discussed as a tool for innovation, productivity, education or digital transformation, and more negatively when connected to labour uncertainty, institutional tensions, legal conflict or broader social risk.

5 Discussion

The findings suggest that public-facing discourse on AI and work changed substantially after the release of ChatGPT. The change is both lexical and structural: before ChatGPT, the discourse was organised mainly around general AI, jobs, data, machines, robots and large technology companies, while the post-ChatGPT discourse became broader and more differentiated, connecting AI to tools, students, writing, productivity, language models, organisational adoption, copyright, prompting and business processes.

This supports the interpretation of a shift from AI as imagined automation toward generative AI as a practical work tool. Earlier debates on AI and work often centred on automation, job replacement and technological unemployment, while more recent research emphasises task-level effects, augmentation, skills, organisational processes and human–AI collaboration (Huang & Rust, 2018; Raisch & Krakowski, 2021; Noy & Zhang, 2023). The present findings reflect this transition: employment concerns remain visible, but they are embedded in a wider discussion about how generative AI is used, adopted, governed and integrated into everyday work.

The post-ChatGPT network also confirms the multidisciplinary and institutionally distributed character of generative AI, which previous studies have linked to research, education, organisations, policy and society more broadly (Dwivedi et al., 2023; Budhwar et al., 2023). Therefore, the public meaning of generative AI is not organised around a single question, such as whether AI will replace jobs, but across several areas in which work is performed, taught, evaluated, regulated and reorganised.

The visibility of education and academic writing is particularly relevant because generative AI became directly accessible for writing, learning, coding, summarising and explanation tasks. Previous studies similarly describe ChatGPT in education as ambivalent: it can support feedback, personalised learning and productivity, but also raises concerns about academic integrity, overreliance, privacy and critical thinking (Tlili et al., 2023; Su & Yang, 2023). The present findings show that this educational discourse is part of AI-and-work discourse, because education shapes future knowledge work and professional expectations about AI use.

The sentiment results further show that post-ChatGPT discourse is neither simply optimistic nor pessimistic. Tool-oriented, education-related, workplace adoption and digital transformation communities are predominantly positive, while communities related to labour uncertainty, research and peer review, legal conflict and social risk are more negative or ambivalent. This is consistent with literature that describes generative AI as combining productivity, business, education and research opportunities with risks related to misinformation, bias, privacy, security, worker uncertainty and ethics (Dwivedi et al., 2023; Budhwar et al., 2023; Ferrara, 2023). The contribution of this study is to show that these framings are theme-dependent rather than evenly distributed across discourse.

Methodologically, the study shows the usefulness of combining frequency analysis, co-occurrence networks, community detection and lexicon-based sentiment analysis. In this case, AI became more frequent and more embedded in a wider semantic space after ChatGPT. Community detection revealed thematic differentiation within this space, while sentiment analysis added an affective layer to the interpretation.

The findings also have practical implications. For organisations, the discourse suggests that generative AI adoption is now discussed less as an abstract automation threat and more through AI literacy, training, workflow redesign, output evaluation and responsible use. This aligns with the European policy context, where the AI Act addresses trustworthy and risk-based AI governance (Regulation (EU) 2024/1689). For educators, the results support treating AI literacy as preparation for knowledge work, while the visibility of copyright, platform use and governance themes points to the need for sector-specific guidance in concrete work settings.

The study contributes by connecting AI-and-work research with public discourse analysis. It does not measure occupational exposure, productivity effects or adoption outcomes directly, but examines how these issues become publicly connected in research-informed media discourse. This matters because public-facing discourse shapes expectations, concerns and institutional responses to emerging technologies.

The findings suggest that the public launch of ChatGPT did not simply intensify existing discourse on AI and work, but marked a turning point in the broader public visibility of generative AI. AI remains connected to automation and labour-market risk, but it is increasingly framed as a practical work tool embedded in writing, learning, research, productivity, organisational transformation and digital work.

6 Conclusion

This paper examined changes in public discourse on artificial intelligence and work before and after the public launch of ChatGPT. Using articles from The Conversation, the study applied text preprocessing, word frequency analysis, co-occurrence networks,

centrality measures, community detection and sentiment analysis.

The results suggest that post-ChatGPT discourse is broader and more differentiated than earlier AI-and-work discourse. The 2025 thematic network further shows that AI discourse had differentiated into multiple fields, including workplace adoption, education, labour-market automation, copyright, prompting and business transformation.

The main contribution of the paper is methodological and interpretive. It shows that text network analysis can reveal how public discourse is structurally organised. The results also show that sentiment is not evenly distributed across AI discourse: themes related to productivity, education and adoption are more positively framed, while themes related to labour-market displacement, institutional tensions and legal conflict are more negative or ambivalent.

The findings should be interpreted with limitations. The corpus is limited to articles from The Conversation, which is suitable for analysing research-informed public commentary but does not represent public opinion, social media, mainstream journalism or academic literature as a whole. The use of a single search query may also have shaped the corpus, and the post-ChatGPT period contains more articles than the pre-ChatGPT period. Methodologically, co-occurrence does not prove conceptual agreement, causality or author intention; community interpretation remains researcher-assigned; and lexicon-based sentiment analysis cannot fully capture irony, negation, domain-specific meanings or whether a concern is endorsed or merely reported. Some smaller 2025 communities may also reflect article-specific themes rather than broad discourse-level patterns.

Future research could expand the corpus to other media outlets, compare research-informed media with mainstream journalism or social media, and use multiple search strategies. Cross-linguistic comparisons would also be valuable, since AI adoption, labour-market concerns and educational debates may differ across national and institutional contexts. Methodologically, future studies could combine co-occurrence networks with topic modelling, transformer-based embeddings or qualitative close reading.

Acknowledgments

This work was funded by the EU NextGeneration under the Juraj Dobrila University of Pula institutional research project number IIP_UNIPU_010133. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Juraj Dobrila University of Pula.

References

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10), P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

Budhwar, P., Chowdhury, S., Wood, G., Aguinis, H., Bamber, G. J., Beltran, J. R., Boselie, P., Cooke, F. L., Decker, S., DeNisi, A., Dey, P. K., Guest, D., Knoblich, A. J., Malik, A., Paauwe, J., Papagiannidis, S., Patel, C., Pereira, V., Ren, S., Rogelberg, S., Saunders, M.N.K., Tung, R.L., Varma, A. (2023). Human resource management in the age of generative artificial intelligence: Perspectives and research directions on ChatGPT. *Human Resource Management Journal*, 33(3), 606–659. https://doi.org/10.1111/1748-8583.12524

Danowski, J. A. (1993). Network analysis of message content. *Progress in Communication Sciences*, 12, 197–221.

Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., & Lakhani, K. R. (2026). Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. *Organization Science*. https://doi.org/10.1287/orsc.2025.21838

Dempere, J., Modugu, K., Hesham, A., & Ramasamy, L. K. (2023). The impact of ChatGPT on higher education. *Frontiers in Education*, 8, 1206936. https://doi.org/10.3389/feduc.2023.1206936

Diesner, J., & Carley, K. M. (2005). Revealing social structure from texts: Meta-matrix text analysis as a novel method for network text analysis. In V. K. Narayanan & D. J. Armstrong (Eds.), *Causal mapping for research in information technology* (pp. 81–108). IGI Global.

Drieger, P. (2013). Semantic network analysis as a method for visual text analytics. *Procedia - Social and Behavioral Sciences*, 79, 4–17. https://doi.org/10.1016/j.sbspro.2013.05.053

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., et al. (2023). So what if ChatGPT wrote it? Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, 71, 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642

Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2024). GPTs are GPTs: Labor market impact potential of LLMs. *Science*, 384(6702), 1306–1308. https://doi.org/10.1126/science.adj0998

Ferrara, E. (2023). Should ChatGPT be biased? Challenges and risks of bias in large language models. *First Monday*, 28(11). https://doi.org/10.5210/fm.v28i11.13346

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30(4), 681–694. https://doi.org/10.1007/s11023-020-09548-1

Freeman, L. C. (1978). Centrality in social networks: Conceptual clarification. *Social Networks*, 1(3), 215–239. https://doi.org/10.1016/0378-8733(78)90021-7

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3), 267–297. https://doi.org/10.1093/pan/mps028

Gmyrek, P., Berg, J., & Bescond, D. (2023). Generative AI and jobs: A global analysis of potential effects on job quantity and quality (ILO Working Paper No. 96). *International Labour Organization*. https://doi.org/10.54394/FHEM8239

Gmyrek, P., Winkler, H., & Garganta, S. (2025). Generative AI and jobs: A refined global index of occupational exposure (ILO Working Paper No. 140). *International Labour Organization*. https://www.ilo.org/publications/generative-ai-and-jobs-refined-global-index-occupational-exposure

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In Kohavi, R., Gehrke, J., DuMouchel, W. and Ghosh, J (Eds), *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 168–177). Association for Computing Machinery. https://doi.org/10.1145/1014052.1014073

Huang, M.-H., & Rust, R. T. (2018). Artificial intelligence in service. *Journal of Service Research*, 21(2), 155–172. https://doi.org/10.1177/1094670517752459

Liu, B. (2012). *Sentiment analysis and opinion mining*. Morgan & Claypool.

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654), 187–192. https://doi.org/10.1126/science.adh2586

OpenAI. (2022, November 30). Introducing ChatGPT. https://openai.com/index/chatgpt/

Raisch, S., & Krakowski, S. (2021). Artificial intelligence and management: The automation–augmentation paradox. *Academy of Management Review*, 46(1), 192–210. https://doi.org/10.5465/amr.2018.0072

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations and Directives (Artificial Intelligence Act). (2024). Official Journal of the European Union, L 2024/1689. https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

Silge, J., & Robinson, D. (2017). *Text mining with R: A tidy approach*. O'Reilly Media.

Su, J., & Yang, W. (2023). Unlocking the power of ChatGPT: A framework for applying generative AI in education. *ECNU Review of Education*, 6(3), 355–366. https://doi.org/10.1177/20965311231168423

The Conversation. (2026). The Conversation. https://theconversation.com

Tlili, A., Shehata, B., Adarkwah, M. A., Bozkurt, A., Hickey, D. T., Huang, R., & Agyemang, B. (2023). What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart Learning Environments*, 10, 15. https://doi.org/10.1186/s40561-023-00237-x

Vaccari, C., & Chadwick, A. (2020). Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media + Society*, 6(1). https://doi.org/10.1177/2056305120903408

Warner, K. S. R., & Wäger, M. (2019). Building dynamic capabilities for digital transformation: An ongoing process of strategic renewal. *Long Range Planning*, 52(3), 326–349. https://doi.org/10.1016/j.lrp.2018.12.001

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge University Press.

Welbers, K., Van Atteveldt, W., & Benoit, K. (2017). Text analysis in R. *Communication Methods and Measures*, 11(4), 245–265. https://doi.org/10.1080/19312458.2017.1387238

Zirar, A., Ali, S. I., & Islam, N. (2023). Worker and workplace Artificial Intelligence (AI) coexistence: Emerging themes and research agenda. *Technovation*, 124, 102747. https://doi.org/10.1016/j.technovation.2023.102747

Booking Platform, Generative AI, or Open Web? An Experimental Comparison of User Experience in Tourism Accommodation Search

Antun Biloš
Josip Juraj Strossmayer University of Osijek
Faculty of Economics and Business
Trg Ljudevita Gaja 7
31000 Osijek, Croatia
antun.bilos@efos.hr

Berislav Andrlić
Josip Juraj Strossmayer University of Osijek
Faculty of Tourism and Rural Development
Vukovarska 17,
34000 Požega, Croatia
bandrlic@ftrr.hr

Vladimir Ristanović
Institute of European Studies
Trg Nikole Pašića 11
Belgrade 11000, Serbia
vladimir.ristanovic@ies.rs

Abstract. *The introduction of generative AI tools has disrupted established patterns of accommodation search, where online travel agencies have long held a dominant position. This study compares the user experience of three search conditions in a controlled experiment with 165 participants across Croatia, Bosnia and Herzegovina, and Serbia. Booking.com achieved the highest pragmatic UX, ChatGPT showed the highest hedonic UX descriptively, though this difference did not reach statistical significance. Decision time did not differ between conditions. Decision-making experience, operationalized as age within the student sample, moderated these effects: older participants reported markedly better experience with AI and worse experience with open-web search, while their evaluation of Booking.com remained stable.*

Keywords. generative AI, user experience, UEQ-S, online travel agencies, ChatGPT, Booking.com, accommodation search, experimental study, decision-making experience, technology adoption

1 Introduction

Online travel agencies (OTAs) such as Booking.com, Airbnb, and Expedia have established themselves as dominant intermediaries in the European hotel sector, where Booking.com alone has captured a 71% share of the European OTA market and continues to expand its dominance (HOTREC, 2024; Fernández et al., 2020; Kalinić et al., 2019). These platforms have refined their interfaces over decades, offering structured search, aggregated reviews, comparison tools, and integrated booking flows that have become the default reference for how consumers expect digital accommodation search to work.

The introduction of ChatGPT in late 2022 disrupted this standard. A generative AI interface based on natural-language dialogue offers an alternative paradigm that does not rely on filters, lists, or photographs, but instead synthesizes information through conversation (Mich & Garigliano, 2023). Early empirical work has documented growing user willingness to consult ChatGPT for travel planning tasks (Biloš & Budimir, 2024; Arora et al., 2025), while controlled comparisons with established alternatives have begun to appear (Luo et al., 2025; De Cicco et al., 2025; Nie et al., 2025). The result is a conditional rather than uniform AI advantage, with effects depending on task type, user characteristics, and design context.

Despite this growing interest, the user experience of generative AI in tourism search has not yet been systematically compared with both major OTA platforms and unstructured open-web search using standardized UX instruments. Standardized UX evaluations of western OTAs such as Booking.com remain limited, and AI-based travel assistants represent an emerging domain with almost no standardized UX evaluations to date. The present study addresses this gap through a controlled experimental comparison of three accommodation search conditions, evaluated using the short version of the User Experience Questionnaire (UEQ-S; Schrepp et al., 2017) alongside behavioral and self-report outcomes. To the authors' knowledge, this is among the first experimental studies to compare an established booking platform, a generative AI interface, and restricted open-web search within the same accommodation-search task using a standardized UX instrument.

Beyond the descriptive comparison, the study examines a moderation mechanism supported by empirical evidence in related research but not yet tested experimentally in this domain. Recent work has shown that experience moderates the predictors of AI-tool

adoption intention; less experienced users rely more on habit and social influence, whereas more experienced users develop more differentiated evaluations (Biloš et al., 2025). Building directly on this finding, the present study extends the moderation logic from self-reported intention to actual user experience in a realistic decision-making task. The hypothesis under test is that decision-making experience, operationalized as age within a student population, reshapes how users evaluate generative AI relative to established alternatives.

The study was conducted with 165 participants across three countries (Croatia, Bosnia and Herzegovina, Serbia) using a standardized accommodation search scenario in Graz, Austria. The remainder of the paper is organized as follows. Section 2 develops the theoretical background and presents the hypotheses. Section 3 describes the experimental design, instrument, sample, and analytical strategy. Section 4 reports the empirical findings, including the moderation analysis and qualitative illustrations. Section 5 discusses theoretical and practical implications and outlines limitations and directions for future research.

2 Theoretical Background

2.1 User Experience and the UEQ-S Framework

User experience (UX) is defined by ISO 9241-210 (2019) as a person's perceptions and responses resulting from the use or anticipated use of a product, system, or service. This conceptualization deliberately spans pragmatic dimensions related to instrumental task performance and hedonic dimensions related to subjective enjoyment and emotional engagement. The User Experience Questionnaire and its short version (UEQ-S) operationalize this distinction through eight semantic differential items that yield two scale scores, Pragmatic Quality and Hedonic Quality, alongside an aggregate score that summarizes overall UX (Schrepp et al., 2017). Recent cross-cultural adaptations of the instrument confirm its applicability across linguistic contexts when the underlying factor structure is preserved (Melissourgos et al., 2025; Juric et al., 2024).

The pragmatic-hedonic distinction is theoretically valuable for comparing digital tools because the two dimensions can be dissociated. A tool may be pragmatically efficient yet hedonically dull, or hedonically engaging yet pragmatically frustrating. Standardized instruments such as UEQ-S have been increasingly applied to consumer-facing digital interfaces (Islami et al., 2023; Jung et al., 2026), although standardized UX evaluations of western online travel agencies and AI-based travel assistants remain limited.

2.2 Generative AI and Established Tools in Consumer Decision-Making

Travel decisions have long been mediated by online channels. Online travel agencies (OTAs) have established a dominant position in accommodation distribution, accounting for over 90% of online hotel reservations in some markets, with Booking.com leading the segment (Fernández et al., 2020; Kalinić et al., 2019). Empirical research on consumer behavior within OTA platforms has identified price, ratings, search position, and informational completeness as key drivers of conversion across the search and booking funnel (Xie et al., 2019; Pinto et al., 2019). The instrumental quality of OTA interfaces, including filters, photographs, and aggregated reviews, has been refined over decades of iterative design.

The introduction of ChatGPT in late 2022 changed the landscape by offering an alternative interaction paradigm based on natural-language dialogue rather than structured filters and lists. Conceptual work mapped the technological potential of large language models for tourism early on (Mich & Garigliano, 2023), and a rapidly expanding empirical literature has begun to examine their actual use. Survey-based research grounded in extended versions of the Technology Acceptance Model (Davis, 1989) and UTAUT2 (Venkatesh, Thong, & Xu, 2012) has identified habit, performance expectancy, and hedonic motivation as the strongest predictors of behavioral intention to use ChatGPT among Generation Z users (Biloš & Budimir, 2024; Arora et al., 2025). Performance expectancy and effort expectancy show inconsistent effects across this literature, suggesting that ease of use is no longer a meaningful differentiator between AI and conventional digital tools.

Experimental work has begun to examine how generative AI compares with established alternatives. Luo et al. (2025) manipulated search type, travel motive, and AI customization level and found that the relative advantage of generative AI over conventional search engines is conditional rather than uniform. De Cicco et al. (2025) compared AI and human advisors across pre-visit and on-site travel stages and reported stage-dependent preferences. Studies on trust have shown that hallucinations and inaccuracies in AI output reduce perceived accuracy and willingness to follow AI recommendations (Kim et al., 2023; Christensen et al., 2025). At the same time, evidence from age-segmented evaluations of LLM-based travel assistants suggests that acceptability and perceived credibility vary across user groups (Nie et al., 2025). The current literature points toward a tool whose UX value is contingent rather than universal, with the contingency factors only beginning to be mapped.

2.3 Decision-Making Experience as a Moderator

A long tradition in behavioral decision research has examined how accumulated experience moderates decision-makers' use of information aids. Older adults with greater consumer experience adapt their search strategies in complex consumer choice tasks and maintain decision quality through accrued domain knowledge (Queen et al., 2012). Prior domain knowledge similarly reduces age-related differences in web search performance (Sanchiz et al., 2017) and interacts with interface design in three-way patterns where experienced users in supportive interface conditions can match or exceed less experienced peers (Zimmer et al., 2015). The dual-process distinction between automatic and deliberative cognition (Kahneman, 2011) provides a theoretical mechanism: experienced users have richer mental representations of available tools and evaluate them through more discriminating, comparative processing, while less experienced users default to undifferentiated heuristic evaluation.

Generational labels such as "digital natives" have been challenged empirically. Hoffmann, Lutz, and Meckel (2014) showed that web experience and education explain differences in online trust more reliably than birth cohort, and cross-national evidence indicates that technology skill and use are distributed on a continuum rather than dichotomized by generation (Bullen & Morgan, 2011). For research on generative AI specifically, this distinction matters because it implies that observed age effects may reflect accumulated decision-making experience rather than generational technological orientation.

Within technology adoption frameworks, experience operates as a formal moderator (Venkatesh et al., 2012) that reshapes which predictors translate into intention or use. Habit, defined as the extent to which behavior becomes automatic through repeated learning (Limayem, Hirt, & Cheung, 2007), accumulates with experience and has been identified as the single strongest predictor of ChatGPT adoption intention in regional Generation Z populations (Biloš & Budimir, 2024). Building directly on this evidence, Biloš et al. (2025) demonstrated that experience itself moderates the predictors of ChatGPT usage intention: less experienced users rely more heavily on habit and social influence, while more experienced users develop differentiated patterns of evaluation. The mechanism documented at the level of behavioral intention should plausibly extend to the level of actual user experience during real interactions with the tool.

2.4 Research Hypotheses

Based on the literature reviewed above, the following hypotheses guide the analysis:

H1. Tool type affects the pragmatic dimension of user experience, with the established Booking platform achieving higher UEQ-S Pragmatic scores than open-web search.

H2. Tool type affects the hedonic dimension of user experience, with the generative AI condition achieving higher UEQ-S Hedonic scores than the established Booking platform.

H3. Decision-making experience, operationalized as age within the student population, moderates the effect of tool condition on user experience.

A secondary research question (RQ1) examines whether tool conditions differ in objective and behavioral outcomes, namely decision time, chosen accommodation price, and stated booking intention.

3 Method

3.1 Design and Participants

The study employed a between-subjects experimental design comparing three accommodation search conditions: (E1) booking through the Booking.com platform, (E2) searching with the ChatGPT conversational interface, and (K) open-web search excluding both online travel agencies and large language model tools. The factorial structure additionally included country (Croatia, Bosnia and Herzegovina, Serbia) and gender as design factors with balanced cell sizes targeted in advance.

Participants were recruited from the student populations of several universities in Croatia, Bosnia and Herzegovina, and Serbia. The final analytic sample included 165 participants (49 from Croatia, 59 from Bosnia and Herzegovina, 57 from Serbia), with a balanced gender distribution (83 men, 82 women) and a mean age of 24.7 years (SD = 5.92, range 19 to 54). The educational profile reflected the recruitment context, with 55.2% reporting completed secondary education, 22.4% completed undergraduate studies, 16.4% completed graduate studies, and the remaining participants distributed across postgraduate and other categories. Allocation to experimental conditions was implemented at the level of small-group sessions. Participation was voluntary and anonymous, with informed consent obtained on the first page of the survey instrument. The study was conducted in accordance with general research ethics principles for studies involving human participants.

3.2 Procedure and Task

All participants received an identical task scenario presented in their native language. Participants were asked to imagine planning a short trip and to find one accommodation option that they would realistically reserve. Task parameters were fixed across all conditions to control for choice complexity: destination Graz (Austria), two adults, two nights from Friday 17 April to Sunday 19 April 2026, maximum total budget of 300 EUR for both nights, location

preferably within 3 km of Hauptplatz, and accommodation type restricted to hotel rooms, private rooms, studios, or apartments (excluding hostel dormitories). Participants were instructed not to complete an actual reservation; the task ended when they indicated their final choice. The protocol communicated a guideline of approximately 15 minutes for task completion, which was not strictly enforced. In practice, 98.8% of participants completed the task within the 15-minute window. The full sample showed a mean decision time of 215.4 seconds (Mdn = 177.0 s, range 21 to 986 s).

The experimental manipulation operated through tool restriction. Participants in E1 used Booking.com exclusively. Participants in E2 used ChatGPT exclusively, remaining within the conversational interface and not opening external websites in separate tabs. Participants in K used open-web search through a search engine and direct websites of accommodation providers or tourism organizations, while explicitly avoiding online travel agencies and large language model interfaces. A minimum decision time of 20 seconds was established as a data quality criterion based on observations from pilot testing, where shorter completions consistently reflected disengaged or hurried responding rather than genuine task performance. Responses below this threshold were excluded prior to analysis. Protocol adherence was monitored through a separate deviations log completed by the researcher after each session, designed to preserve the anonymity of participant responses while documenting any procedural irregularities. The number of recorded deviations was small and did not show a systematic pattern across conditions. The analytic sample of 165 participants reflects the dataset following these data quality checks.

3.3 Measures

Decision time was recorded automatically by the survey platform (Alchemer) as the duration spent on the task page, from the moment the task was displayed until participants clicked the corresponding button to indicate they had made their decision.

User experience was measured using the short version of the User Experience Questionnaire (UEQ-S; Schrepp et al., 2017). The UEQ-S comprises eight items in semantic differential format on a seven-point scale, organized into two dimensions of four items each: Pragmatic Quality (obstructive/supportive, complicated/easy, inefficient/efficient, confusing/clear) and Hedonic Quality (boring/exciting, not interesting/interesting, conventional/inventive, usual/leading-edge). Following standard UEQ practice, raw responses were transformed to the official metric ranging from -3 to +3 by subtracting 4 from each item score, with values above +0.8 indicating positive evaluations and values below -0.8 indicating negative evaluations. The instrument was administered in the language version corresponding to each participant's country of residence. The Croatian items followed the order of the original UEQ-S list with adjustments for cross-country comparability with Bosnian and Serbian versions, departing in some pair formulations from the validated Croatian UEQ version (Juric et al., 2024). Cross-cultural adaptation of UEQ-S for related linguistic contexts has recently been reported by Melissourgos et al. (2025), supporting the legitimacy of regional adaptations when the underlying factor structure is preserved.

Behavioral and evaluative outcomes following the task included: total price of the chosen accommodation in euros (open numeric entry), accommodation type, booking intention measured as the perceived likelihood of actually reserving the chosen accommodation (1 = not at all likely, 7 = extremely likely), perceived task ease (1 = very difficult, 7 = very easy), and prior familiarity with the destination (1 = not familiar at all, 5 = very familiar). An open-ended question invited participants to comment on what helped or hindered their decision.

3.4 Analytical Strategy

Statistical analyses were conducted using SPSS 29.0.2.0 and Microsoft Excel. Analyses proceeded in three stages. First, between-condition differences on the seven primary outcome variables were tested using one-way ANOVA, with effect sizes reported as omega squared (ω^2) to provide unbiased estimates given unequal cell sizes. Second, the central hypothesis on the moderating role of decision-making experience was tested using hierarchical linear regression, with age entered as a continuous moderator interacting with experimental condition (dummy-coded with K as reference). Median split visualizations are used for illustrative purposes only, while the statistical inference relies on the continuous specification. Third, robustness across the three countries was examined by re-estimating the central interaction model with country dummies and country × condition terms, following recent recommendations in cross-cultural OTA research (Filieri et al., 2021). Open-ended responses (v09) were thematically reviewed to identify illustrative quotations supporting the quantitative findings. Levene's test indicated unequal group variances for two outcomes (task ease and UEQ-S Pragmatic Quality); Welch's robust ANOVA yielded equivalent conclusions for these variables (Welch's $F = 15.06$ and 5.88, both $p < .01$), and standard ANOVA results are reported throughout.

Generative AI tools (Claude Opus 4.7, ChatGPT 5.5, Gemini 3.1 Pro) and the literature discovery tool SciSpace were consulted at various stages for support tasks such as study design refinement, literature mapping, and drafting iterations of the manuscript. Grammarly was used for language editing. All scientific decisions, statistical analyses, interpretations, and final wording remained the responsibility of the authors.

4 Results

4.1 Main Effects of Tool Type

Table 1 reports descriptive statistics and one-way ANOVA results for the seven primary outcome variables across the three experimental conditions.

Table 1. Means and one-way ANOVA results across tool conditions

Variable	Booking	ChatGPT	Open-web	F (p)	ω²
Decision time (s)	211.3 (197.6)	212.8 (133.6)	223.4 (173.5)	0.08 (.925)	< .001
Price (€)	203.7 (62.4)	214.4 (52.0)	217.7 (51.7)	1.00 (.370)	< .001
Booking intention (1–7)	5.52 (1.16)	5.15 (1.22)	4.80 (1.48)	4.59 (.012)	.042
Task ease (1–7)	6.40 (0.91)	5.90 (1.45)	5.28 (1.23)	12.91 (<.001)	.126
UEQ-S Pragmatic (–3 to +3)	1.84 (1.07)	1.61 (1.42)	1.03 (1.38)	5.90 (.003)	.056
UEQ-S Hedonic (–3 to +3)	0.76 (1.27)	1.22 (0.92)	0.77 (1.03)	2.99 (.053)	.024
UEQ-S Overall (–3 to +3)	1.30 (1.02)	1.42 (1.01)	0.90 (1.08)	3.51 (.032)	.030

Note. Booking n = 67, ChatGPT n = 48, Open-web n = 50. Means with standard deviations in parentheses. F values report F(2, 162). Effect sizes (ω²) reported as <.001 indicate negligible effects.

The conditions differed substantially on subjective task ease (ω² = .126) and on UEQ-S Pragmatic Quality (ω² = .056). Booking achieved the highest scores on both measures, with Open-web search performing the weakest (Fig. 1). Booking intention followed the same ordering, although with a smaller effect size (ω² = .042). The pattern on UEQ-S Hedonic Quality reversed the ordering on the pragmatic dimension, with ChatGPT scoring highest (M = 1.22) and Booking lowest (M = 0.76). On the composite UEQ-S Overall score, ChatGPT slightly outperformed Booking, with Open-web again being the lowest. Tukey HSD post-hoc tests confirmed that Booking significantly exceeded Open-web on UEQ-S Pragmatic Quality (mean difference = 0.81, p = .003). While ChatGPT showed a higher mean on UEQ-S Hedonic Quality than Booking (mean difference = 0.47), this difference did not reach statistical significance (Tukey HSD p = .069).

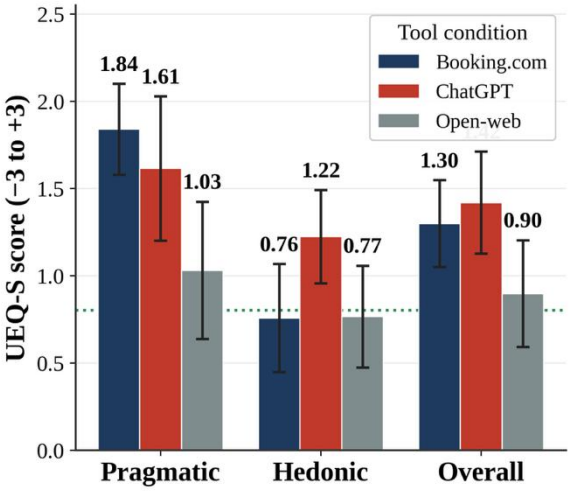


Figure 1. UEQ-S Pragmatic, Hedonic, and Overall scores across tool conditions, with 95% confidence intervals. The dotted reference line marks the +0.8 benchmark for positive evaluations.

Two outcomes showed no statistically reliable variation between tools. The price of the chosen accommodation was similar across conditions, with means ranging from 204 to 218 EUR, well below the 300 EUR task limit. Decision time was likewise indistinguishable between tools (means ranging from 211 to 223 seconds), with ω² approaching zero. Given the right-skewed distribution of raw decision time, the analysis was repeated on log-transformed values; the conclusion remained unchanged (F(2,162) = 0.25, p = .781).

4.2 The Fluency-Time Disconnect

The combined pattern in Table 2 reveals a notable disconnect between objective and subjective task indicators (Fig. 2). Although participants spent virtually identical amounts of time on the task across conditions, their subjective evaluations of task ease differed by more than 1 full-scale point on a 7-point scale (Booking: 6.40; ChatGPT: 5.90; Open-web: 5.28). The within-sample correlation between decision time and perceived task ease was negligible (r = -.08, p = .288), confirming that the time spent on the task carried little information about how the experience felt to participants. Tool conditions did not differ in how long they took, but they clearly differed in how easy and pragmatically well-functioning they were perceived to be. This disconnect motivates the central question that follows: if subjective experience varies independently of objective effort, what individual-level factors shape the magnitude and direction of these subjective evaluations?

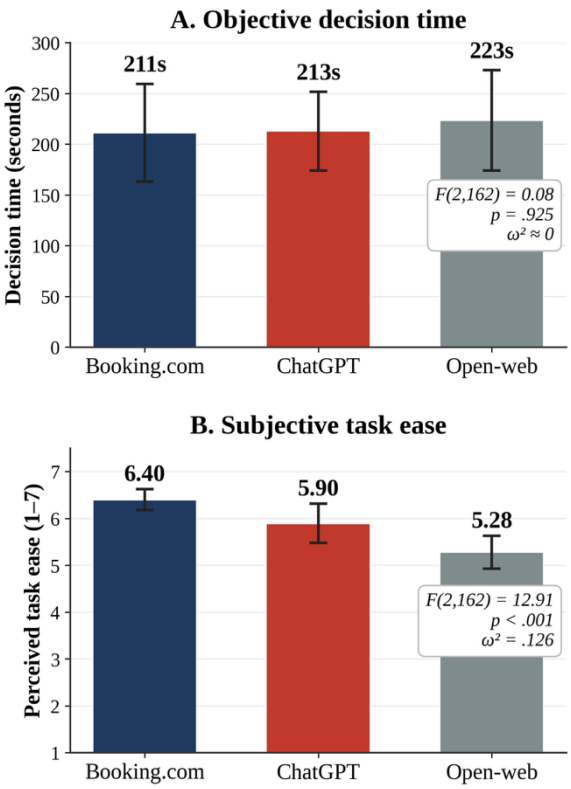


Figure 2. The fluency-time disconnect. Panel A shows decision time in seconds. Panel B shows perceived task ease on a 1–7 scale. Error bars represent 95% confidence intervals.

4.3 Decision-Making Experience as a Moderator

The central hypothesis predicted that decision-making experience would moderate the effect of tool condition on user experience. Hierarchical regression analyses tested this prediction on the two UEQ-S outcomes that showed reliable between-tool differences: Pragmatic Quality and Overall UX. Age was used as a continuous proxy for accumulated decision-making experience and was mean-centered before model estimation. Tool conditions were dummy-coded with Open-web (K) as the reference category. Gender was retained as a covariate. Because Open-web was used as the reference condition, the age coefficient represents the age slope in the restricted open-web condition, while the interaction terms indicate how much the Booking and ChatGPT age slopes differ from that reference slope. Thus, a positive ChatGPT × Age interaction means that the association between age and UX was more positive in the ChatGPT condition than in the restricted open-web search.

Table 2. Hierarchical regression predicting UEQ-S Pragmatic Quality and UEQ-S Overall

Step / Predictor	B	SE	t	B	SE	t
	UEQ-S Pragmatic			*UEQ-S Overall*		
Step 1: Controls						
Age (centered)	.02	.02	0.92	.01	.01	0.79
Gender (Male=1)	-.56	.21	-2.73**	-.29	.17	-1.76
R^2	**.045**			**.020**		
Step 2: Main effects (ΔR^2)	**.063 ****			**.041 ***		
Booking (vs Open-web)	.78	.24	3.30**	.38	.19	1.98*
ChatGPT (vs Open-web)	.58	.26	2.27*	.53	.21	2.53*
Step 3: Tool × Age (ΔR^2)	**.087 *****			**.076 ****		
Booking × Age	.12	.04	2.99**	.09	.03	2.64**
ChatGPT × Age	.21	.05	4.06***	.16	.04	3.67***

Note. * p < .05, ** p < .01, *** p < .001

The interaction terms made a substantial contribution to the model in both cases. For UEQ-S Pragmatic, adding the two Tool × Age interactions increased explained variance from 10.8% to 19.5%, with both interaction terms reaching statistical significance, and the ChatGPT × Age slope showing the strongest effect in the model (B = .21, t = 4.06, p < .001). The pattern was parallel for UEQ-S Overall (final R^2 = .137).

To interpret the interactions, simple slopes of age were estimated within each tool condition, controlling for gender (Fig. 3). In the Booking condition, age was unrelated to UEQ-S Pragmatic (B = +.02, p = .207). In the ChatGPT condition, age showed a strong positive slope (B = +.12, t = 2.76, p = .008), indicating that older participants reported markedly better pragmatic UX than younger ones. In the Open-web condition, the slope reversed sign and reached statistical significance in the opposite direction (B = -.10, t = -2.48, p = .017), with older participants reporting worse pragmatic UX than younger ones. The same crossover pattern held for UEQ-S Overall (Booking B = +.02, p = .256; ChatGPT B = +.09, p = .009; Open-web B = -.07, p = .037).

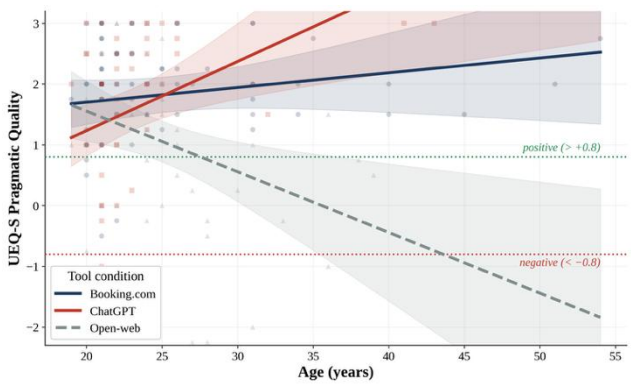


Figure 3. Tool × Age interaction on UEQ-S Pragmatic Quality. Lines represent regression-

predicted scores at the sample mean for gender, with shaded bands indicating 95% confidence intervals.

In summary, Booking provided a stable user experience across the age range observed in the sample, with younger and older participants converging on the same evaluation. ChatGPT and Open-web search behaved in the opposite manner: as accumulated decision-making experience increased, ChatGPT was evaluated more favorably while Open-web was evaluated less favorably. The effect of tool condition on user experience depended systematically on participant experience rather than operating uniformly across the sample.

4.4 Robustness Across Countries

To assess the generality of the moderation pattern, the Tool × Age regression was re-estimated separately for each national subsample. A statistically significant Tool × Age interaction was reproduced in each country, although the specific term reaching significance varied: ChatGPT × Age in Croatia (B = +0.24, p = .011) and Serbia (B = +0.73, p = .007), and Booking × Age in Bosnia and Herzegovina (B = +0.12, p = .030). Cell sizes ranged from 14 to 28 participants per condition × country, limiting statistical power for individual subsample tests, and the magnitude of the Serbian coefficient should be interpreted with caution given the small subsample size. The cross-country replication of the directional pattern is consistent with the pooled finding.

4.5 Qualitative Illustrations

Open-ended responses were thematically reviewed to identify quotations that illustrated the quantitative pattern (quotations were translated into English by the authors). Three themes emerged that map directly onto the regression findings.

First, participants in the Open-web condition described systematic difficulty avoiding the gravitational pull of the Booking platform, even when explicitly instructed to do so. Comments such as "Booking kept getting in my way" (Serbia, age 23) and "I had Booking offers showing up although I was not allowed to use them" (Serbia, age 25) suggest that the Open-web condition was not simply less efficient but actively effortful, which is consistent with its lowest pragmatic UX score and the negative age slope on this outcome.

Second, the youngest participants in the ChatGPT condition expressed explicit distrust of generative AI. Statements such as "I always doubt the results of AI tools and would rather check accommodations manually" (Bosnia and Herzegovina, age 22) and "The accuracy of LLM information is questionable to me in general" (Croatia, age 21) align with the lower UX evaluations among younger participants in this condition. By contrast, slightly older participants in the same condition described positive engagement with the conversational interface, including formulations such as "ChatGPT was a piece of cake to use, that is how I will look for accommodation in the future" (Serbia, age 24).

Third, one participant explicitly noted that ChatGPT proposed only a narrow subset of accommodation types, which echoes the descriptive observation that hotels dominated the chosen options in the ChatGPT condition. This kind of constraint on the choice space, although not the focus of the present analysis, supports the broader interpretation that each tool reshapes not only how users feel about the search but also what they actually consider.

5 Discussion

5.1 Synthesis of Findings

Three findings stand out from the experimental comparison. The pragmatic dimension of user experience varied substantially across tools, with Booking achieving the highest scores and open-web search the lowest, while ChatGPT fell between the two. The hedonic dimension reversed this ordering, with ChatGPT scoring highest. Decision time showed virtually no variation between tools despite large differences in subjective task ease, which produced a near-zero correlation between objective time spent and perceived ease. Most centrally for the contribution of this paper, decision-making experience moderated the effect of tool condition on UX in the predicted direction. Older participants reported markedly better pragmatic UX with ChatGPT and markedly worse pragmatic UX with open-web search than younger participants, while their evaluation of Booking remained stable across the age range observed in the sample. The same crossover pattern was reproduced in UEQ-S Overall and maintained the same direction across the three national subsamples. Taken together, these findings provide clear support for H1 and H3. H2 was not supported. Although ChatGPT showed a higher mean on UEQ-S Hedonic Quality than Booking.com, the difference did not reach statistical significance (Tukey HSD p = .069). This suggests that the hedonic appeal of generative AI should not be overstated. In this task context, accommodation search remained strongly instrumental, and participants may have evaluated enjoyment through the lens of decision confidence rather than novelty alone.

5.2 Theoretical Contributions

The moderation finding extends the recent demonstration that experience moderates the predictors of generative AI adoption intention (Biloš et al., 2025) into the domain of actual user experience. The earlier survey-based work showed that less experienced users rely more heavily on habit and social influence in

forming their adoption intentions. The present experimental results indicate a parallel pattern in real interaction: less experienced users showed relatively flat UX evaluations across tool conditions, whereas more experienced users discriminated sharply between AI, the established OTA, and unstructured open-web search. The convergence of self-reported intention and behavioral UX evidence strengthens the case for experience as a fundamental moderator of how users translate AI capabilities into perceived value. One possible explanation is that relatively older participants within the student sample benefited more from ChatGPT because the conversational interface reduced the need to manage fragmented search results, compare multiple sources manually, and avoid dominant OTA links. In contrast, restricted open-web search may have imposed a greater coordination burden, especially because participants had to search the web while deliberately avoiding the platforms that typically structure accommodation searches.

The fluency-time disconnect carries both methodological and substantive implications for UX research. Tools that produced very different subjective experiences required statistically indistinguishable amounts of time to complete the same task. Inferring tool efficiency from objective time measurements, or inferring user experience from objective time savings, would have produced misleading conclusions in this dataset. The result is consistent with broader work on processing fluency, which has shown that subjective ease can be experimentally dissociated from objective performance. For comparative evaluations of AI and conventional digital tools, this dissociation suggests that subjective UX instruments and objective performance measures address complementary rather than overlapping questions.

A third contribution concerns the position of generative AI in the existing ecosystem. The pattern of pragmatic-hedonic dissociation observed in our data is consistent with prior UTAUT2 evidence (Biloš & Budimir, 2024): hedonic motivation has been identified as a key driver of ChatGPT adoption, although our test of the corresponding hedonic UX advantage in actual interaction did not reach statistical significance, the descriptive pattern was directionally consistent with this prediction. At the same time, established OTA platforms retain a clear pragmatic advantage rooted in decades of refinement, including filters, photos, and reviews. The empirical question raised by Mich and Garigliano (2023) regarding whether AI will replace OTA platforms therefore receives a nuanced answer in the present data: it depends on which UX dimension is at stake and on which segment of users is considered. The results therefore suggest that Booking.com and ChatGPT should not be interpreted as direct substitutes on a single UX continuum: Booking.com currently offers stronger pragmatic assurance, while ChatGPT offers a more engaging but not statistically superior hedonic experience.

5.3 Practical Implications

For tourism platforms, the findings suggest two priorities. First, OTA platforms retain a clear advantage on pragmatic dimensions and should defend this advantage by continuing to invest in structured affordances rather than imitating AI-style interaction. Second, the segment of users for whom generative AI is most attractive is not the youngest population but those with somewhat greater accumulated decision-making experience. Marketing and integration strategies for AI-augmented search should reflect this segmentation rather than assume uniform appeal across age cohorts. For AI tool designers, the disconnect between subjective ease and objective time indicates that hedonic engagement does not automatically translate into productivity gains, and that pragmatic features such as visual content, link validity, and filtering remain necessary complements to conversational fluency.

5.4 Limitations

Several limitations should be kept in mind when interpreting these results. The sample was drawn from student populations across three countries, and the age range observed (19 to 54 years, with most participants under 30) means that the observed moderation describes the transition from undergraduate to early professional life rather than a full generational comparison. Therefore, the findings should not be generalized to the full population of accommodation bookers, older tourists, or less digitally active users. Rather, they describe how relatively young and educated users evaluate competing digital search tools under controlled task conditions. The study used a single destination, a single accommodation task, and a single price ceiling, which may limit the external validity of specific descriptive findings such as the dominance of hotels in chosen accommodations. The cross-sectional design does not allow conclusions about how repeated use would change the observed UX patterns. ChatGPT was used in a single configuration, and rapid iteration of generative AI tools means that specific findings may date faster than the underlying theoretical pattern.

5.5 Future Research

Three directions appear particularly promising. First, replication on non-student populations with broader age ranges would test whether the moderation pattern generalizes beyond the early professional segment. Second, longitudinal designs that track UX over repeated use would show whether the dissociation between objective time and subjective experience persists or diminishes as users build mental models of the tools. Third, comparative experiments with additional generative AI interfaces (Gemini, Claude, Perplexity) and different tourism task types (route

planning, local activities, multi-stop itineraries) would clarify whether the patterns observed for ChatGPT in accommodation search reflect properties of the tool category or specific design choices of one platform.

6 Conclusion

The user experience of generative AI in accommodation search is neither uniformly superior nor inferior to established alternatives, but rather depends on the UX dimension examined and the user's prior decision-making experience. ChatGPT showed a directional advantage on hedonic UX, while Booking.com retains a clear pragmatic advantage built over decades of interface refinement. The moderation pattern suggests that AI integration in tourism platforms should consider user segmentation rather than assume uniform appeal. The dissociation between perceived ease and objective time spent further indicates that comparative evaluations of AI and conventional digital tools require both UX instruments and behavioral measures, since the two address complementary questions.

Acknowledgments

This research is funded by the Croatian Science Foundation (HRZZ) research project "Tourism 5.0: Customer Satisfaction through the Digital Transformation of Tourist Destinations", grant no. IP-2025-02-1899 (Croatian Science Foundation).

References

Arora, N., Manchanda, P., Aggarwal, A., & Maggo, V. (2025). Tapping generative AI capabilities: A study to examine continued intention to use ChatGPT in the travel planning. Asia Pacific Journal of Tourism Research, 30(10), 1303–1322. https://doi.org/10.1080/10941665.2024.2405134

Biloš, A., & Budimir, B. (2024). Understanding the adoption dynamics of ChatGPT among Generation Z: Insights from a modified UTAUT2 model. Journal of Theoretical and Applied Electronic Commerce Research, 19(2), 863–879. https://doi.org/10.3390/jtaer19020045

Biloš, A., Turkalj, D., & Kelić, I. (2025). The role of experience in AI adoption: The moderating effect on predictors of ChatGPT usage intention among Generation Z. In Proceedings of the 9th FEB International Scientific Conference (pp. 289–298). University of Maribor Press. https://doi.org/10.18690/um.epf.5.2025.27

Bullen, M., & Morgan, T. (2011). Digital learners not digital natives. La Cuestión Universitaria, 7, 60–68.

Christensen, J., Hansen, J. M., & Wilson, P. (2025). Understanding the role and impact of Generative Artificial Intelligence (AI) hallucination within consumers' tourism decision-making processes. Current Issues in Tourism, 28(4), 545–560. https://doi.org/10.1080/13683500.2023.2300032

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly, 13(3), 319–340.

De Cicco, R., Francioni, B., Dini, M., Curina, I., Filieri, R., & Cioppi, M. (2025). Generative AI and human advice: User perceptions, intentions, and behavior across pre-visit and on-site travel stages. Journal of Travel Research. https://doi.org/10.1177/00472875251364222

Fernández, J. M., Vall-Llosera, L., & Moya, D. H. (2020). Analysis of OTA impact on hotel reservations: Case study: Bogotá. Turismo y Sociedad, 27, 145–159. https://doi.org/10.18601/01207555.n27.08

Filieri, R., & Mariani, M. (2021). The role of cultural values in consumers' evaluation of online review helpfulness: A big data approach. International Marketing Review, 38(6), 1267–1288. https://doi.org/10.1108/IMR-07-2020-0172

HOTREC. (2024). European hotel distribution study 2024: Results for the reference year 2023 (R. Schegg, Author). HOTREC & HES-SO Valais-Wallis. https://www.hotrec.eu/media/static/files/import/all_news_2024_2024_21/hotrec-distribution-study-2024.pdf

Hoffmann, C. P., Lutz, C., & Meckel, M. (2014). Digital natives or digital immigrants? The impact of user characteristics on online trust. Journal of Management Information Systems, 31(3), 138–171.

ISO. (2019). ISO 9241-210:2019 Ergonomics of human-system interaction. Part 210: Human-centred design for interactive systems. International Organization for Standardization.

Islami, R., Hilabi, S. S., & Hananto, A. (2023). Analisis user experience aplikasi Traveloka dan Tiket.Com menggunakan metode User Experience Quesionnaire. REMIK: Riset dan E-Jurnal Manajemen Informatika Komputer, 7(1), 497–505. https://doi.org/10.33395/remik.v7i1.12106

Jung, E., Shi, S., & Rew, D. (2026). Unlocking trust: How UX builds trust and engagement in online travel agencies. Tourism and Hospitality Research. https://doi.org/10.1177/14673584261418145

Juric, M., Pehar, F., & Peša Pavlović, N. (2024). Translation, psychometric evaluation, and

validation of the Croatian version of the User Experience Questionnaire (UEQ). International Journal of Human–Computer Interaction, 40(7), 1644–1657. https://doi.org/10.1080/10447318.2022.2147713

Kahneman, D. (2011). Thinking, fast and slow. Farrar, Straus and Giroux.

Kalinić, Z., & Novaković, M. (2019). Online accommodation booking habits and attitudes of Serbian travelers. Menadžment u Hotelijerstvu i Turizmu, 7(2), 11–23. https://doi.org/10.5937/menhottur1902011K

Kim, J. H., Kim, J., Park, J., Kim, C., & Jhang, J. (2023). When ChatGPT gives incorrect answers: The impact of inaccurate information by generative AI on tourism decision-making. Journal of Travel Research. https://doi.org/10.1177/00472875231212996

Limayem, M., Hirt, S. G., & Cheung, C. M. K. (2007). How habit limits the predictive power of intention: The case of information systems continuance. MIS Quarterly, 31(4), 705–737.

Luo, X., Xu, D., Li, Y., & Wan, L. C. (2025). Advancing information search through GenAI: The roles of search type, travel motive and GenAI customization level. International Journal of Contemporary Hospitality Management, 37(5), 1725–1743. https://doi.org/10.1108/IJCHM-06-2024-0941

Melissourgos, G., & Katsanos, C. (2025). UEQ-GR and UEQ-S-GR: Towards a Greek adaptation of the User Experience Questionnaire and its short version. In CHIGreece ’25: Proceedings of the 3rd International Conference of the ACM Greek SIGCHI Chapter (203–211). https://doi.org/10.1145/3749012.3749059

Mich, L., & Garigliano, R. (2023). ChatGPT for e-tourism: A technological perspective. Information Technology & Tourism, 25, 1–12.

Nie, F., Yuhan, S., Hou, T., Shi, S., & Chen, X. (2025). Enhancing tourist experience across age groups with large language models: Evaluating ChatGPT as a virtual tour guide. Proceedings of the ACM on Smart Tourism. https://doi.org/10.1145/3745238.3745457

Pinto, P., Castro, C., & Mota, A. (2019). Online travel agencies: Factors influencing tourists' purchase decisions. Tourism & Management Studies, 15(2), 7–20.

Queen, T. L., Hess, T. M., Ennis, G. E., Dowd, K., & Grühn, D. (2012). Information search and decision making: Effects of age and complexity on strategy use. Psychology and Aging, 27(4), 817–824.

Sanchiz, M., Chin, J., Chevalier, A., Fu, W.-T., Amadieu, F., & He, J. (2017). Searching for information on the web: Impact of cognitive aging, prior domain knowledge and complexity of the search problems. Information Processing and Management, 53(1), 281–294. https://doi.org/10.1016/j.ipm.2016.09.003

Schrepp, M., Hinderks, A., & Thomaschewski, J. (2017). Design and evaluation of a short version of the User Experience Questionnaire (UEQ-S). International Journal of Interactive Multimedia and Artificial Intelligence, 4(6), 103–108.

Venkatesh, V., Thong, J. Y. L., & Xu, X. (2012). Consumer acceptance and use of information technology: Extending the unified theory of acceptance and use of technology. MIS Quarterly, 36(1), 157–178.

Xie, K. L., & Lee, Y. J. (2020). Hotels at fingertips: Informational cues in consumer conversion from search, click-through, to book. Journal of Hospitality and Tourism Technology, 11(1), 49–67. https://doi.org/10.1108/JHTT-03-2017-0026

Zimmer, J. C., Tams, S., Craig, K., Thatcher, J., & Pak, R. (2015). The role of user age in task performance: Examining curvilinear and interaction effects of user age, expertise, and interface design. Journal of Business Economics, 85(4), 367–394.

Designing Olivia AI: User Expectations for a Trustworthy GDPR Compliance Assistant

Anamarija Mladinić
University of Zagreb
Faculty of Organization and Informatics
Pavlinska 2, 42000 Varaždin, Croatia
amladinic25@student.foi.hr

Abstract. *This paper examines whether Olivia, a Croatian GDPR compliance platform, should evolve into an AI-assisted compliance assistant. Based on 686 responses from registered Olivia users, the study analyses satisfaction with the existing tool, expectations for AI-supported GDPR guidance, and the need for a module assessing GDPR compliance in concrete AI tool use. The results show strong demand for such a module and confirm that trust safeguards, especially official sources and clear limitations, are essential for adoption. The paper argues that Olivia AI should be designed as a controlled, source-based compliance workflow, not as a generic chatbot.*

Keywords. GDPR, artificial intelligence, AI governance, data protection, SMEs

1 Introduction

Since 25 May 2018, the General Data Protection Regulation (GDPR) has been fully applicable across the European Economic Area (European Parliament and Council, 2016). For many smaller organisations, the difficult part has never been merely reading the Regulation, but translating it into everyday work: choosing a lawful basis, informing data subjects, keeping records, assessing risks, managing processors and reacting to data breaches. This is particularly relevant in Croatia, where micro, small and medium-sized enterprises (SMEs) dominate the economy and crafts also represent a large part of business activity. These organisations often process personal data daily, but do not necessarily have internal legal, IT or data protection capacity.

Earlier Croatian evidence already pointed to an implementation gap. In the EU-funded Awareness Raising Campaign for SMEs (ARC I) project, aimed at supporting Croatian SMEs in their efforts to comply with the GDPR, many SMEs recognised that GDPR compliance mattered for business, but far fewer had reviewed their processing activities or felt confident that their staff had enough information to act in line with GDPR obligations. The same study also showed strong demand for a practical and free digital tool (Mladinić, Vukić, & Rončević, 2023). Olivia was developed in that context by the Croatian Personal Data Protection Agency (AZOP) and the Faculty of Organization and Informatics (FOI) as part of the ARC II project.

Olivia is a web-based GDPR tool designed as a free and practical platform for GDPR awareness, e-learning and compliance support, primarily for SMEs (Olivia GDPR ARC, 2025). The current platform includes GDPR learning modules, quizzes, certificates, self-assessment questionnaires, practical templates and tools supporting the preparation of privacy notices, consent forms, legitimate interest assessments, records of processing activities, templates and other compliance documents. However, the platform is still largely modular and user-driven: users must know which topic is relevant, which module to open and how to connect different outputs into a coherent compliance workflow.

This limitation becomes particularly relevant in the context of tools based on artificial intelligence (AI). Organisations using AI need to assess several connected GDPR questions at once, including whether personal data are processed, which legal basis applies, whether data subjects are informed, whether profiling or automated decision-making is involved, whether a data protection impact assessment is required and which safeguards should be documented. This raises the question of whether Olivia should evolve from a digital GDPR support tool into a more interactive, AI-assisted compliance assistant.

Since 2025, the regulatory and technological context has shifted substantially, particularly due to the wider adoption of AI tools and the evolving EU digital regulatory framework. Organisations that still struggle with basic GDPR questions increasingly use AI tools for analytics, automation, customer interaction or internal productivity (Croatian Employers' Association, 2025; Croatian Chamber of Economy, 2024). This creates a more specific compliance

problem: not simply how to comply with the GDPR, but how to understand GDPR duties when a concrete AI tool processes, stores or helps generate information about individuals. The key issue is therefore not whether Olivia should use AI, but under what governance conditions users would trust such support.

The aim of this paper is to determine the extent to which users recognise the need for a tool that would help organisations understand and assess GDPR obligations linked to concrete AI tools or AI systems, and which elements of usefulness, trust and regulatory assurance should be embedded in its design so that Olivia can become a responsible digital innovation in the field of GDPR and AI compliance.

Accordingly, the main research question is: to what extent do users of the Olivia platform recognise the need to develop an AI/GDPR self-assessment tool that would help organisations understand and verify GDPR obligations related to the use of specific AI tools or artificial intelligence systems? The study tests four hypotheses focused on users' need for an AI/GDPR self-assessment module, its relationship with adoption readiness, and the role of trust safeguards:

H1: Users of the Olivia platform express a high need for an AI/GDPR self-assessment module that supports both understanding GDPR obligations related to AI tools and assessing GDPR compliance in concrete AI tool use.
H2: A higher need for an AI/GDPR self-assessment module is positively associated with users' readiness to use AI Olivia.
H3: The need for an AI/GDPR self-assessment module significantly predicts users' readiness to use AI Olivia, even after controlling for organisational and user-related characteristics.
H4: Trust safeguards, particularly the inclusion of official sources, relevant GDPR provisions and clear limitations of the tool, are positively associated with users' readiness to use AI Olivia.

The rest of the paper is structured as follows. Section 2 presents the theoretical and regulatory background, focusing on the relationship between GDPR, AI governance and technology acceptance. Section 3 describes the methodology, including the survey design, sample, variables and statistical methods. Section 4 presents the results of the descriptive, correlational, regression and thematic analyses. Section 5 discusses the findings in relation to the hypotheses and derives design implications for Olivia AI. Section 6 concludes the paper by summarising the scientific and practical contribution of the study.

2 Theoretical Background and Research Context

The European Union (EU) digital regulatory environment is dense and still moving. Alongside the GDPR, organisations must now consider the AI Act, the Data Act, the Data Governance Act, the Digital Services Act, the NIS2 Directive and sector-specific rules, creating an increasingly complex compliance environment (Graux et al., 2025). The Digital Omnibus package confirms that the EU is trying to simplify and clarify this regulatory environment, but it does not remove organisational responsibility (European Commission, 2025). Simplification does not mean that controllers may stop assessing legal basis, transparency, data subject rights, data protection impact assessment (DPIA) requirements or security measures. On the contrary, clearer rules increase the expectation that organisations will be able to operationalise them.

This question is central for AI use. If personal data are used in the development or operation of AI systems, organisations need to understand the purpose of processing, lawful basis, transparency duties, possible rights of objection, risks to data subjects and the need for technical and organisational safeguards (European Data Protection Board, 2024; Sartor & Lagioia, 2020). The AI Act adds further obligations, especially for high-risk systems, but it does not replace the GDPR; rather, the two regimes operate as complementary layers of EU digital governance (European Parliament and Council, 2024; Hohmann & Kollár, 2025; Holtz & Ledendal, 2026; Sartor & Lagioia, 2020). A tool such as Olivia could be particularly important for Croatian startups and other organisations developing high-risk AI systems, since bringing such products to the EU market requires not only demonstrating compliance with the AI Act, but also ensuring and documenting compliance with the GDPR whenever the system involves the processing of personal data. A compliance tool dealing with AI cannot be limited to AI Act classification; it must also ask whether and how personal data are processed, because GDPR obligations apply whenever AI development or deployment involves personal data (European Data Protection Board, 2024; Croatian Personal Data Protection Agency, 2025).

For SMEs and other smaller organisations this is a practical problem. They may use commercially available AI tools without fully understanding whether prompts, documents, customer records or employee data are processed, stored or reused for model improvement. They may also rely on vendors without sufficiently assessing controller-processor roles, transfers outside the EEA, retention periods or contractual safeguards. A general statement that AI must be used lawfully is not operationally useful. What smaller organisations need is a sequence of concrete questions that translates legal requirements into an assessment workflow.

The survey instrument partly follows the logic of the Technology Acceptance Model, especially perceived usefulness and behavioural intention to use

(Davis, 1989; Venkatesh & Davis, 2000). However, a GDPR/AI assistant is a legally sensitive tool, so usefulness alone is insufficient. Trust, source reliability and clear limitations are part of adoption. This is consistent with work on trust in automation and intelligent systems acceptance, as well as broader discussions on transparency and the black-box problem in AI (Lee & See, 2004; Vorm & Combs, 2022; Zerilli et al., 2022; von Eschenbach, 2021). In the context of regulatory compliance, a wrong but confident answer, or an explanation that appears convincing but is misleading, can create legal risk (Lakkaraju & Bastani, 2020). Therefore, adoption readiness must be analysed together with governance safeguards.

The point of Olivia AI should not be to make the existing platform sound more like a chatbot. Its value would be in helping users move from general GDPR information to a structured assessment of what they actually need to check when using AI tools. For that reason, the system should be based on verified sources, clear limits, traceable outputs and safeguards against overconfident legal answers. The paper therefore connects three issues that cannot be separated in this context: whether users would accept such a tool, whether it would support GDPR compliance in practice, and whether it could be governed responsibly. In this sense, the research links technology acceptance, data protection compliance and AI governance (Batool et al., 2024; Papagiannidis et al., 2025; Raji et al., 2020).

This approach also reflects a broader shift in digital regulation. The EU increasingly expects organisations to translate abstract legal requirements into documented governance processes. In the GDPR, this is visible through accountability, records of processing activities, data protection impact assessment, processor management and breach documentation. In the AI Act, the same logic appears through risk classification, transparency duties, technical documentation, human oversight and post-market monitoring for higher-risk systems. Although Olivia AI would not itself be designed as a high-risk AI system, its subject matter is high-stakes because it guides users through legal and organisational choices. The design must therefore make legal uncertainty visible rather than hide it behind a fluent answer, and should be aligned with responsible AI governance and auditability practices (Papagiannidis et al., 2025; Raji et al., 2020).

For that reason, the proposed module is not a substitute for legal advice. Its value lies in helping non-expert users identify the relevant issues before they make decisions. In practice, this means asking better questions: Is personal data entered into the tool? Is the AI provider a processor, controller or independent controller? Are prompts stored? Are they used to train the model? Is there a transfer outside the EEA? Is the output used to make decisions about individuals? These questions are the entry points for GDPR analysis and for deciding whether additional expert review is required.

3 Methodology

An online questionnaire was conducted through the European Commission's EU Survey platform from 11 to 20 May 2026. The invitation was sent to 7,314 registered active Olivia users, with a reminder sent on 18 May. In total, 686 valid responses were collected, corresponding to a response rate of 9.38%. All closed questions were mandatory, so questions 1 to 19 had no missing answers. The sample is sufficiently large for descriptive and inferential statistical analysis; however, it should be interpreted as a sample of responding Olivia users rather than a representative sample of all Croatian organisations. The survey data used in this paper are based on the Olivia User Survey conducted in 2026 (Olivia User Survey, 2026). The questionnaire included 20 questions organised into several blocks. The first block covered respondent and organisational profile: type of organisation, role of the respondent, existence of a data protection officer (DPO) or other GDPR support and frequency of Olivia use. The second block examined the purposes for which Olivia had been used, such as understanding GDPR obligations, lawful bases, data subject rights, records of processing activities, security measures, privacy notices, video surveillance, cookies and webinars. The third block measured satisfaction with the existing Olivia platform. The fourth block measured perceived usefulness of a future AI version of Olivia. The final block measured trust safeguards, concern about incorrect or oversimplified AI answers and readiness to use AI Olivia.

Likert-scale answers were coded from 1 to 5. Data preparation and manual coding of open-ended answers were conducted in Microsoft Excel. Quantitative analyses were conducted in jamovi Desktop 2.7.30. The analyses included descriptive statistics, reliability analysis using Cronbach's alpha, Spearman correlations, Kruskal-Wallis and Mann-Whitney tests, and binary logistic regression. Statistical significance was assessed at the 5% level ($p < 0.05$).

Several composite variables were created. Satisfaction with the existing Olivia platform was measured through six items. Perceived usefulness of AI Olivia was measured through items addressing AI-supported guidance, search time, tailored recommendations, checklists, forms and DPIA screening. Trust in safeguards was measured through items addressing official sources, GDPR provisions, links to official materials and clear limits of the tool. Questions 14 and 15 were combined into a two-item construct measuring the need for a module assessing GDPR compliance in the use of AI tools. Question 19 was used as the main dependent variable for adoption readiness.

Because Likert items are ordinal, Spearman's rho was used to test associations between the need construct, trust safeguards, concern about incorrect answers and readiness to use AI Olivia. Group differences were examined with non-parametric tests.

A binary logistic regression was then used to test whether the need for the AI-tool GDPR compliance module predicts readiness to use AI Olivia. Answers 4 and 5 on Question 19 were coded as readiness to use, while answers 1 to 3 were coded as neutral or negative. The model included organisational type, GDPR support, frequency of use, respondent role, trust safeguards and concern about incorrect answers as controls.

The open-ended Question 20 asked respondents to provide additional suggestions for improving Olivia or developing the future AI virtual assistant. Responses were coded manually using thematic coding. Empty or non-substantive answers were removed. The unit of analysis was the individual comment. A primary thematic code was assigned to each substantive comment, with additional codes used where comments covered more than one theme. This qualitative component was used to interpret and deepen the quantitative findings.

The coding was used as an interpretative layer to help identify whether the numerical support for AI Olivia was unconditional or whether users attached specific expectations to it. The latter proved to be the case. The most frequent themes were not novelty, automation or speed, but trust, accuracy, official sources, human oversight and practical guidance. This qualitative evidence is important because it prevents overinterpreting high agreement with AI functionality as unqualified enthusiasm for automation.

4 Results

The respondent structure shows that Olivia is used both by its original target group and by a wider range of organisations. When micro, small and medium-sized enterprises are grouped together, SMEs represent the largest share of respondents, with 250 responses, or 36.45% of the sample. Large enterprises account for 28.72%, while public authorities represent 24.93%. Crafts and self-employed activities account for 3.64%, non-profit organisations for 3.35%, and other organisations for 2.92%. This confirms that Olivia is relevant for SMEs, but also that its use extends well beyond them.

Data protection officers represented the largest respondent group (30.61%), followed by employees (23.03%), administrative or business staff (13.41%) and lawyers (13.27%). Most respondents came from organisations with an appointed DPO (67.93%) or another legal/GDPR support function (17.49%). This suggests that the sample includes many users who are already aware of data protection obligations.

Olivia is most often used occasionally. Almost half of the respondents used it several times per year (47.23%), while 34.69% had used it only once. The most common reason for using Olivia was understanding basic GDPR obligations (60.50%), followed by viewing webinars or presentations (36.88%) and support in determining the lawful basis for processing (27.41%). Other reported purposes included data subject rights, records of processing activities, security measures, privacy notices, breaches and specific topics such as video surveillance or cookies. This confirms that Olivia functions both as an awareness tool and as practical compliance support.

Satisfaction with the existing platform was high. A total of 87.61% agreed that Olivia is useful as a digital tool for GDPR information, 86.30% agreed that it is easy to use and 80.91% agreed that it helps them understand what their organisation should do to comply with the GDPR. The somewhat lower, but still positive, score for everyday practical usefulness for small and medium-sized organisations (72.74%) indicates room for improvement. The satisfaction scale had a mean of 4.12, median of 4.17, standard deviation of 0.69 and Cronbach's alpha of 0.920, indicating very high internal consistency.

Expectations for AI Olivia were also high. Users supported functions that would reduce search time (83.67%), provide guidance adapted to organisation type and processing context (83.24%), support basic forms (83.23%), support DPIA screening (82.66%) and provide interactive GDPR checklists (80.32%). The perceived usefulness scale had a mean of 4.19, median of 4.22, standard deviation of 0.76 and Cronbach's alpha of 0.964. This very high reliability suggests that users see these AI functions as connected elements of a broader AI-assisted compliance tool.

The strongest result for the research question comes from Questions 14 and 15. A total of 79.44% agreed that Olivia support would be useful for understanding GDPR obligations related to AI systems or AI tools, including lawful basis, transparency, data subject rights and security. Similarly, 78.86% agreed that an interactive assessment of GDPR compliance in the use of a concrete AI tool would be useful. These two items formed a reliable two-item construct, with a mean of 4.13, median of 4.00, standard deviation of 0.82 and Cronbach's alpha of 0.905. The result shows that users do not only want general AI guidance; they want a concrete module for assessing GDPR obligations in specific AI tool use.

Trust was the most emphatic user requirement. The highest agreement in the whole questionnaire concerned official sources: 86.74% would trust AI Olivia more if answers included official sources, relevant GDPR provisions or links to additional official materials. In addition, 72.31% would trust the tool more if it clearly stated that it provides general information rather than individual legal advice. Concern about incorrect or oversimplified answers was also substantial: 52.63% agreed that such concern could deter them from using AI Olivia. Nevertheless, 82.36% stated that they would be ready to use AI Olivia if it had useful functionalities and clear safeguards. The key descriptive findings are summarised in Table 1.

Table 1. Key descriptive findings on existing Olivia and expected AI support

Item	% Agreement (4 or 5)	Mean
Existing Olivia is useful as a digital GDPR information tool	87.61%	4.33
An AI upgrade of Olivia would be useful	78.28%	4.18
Need for support in understanding GDPR obligations related to AI tools (Q14)	79.44%	4.14
Need for an interactive assessment of concrete AI tool use (Q15)	78.86%	4.13
More trust if answers include official sources and GDPR provisions	86.74%	4.43
Readiness to use AI Olivia (Q19)	82.36%	4.15

Spearman correlations confirmed the link between module need and adoption readiness. The need construct was positively and significantly associated with readiness to use AI Olivia (rho = 0.585, $p < 0.001$). Questions 14 and 15 were also individually associated with readiness (rho = 0.559 and rho = 0.568, both $p < 0.001$). The need construct was positively associated with trust safeguards (rho = 0.497, $p < 0.001$), but it was not significantly associated with concern about incorrect answers (rho = -0.036, p = 0.352). Group comparisons showed that the need for the module did not significantly differ by organisation type. The need was present across SMEs, large enterprises, public authorities and other organisations. A statistically significant but small difference was found according to GDPR support in the organisation (H = 11.954, p = 0.018), with the lowest need among respondents who were unsure whether their organisation had GDPR support. Frequency of Olivia use mattered more clearly: occasional and regular users expressed a higher need for the module than one-time users (H = 15.969, $p < 0.001$). Differences by respondent role were not statistically significant (H = 10.996, p = 0.089).

Binary logistic regression showed that need for the AI-tool GDPR compliance module significantly predicted readiness to use AI Olivia (OR = 3.69, 95% CI: 2.55-5.34, $p < 0.001$), controlling for organisational and user-related characteristics. Trust safeguards were also a significant predictor (OR = 2.20, 95% CI: 1.53-3.16, $p < 0.001$). Concern about incorrect or oversimplified answers was not significant (OR = 1.06, p = 0.682). Model quality was good, with AUC = 0.853 and Nagelkerke R^2 = 0.403. Predicted probabilities illustrate the effect: when module need is at level 3, readiness is about 57%; at level 4, about 83%; and at level 5, about 95%. The main regression estimates are presented in Table 2.

Table 2. Binary logistic regression predicting readiness to use AI Olivia

Predictor	OR = odds ratio	95% CI = confidence interval	p
AI/GDPR self-assessment need (Q14+Q15)	3.69	2.55-5.34	<0.001
Governance trust (Q16+Q17)	2.20	1.53-3.16	<0.001
Risk concern (Q18)	1.06	0.82-1.36	0.682

The open-ended comments supported the quantitative findings. Of 686 respondents, 174 left some response to Question 20; after removing non-substantive answers, 85 substantive comments remained. The most frequent theme was trust, sources, accuracy and human oversight (36.5% of coded comments). Other themes included guided procedures, self-assessment and documents (16.5%), practical examples and concrete interpretations (12.9%), AI functionalities and contextual adaptation (10.6%), and scepticism or risks linked to AI (10.6%). One comment was particularly important because it suggested documenting and publishing the development of Olivia AI as an example of compliance for other AI deployers and providers. This expands Olivia's relevance beyond a user-facing tool: it could also serve as a demonstrator of responsible AI governance in the public interest. The distribution of primary qualitative themes is shown in Table 3.

Table 3. Main themes from open-ended responses

Code	Main theme	Frequency	% of substantive comments
T4	Trust, sources, accuracy and human oversight	31	36.47%
T2	Guided procedures, self-assessment and documents	14	16.47%
T1	Practical examples and concrete interpretations	11	12.94%
T3	AI functionality and adaptation to user context	9	10.59%
T7	Scepticism and AI-related risks	9	10.59%
T5	Education, webinars, updates and continuous support	6	7.06%

T6	User experience and technical improvements	5	5.88%

Taken together, the descriptive, inferential and qualitative results point in the same direction. Users want structured support for decisions that are currently difficult to make, especially when AI tools are involved. The strongest demand is therefore located at the intersection of practical usefulness and governance safeguards.

5 Discussion

The findings support all four hypotheses. H1 is supported by strong agreement with the two items measuring the need for AI/GDPR support: 79.44% of respondents agreed that Olivia support would be useful for understanding GDPR obligations related to AI systems or AI tools, and 78.86% agreed that an interactive assessment of GDPR compliance in concrete AI tool use would be useful. H2 is supported by the significant positive correlation between the AI/GDPR self-assessment need construct and readiness to use AI Olivia (rho = 0.585, $p < 0.001$). H3 is supported by the binary logistic regression model, where the need for the module remains a significant predictor of adoption readiness after controls (OR = 3.69, $p < 0.001$). H4 is supported by the significant role of trust safeguards in predicting readiness to use AI Olivia (OR = 2.20, $p < 0.001$). The empirical answer to the research question is therefore clear: Olivia users strongly recognise the need for a module that helps organisations understand and assess GDPR obligations related to concrete AI tools or AI systems. At the same time, the results draw an important boundary. Users do not want a confident but uncontrolled chatbot. They want a tool that cites official sources, explains its limits, avoids false legal certainty and supports structured assessment. This is not a minor design preference. In a regulatory domain, a convincing but wrong answer can mislead the user into non-compliance. The future system therefore needs to be designed around legal reliability, not only conversational convenience.

The findings also fit an extended technology-acceptance logic. Perceived usefulness clearly matters, as shown by the strong support for AI functions and the predictive effect of the AI-tool assessment need. However, usefulness is not enough. Users are willing to adopt AI support when usefulness is combined with safeguards: official sources, GDPR references, links to additional materials and clear limits. This is especially relevant because more than half of respondents expressed concern that incorrect or oversimplified answers could deter them from using AI Olivia. Users are not rejecting AI; they are asking for controlled AI.

These results have direct implications for the design of Olivia AI. The upgrade should turn Olivia into a guided compliance workflow. The system could first ask users about the organisation, sector, role in processing and type of processing activity, and then direct them to relevant compliance steps. If the user does not inform data subjects, Olivia should lead to the privacy notice module. If legitimate interest is used, it should trigger the legitimate interest assessment. If AI or high-risk processing is involved, it should trigger DPIA-related questions.

The most distinctive functionality should be a module assessing GDPR compliance in the use of AI tools. The proposed workflow is shown in Fig. 1.

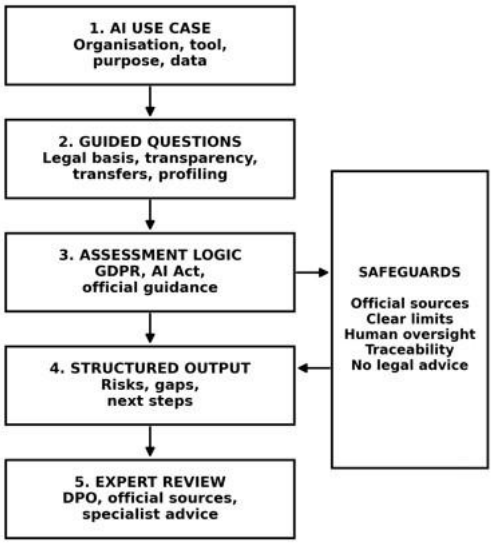


Figure 1. Proposed Olivia AI self-assessment workflow

This module should not assess “AI” in general, but a specific use of a specific AI tool in a specific organisational context. It should ask whether personal data are entered into the tool, which categories of data are processed, whether special categories of data are involved, whether the tool is used for profiling or decision-making, whether input data are used for model training, where data are stored, which provider is involved, whether transfers outside the European Economic Area (EEA) occur and whether data subjects are properly informed. The output should not state that the user is “GDPR compliant”. It should provide a structured overview of risks, missing information and recommended next steps.

Technically, the findings support a hybrid model rather than a general-purpose generative chatbot. Such a system should combine a curated and versioned knowledge base, retrieval-augmented generation, rule-based decision logic, structured questionnaires, controlled templates, document analysis and human oversight. The language model should help explain, navigate and communicate, but it should not be treated as the source of legal truth. If the system cannot identify a reliable source or if the case is too complex, the safer response is to state the limitation and refer the user to official sources, a DPO or expert advice. This

is consistent with the need for traceability, internal auditing and governance controls in AI systems (Papagiannidis et al., 2025; Raji et al., 2020).

The study also indicates that Olivia AI could have a broader role. If its development process is documented, including risk assessment, DPIA, AI Act classification, knowledge-based governance, red-teaming, hallucination testing, prompt-injection controls, versioning and incident handling, Olivia could become a practical example of governance-by-design. This would be particularly valuable because many organisations are being told to govern AI responsibly, but lack examples of how to document and operationalise such governance in practice (Batool et al., 2024; Papagiannidis et al., 2025; Raji et al., 2020).

Such documentation would also be aligned with the expectations users expressed in the survey. If users expect official sources, clear limits and human oversight in the answers they receive, the same standards should apply to the system behind those answers. The governance of the tool should therefore be visible at least at a high level: what sources are used, how they are updated, when the system refuses to answer, how errors are reported and how sensitive user inputs are protected. This would reduce the risk of automation bias and make the platform more credible for both SMEs and more mature organisations.

A further implication concerns the role of public institutions in digital innovation. Olivia was developed by a supervisory authority and an academic partner, not by a commercial vendor. That gives the compliance tool a different legitimacy, but also creates higher expectations. A public-interest AI compliance assistant should be more conservative than a commercial productivity chatbot. It should prefer a limited but reliable answer to a broad but unsupported one.

The study has limitations. Respondents were registered Olivia users and participation was voluntary, so the findings cannot be generalised to all Croatian organisations. Many respondents already came from organisations with a DPO or another GDPR support function, which may increase the level of awareness in the sample. The survey measures expectations and readiness, not actual behaviour after deployment. Future research should therefore include prototype testing, usability testing, analysis of real user behaviour and evaluation of whether the module improves concrete compliance decisions.

6 Conclusion

The study shows that Olivia has a strong basis for further development. Users value the existing platform and clearly support AI-assisted GDPR guidance. The most important finding is the demand for a module that helps assess GDPR compliance in the use of concrete AI tools. This need is strongly associated with, and significantly predicts, readiness to use AI Olivia.

However, adoption depends on trust. Users expect official sources, GDPR references, clear limitations and safeguards against incorrect or oversimplified answers. The future Olivia AI should therefore be designed as a hybrid compliance system with a controlled knowledge base, RAG, structured questions, rule-based logic, validated templates, human oversight and clear boundaries.

The scientific contribution of this paper lies in connecting technology acceptance, GDPR compliance and AI governance in the specific context of an AI-assisted compliance tool. The study provides empirical evidence that users recognise a specific need for a module assessing GDPR compliance in the use of concrete AI tools. It further shows that adoption readiness is shaped by perceived usefulness and by trust safeguards, including official sources, GDPR references and clear limitations of the tool.

The practical contribution lies in translating these findings into design requirements for Olivia AI. The results suggest that Olivia AI should be developed as a controlled, source-based compliance workflow combining structured questions, verified legal sources, rule-based logic, validated templates, human oversight and clear boundaries. If developed transparently and properly documented, Olivia AI could serve both as a support tool for organisations and as a practical example of responsible AI governance in public interest.

References

Batool, A., Zowghi, D., & Bano, M. (2024). *Responsible AI governance: A systematic literature review*. arXiv. https://arxiv.org/abs/2401.10896

Croatian Chamber of Economy. (2024). *Analysis of the perception and use of artificial intelligence in the Croatian economy*. Retrieved from https://hgk.hr/documents/f2024-ai-studija-baf674e1e6c8c834.pdf

Croatian Employers' Association. (2025). *How the Croatian economy perceives and uses artificial intelligence?* Retrieved from https://www.hup.hr/EasyEdit/UserFiles/2025/a1/hup-b2b-ai-istrazivanje-2025.pdf

Croatian Personal Data Protection Agency. (2025). *Data Protection and Artificial Intelligence*: *Frequently Asked Questions*. Retrieved from https://azop.hr/umjetna-inteligencija-i-zastita-osobnih-podataka/

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly, 13*(3), 319-340. https://doi.org/10.2307/249008

European Commission. (2025). *Digital Omnibus Regulation Proposal*. Retrieved from

https://digital-strategy.ec.europa.eu/en/library/digital-omnibus-regulation-proposal

European Data Protection Board. (2024). Opinion 28/2024 on certain data protection aspects related to processing personal data in the context of AI models. Retrieved from https://www.edpb.europa.eu/system/files/2024-12/edpb_opinion_202428_ai-models_en.pdf

European Parliament and Council. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council (General Data Protection Regulation). Official Journal of the European Union, L 119. Retrieved from https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng

European Parliament and Council. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union. Retrieved from https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

Graux, H., Garstka, K., Murali, N., Cave, J., & Botterman, M. (2025). *Interplay between the AI Act and the EU digital legislative framework. European Parliament*. Retrieved from https://www.europarl.europa.eu/thinktank/en/document/ECTI_STU(2025)778575

Hohmann, B., & Kollár, G. (2025). Reflections on the data protection compliance of AI systems under the EU AI Act. *Cogent Social Sciences, 11(1), 2560654.* https://doi.org/10.1080/23311886.2025.2560654

Holtz, H. M., & Ledendal, J. (2026). AI data governance–overlaps between the AI Act and the GDPR. *Law, Innovation and Technology, 18(1), 380–409.* https://doi.org/10.1080/17579961.2026.2633677

Lakkaraju, H., & Bastani, O. (2020). "How do I fool you?" Manipulating user trust via misleading black-box explanations. In *Proceedings of the AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society* (pp. 79-85). https://doi.org/10.1145/3375627.3375833

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50-80. https://doi.org/10.1518/hfes.46.1.50_303

Mladinić, A., Vukić, Z., & Rončević, A. (2023). GDPR compliance challenges in Croatian micro, small and medium-sized enterprises. *Pravni vjesnik, 39*(3-4), 53-75. https://doi.org/10.25234/pv/23972

Olivia GDPR ARC. (2025). *Olivia – web tool for GDPR compliance*. Retrieved from https://olivia-gdpr-arc.eu/hr

Olivia User Survey. (2026). *Questionnaire on the use of Olivia and users' needs for the development of an AI virtual assistant for GDPR compliance*. Retrieved from https://arc-rec-project.eu/istrazivanje-olivia/

Papagiannidis, E., Mikalef, P., & Conboy, K. (2025). Responsible artificial intelligence governance: A review and research framework. *The Journal of Strategic Information Systems, 34(2), 101885*. https://doi.org/10.1016/j.jsis.2024.101885

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 33-44*. https://doi.org/10.1145/3351095.3372873

Sartor, G., & Lagioia, F. (2020). *The impact of the General Data Protection Regulation (GDPR) on artificial intelligence*. European Parliament. Retrieved from https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2020)641530

Venkatesh, V., & Davis, F. D. (2000). A theoretical extension of the Technology Acceptance Model: Four longitudinal field studies. *Management Science, 46*(2), 186-204. https://doi.org/10.1287/mnsc.46.2.186.11926

von Eschenbach, W. J. (2021). Transparency and the black box problem: Why we do not trust AI. *Philosophy & Technology, 34*(4), 1607-1622. https://doi.org/10.1007/s13347-021-00477-0

Vorm, E. S., & Combs, D. J. Y. (2022). Integrating transparency, trust, and acceptance: The Intelligent Systems Technology Acceptance Model (ISTAM). *International Journal of Human-Computer Interaction, 38*, 1828-1845. https://doi.org/10.1080/10447318.2022.2070107

Zerilli, J., Bhatt, U., & Weller, A. (2022). How transparency modulates trust in artificial intelligence. *Patterns, 3*(4), 100455. https://doi.org/10.1016/j.patter.2022.100455

Cybersecurity and Patient Data Protection in Healthcare: Research Trends and Organisational Challenges in the Context of NIS2

Rikardo Protrka
University North
Trg dr. Žarka Dolinara 1, Koprivnica, Croatia
rikardo.protrka@unin.hr

Antonija Srok
University of Rijeka, Faculty of Economics and Business
Department of Quantitative Economics
Ivana Filipovića 4, Rijeka, Croatia
antonija.srok@efri.uniri.hr

Abstract. *Healthcare digitalisation has increased the use of patient data, but also created new cybersecurity risks. This paper examines recent research on healthcare cybersecurity and patient data protection in the context of NIS2, adopted in 2022 to strengthen cybersecurity obligations. We analysed 1,951 WoSCC articles published from 2023 to 2025 using bibliometric analysis, keyword co-occurrence and topic-trend analysis, and full-text content analysis of the ten most-cited articles. Article output increased, with security, blockchain, privacy, cybersecurity, and authentication central to the field. The findings suggest that secure digital healthcare also requires, in practice, organisational preparedness, interoperability, patient trust, governance, and implementation resources.*

Keywords. digital health, electronic health records, interoperability, cybersecurity, NIS2 Directive, healthcare management

1 Introduction

The digitalisation of healthcare has made patient data more widely used, more connected, and more difficult to protect. Electronic health records (EHRs), telemedicine services, interoperable platforms, artificial intelligence (AI)-based tools, and connected medical devices can improve access to information and support more coordinated care. At the same time, they increase the number of systems, users, and data flows that need to be secured.

In healthcare, a cybersecurity incident is not simply a data-protection issue. It can also become an operational and clinical problem: electronic records may become unavailable, clinical workflows may be interrupted, treatment may be delayed, and patients' trust in digital services may weaken. Cybersecurity is therefore closely linked to patient safety and the continuity of care. It also has an organisational dimension: healthcare providers need appropriate governance arrangements, technical safeguards, trained staff, and sufficient resources to manage risks effectively (Carello et al., 2023). The importance of these issues has increased following the adoption of the NIS2 Directive by the European Union in 2022 (European Commission, 2026). The Directive strengthens cybersecurity obligations for essential sectors, including healthcare, and places greater emphasis on risk management, organisational responsibility, and the resilience of information systems. Alongside the General Data Protection Regulation (GDPR) and other relevant regulatory frameworks, it highlights the need to protect sensitive patient data while maintaining the reliable operation of healthcare services (European Parliament and Council of the European Union, 2025).

Research on healthcare cybersecurity has expanded in recent years, but the literature remains spread across partly separate strands, including secure data sharing, EHRs, AI-based applications, connected devices, interoperability, and patient trust. Technical solutions remain important, but their practical value also depends on whether healthcare organisations have the governance, infrastructure, staff capacity, and resources needed to implement them safely (Lampropoulos et al., 2023). Given the breadth and rapid development of the field, bibliometric analysis is useful for identifying where research activity is concentrated and how topics are connected. It provides a systematic way to map the expanding body of research, identify publication and citation patterns, and examine how the main areas of interest are developing (Donthu et al., 2021). Bibliometric methods have been widely applied across scientific fields and, more recently, in digital-health research, where technological, clinical, organisational, and regulatory issues frequently overlap.

The purpose of this study is to map recent research on cybersecurity and information security in healthcare, with particular attention to the protection

and secure management of patient data. We analyse 1,951 articles from the Web of Science Core Collection (WoS CC) published between 2023 and 2025. We combine bibliometric performance analysis with keyword co-occurrence and topic-trend analyses, complemented by a full-text content analysis of the ten most-cited articles. This combined approach allows us to examine both the broader structure of the research field and the issues emphasised in its most influential articles. We describe the development of the field, identify its leading contributors and key research topics, and assess how closely the current literature relates to areas relevant to NIS2 implementation, including secure data governance, risk management, interoperability, organisational preparedness, and system resilience.

In doing so, the paper examines whether recent healthcare cybersecurity research reflects the practical conditions under which secure digital healthcare must be implemented. To address these aims, we answer the following research questions:

RQ1. How has research on cybersecurity and patient data protection in healthcare evolved from 2023 to 2025 in terms of scientific production?

RQ2. Which countries, affiliations, journals, authors, and articles are the most prominent contributors to this research field?

RQ3. Which author keywords, keyword co-occurrence patterns, and topic trends shape the current research landscape of cybersecurity and patient data protection in healthcare?

RQ4. Which technical, patient-related, organisational, and resource-related issues emerge from the most influential articles?

RQ5. How do the identified research themes relate to areas relevant to the implementation of the NIS2 Directive, particularly secure data governance, risk management, interoperability, organisational preparedness, and the resilience of healthcare information systems?

The remainder of the paper is organised as follows. Section 2 describes the data collection and preparation process. Section 3 presents the methodology used for the bibliometric and content analyses. Section 4 presents and discusses the results. Section 5 summarises the main conclusions, outlines the study limitations, and identifies directions for future research.

2 Data Collection

The data collection procedure followed established guidelines for bibliometric research (e.g., Donthu et al., 2021). We retrieved bibliographic records from the Wos CC, which is widely used in bibliometric studies because of its broad coverage of scientific journals and citation indexes (e.g., Pranckutė, 2021). We conducted a topic-based search covering article titles, abstracts, and author keywords. The search was based on the following query:

*((NIS*2 directive) OR (network and information systems directive) OR (Directive 2022/2555) OR cybersecurit* OR (information securit*)) AND health* AND patient**

We retrieved the records on 8 June 2026 and restricted the search to English-language articles, review articles, and proceedings papers published between 2023 and 2025. The initial search yielded 1,986 articles. We imported the bibliographic records into R and assessed metadata completeness before conducting the analysis. Where possible, we used DOI-based lookup procedures to complete missing information. We also standardised cited references and merged duplicate citation variants, primarily on the basis of DOI information. After additional cleaning based on publication year, the final dataset included 1,951 articles published between 2023 and 2025. This final dataset formed the basis for the descriptive, bibliometric, and content analyses.

3 Methodology

We conducted the bibliometric analysis in R using the bibliometrix package and its Biblioshiny web interface (Aria & Cuccurullo, 2017). Following Donthu et al. (2021), the analysis combined performance analysis and science mapping.

The performance analysis examined annual scientific production, citation impact, document types, leading journals, authors, affiliations, countries, and patterns of international collaboration. At the journal level, we examined journal output and the concentration of articles across core journals using Bradford's Law. We also identified the ten most-cited articles and analysed their full texts qualitatively.

The science-mapping analysis focused on the conceptual structure of the field. We examined author-keyword frequencies and conducted a keyword co-occurrence analysis to identify relationships between research topics. We also analysed topic trends to examine how the focus of the literature developed during the observed period.

To complement the bibliometric analysis, we conducted a qualitative full-text content analysis of the ten most-cited articles. We identified recurring technical, patient-related, organisational, and resource-related issues and grouped them into four analytical categories. We then mapped these categories interpretatively to areas relevant to the implementation of the NIS2 Directive in healthcare, including secure data governance, risk management, interoperability, organisational preparedness, and the resilience of healthcare information systems.

4 Results and Discussion

4.1 Performance Analysis

The final dataset comprised 1,951 articles published between 2023 and 2025 across 931 sources. The average article age was 1.88 years, which reflects the intentionally recent publication window. The articles included 72,803 cited references and received an average of 7.92 citations per article. The dataset included 8,950 authors from 93 countries and 3,335 reported affiliations, with an average of 5.34 co-authors per article. International co-authorship accounted for 32.34% of the articles, showing that almost one-third of the articles involved cross-country collaboration. The dataset h-index was 50. Table 1 summarises the main characteristics of the dataset. For consistency, all records included in the analysis are referred to as articles throughout the paper, irrespective of their specific document type. The dataset consisted mainly of articles (1,461; 74.88%), followed by proceedings papers (253; 12.97%) and review articles (237; 12.15%).

Table 1. Main characteristics of the dataset

Indicator	Value
Publication period	2023–2025
Articles	1,951
Sources	931
Annual growth rate (%)	20.20
Average article age	1.88
Average citations per article	7.92
Cited references	72,803
Author keywords	5,884
Authors	8,950
Average co-authors per article	5.34
International co-authorship (%)	32.34
Dataset h-index	50
Number of countries	93
Number of affiliations	3,335

Fig. 1 visualises annual scientific production, while Table 2 presents the corresponding citation indicators. The number of articles rose from 508 in 2023 to 709 in 2024 and 734 in 2025. Most of this increase occurred between 2023 and 2024, when the number of articles grew by 39.6%. Growth continued in 2025, although at a slower rate of 3.5%. Overall, scientific production increased by 44.5% between 2023 and 2025, corresponding to an annual growth rate of 20.2%.

Articles published in 2023 had received an average of 13.46 citations per article by the time of data extraction, compared with 8.36 citations for articles published in 2024 and 3.65 for those published in 2025. This decline is expected because more recent articles have had less time to accumulate citations. Using approximate citation windows adjusted to the extraction date, the mean annual citation rate declined from approximately 3.92 citations per article in 2023 to 3.44 in 2024 and 2.55 in 2025.

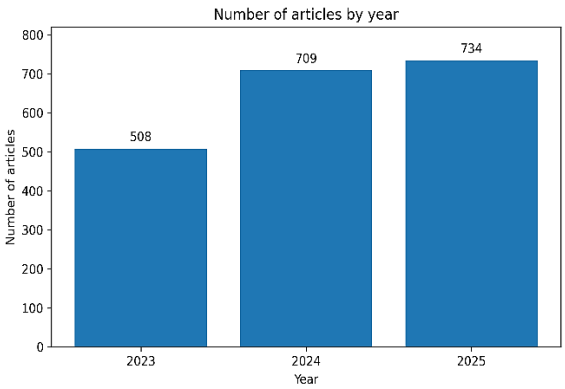


Figure 1. Annual scientific production

Table 2. Average citations by publication year

Year	Art. (n)	Cit. / art.	Cit. window (yrs)	Cit./art. /yr
2023	508	13.46	3.43	3.92
2024	709	8.36	2.43	3.44
2025	734	3.65	1.43	2.55

4.2 Leading countries, Affiliations, and Authors

Table 3 presents the leading countries based on the corresponding authors' affiliations. India accounted for the largest number of articles (367; 18.8%), followed by China (259; 13.3%) and the United States (227; 11.6%). Among the five most productive countries, the United Kingdom had the highest proportion of internationally co-authored articles (45.2%), followed by Saudi Arabia (40.6%) and China (29.0%). The broader geographical distribution of article production is shown in Fig. 2.

Table 3. Leading countries by corresponding author

Country	Art. (n)	Share (%)	International co-authorship (%)
India	367	18.8	20.2
China	259	13.3	29.0
United States	227	11.6	19.4
Saudi Arabia	96	4.9	40.6
United Kingdom	73	3.7	45.2

The leading affiliations and authors are summarised in Table 4. King Saud University was the most productive affiliation, with 42 articles, followed by

Sichuan University with 38 and the University of California San Diego with 36. At the author level, Zhang Y contributed the largest number of articles (17), followed by Wang Y (15) and Li J (12).

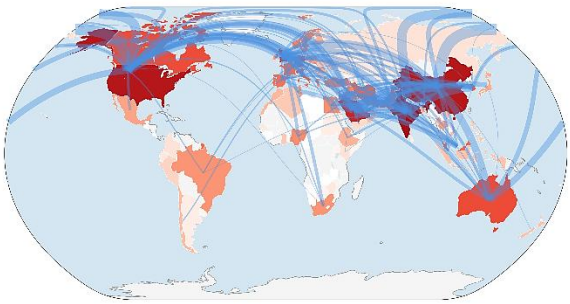


Figure 2. Geographical distribution of articles by corresponding author country

Table 4. Leading affiliations and authors

Affiliation	Art. (n)	Author	Art. (n)
King Saud University	42	Zhang Y.	17
Sichuan University	38	Wang Y.	15
University of California San Diego	36	Li J.	12
Princess Nourah bint Abdulrahman University	32	Kumar K.	11
University of Pennsylvania	31	Wang H.	10

4.3 Influential Journals

The articles were distributed across a large number of journals, indicating that research on cybersecurity and patient data protection in healthcare spans both technical and health-related fields. As shown in Table 5, IEEE Access was the most productive journal, with 51 articles, followed by the Journal of Medical Internet Research with 49 articles and Scientific Reports with 39 articles. Other prominent journals included the IEEE Internet of Things Journal, Sensors, and Electronics. Fig. 3 presents the development of article output among the leading journals during the observed period.

Table 5. Most productive journals

Journal	Art. (n)
IEEE Access	51
Journal of Medical Internet Research	49
Scientific Reports	39
IEEE Internet of Things Journal	28
Sensors	27
Electronics	25
International Journal of Advanced Computer Science and Applications	24
Applied Sciences	22
Cureus Journal of Medical Science	22
BMC Health Services Research	21

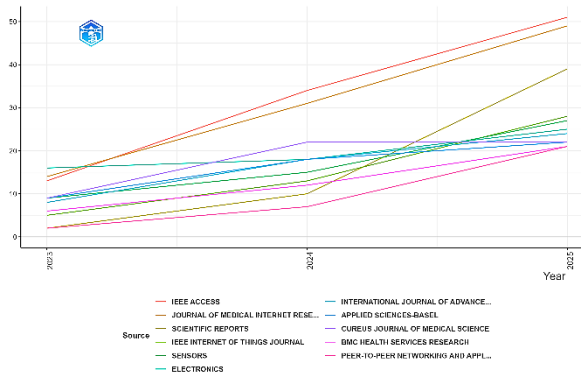


Figure 3. Production of the leading journals over time

Even the most productive journal accounted for only 2.6% of all articles, confirming that the literature is relatively dispersed. However, Bradford's Law identified a smaller group of 38 core journals, representing 4.1% of all sources, which published approximately one-third of the articles. Beyond this core group, the remaining articles were distributed across a much larger number of journals.

4.4 Keyword Analysis and Topic Trends

The most frequent author keywords are presented in Table 6, while Fig. 4 shows their co-occurrence network. Security was the most frequently used keyword, appearing 396 times, followed by blockchain (293), privacy (278), healthcare (236), and cybersecurity (166). Other prominent keywords included internet, artificial intelligence, health, care, and authentication. These results place security, privacy, blockchain, cybersecurity, and authentication at the centre of the field, while broader terms such as healthcare, health, and care reflect the search scope and clinical context.

Table 6. Most frequent author keywords

Keyword	Freq.
Security	396
Blockchain	293
Privacy	278

Keyword	Freq.
Healthcare	236
Cybersecurity	166
Internet	163
Artificial intelligence	142
Health	123
Care	122
Authentication	103

The keyword co-occurrence network reveals several closely connected thematic areas. Security, privacy, blockchain, and authentication form a central group of concepts, while artificial intelligence and related data-driven technologies appear as an increasingly important strand of research.

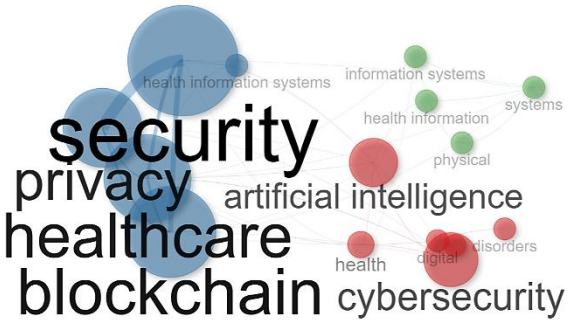


Figure 4. Keyword co-occurrence network

Fig. 5 presents the temporal development of the main trend topics. Security, blockchain, and privacy remained prominent throughout the observed period. AI-related terms became more visible in the later part of the period, with median year of occurrence being 2025. By contrast, COVID-19, mobile health, and privacy protection were more strongly associated with the earlier part of the analysed period. The results suggest a shift from broader digital-health and privacy concerns towards AI-related security challenges and privacy-preserving approaches.

Figure 5. Trend topics in cybersecurity and patient data protection research

4.5 Content Analysis of the Most-Cited Articles

To identify the main practical issues raised in the most-cited literature, we conducted a qualitative full-text content analysis of the ten most-cited articles in the dataset. As shown in Table 7, these articles had received between 119 and 289 citations in the Web of Science Core Collection by the date of data extraction. The selected articles address a diverse range of technologies and implementation settings. AI-related issues are particularly prominent, including privacy-preserving AI, generative AI, large language models (LLMs), and AI-enabled wearable sensors. Other articles focus on metaverse healthcare, internet of medical things (IoMT) vulnerabilities, medical-image encryption, and the adoption of personal health records.

Table 7. Most-cited articles included in the content analysis

Author (yr)	Main focus	Cit.
Khalid et al. (2023)	Privacy-preserving AI and secure health-data use	289
Chengoden et al. (2023)	Metaverse healthcare, interoperability, privacy, and cost	230
Eggmann et al. (2023)	LLM use in dental medicine, reliability, and cybersecurity	204
Paul et al. (2023)	Digital healthcare, IoMT vulnerabilities, and cyber-risk controls	172
Chen and Esmaeilzadeh (2024)	Generative-AI risks across the system life cycle	170
Wu et al. (2023)	Secure transmission and encryption of medical images	168
Jeyaraman et al. (2023)	AI ethics, accountability, transparency, and data stewardship	166
Chen et al. (2025)	Wearable sensors, AI integration, reliability, and infrastructure constraints	124
Wang and Zhang (2024)	LLM applications, data security, hallucinations, bias, and accountability	122
Alsyouf et al. (2023)	Patient adoption of personal health records: privacy, security, and usability	119

Despite their different areas of focus, the selected articles identify several recurring concerns. The most prominent issues relate to the secure use and management of patient data and the growing complexity of digital healthcare systems. The articles identify vulnerabilities affecting EHRs, personal health

records, AI applications, connected devices, wearable sensors, and emerging virtual-care environments. Although these technologies may improve access, monitoring, and clinical decision support, their safe implementation depends on reliable data, privacy protection, secure data transmission, interoperability, appropriate organisational controls, and sufficient technical and human resources. Based on the full-text analysis, we grouped the identified issues into four analytical categories: technical, patient-related, organisational, and resource-related issues. Table 8 summarises the main issues within each category.

Table 8. Main issues identified in the selected articles

Category	Main issues
Technical issues	Protection of sensitive health data; cyberattacks affecting EHRs, connected devices, AI systems, and data transmission; fragmented and non-standardised data; limited interoperability; inaccurate or biased AI outputs; insufficient explainability
Patient-related issues	Privacy and confidentiality concerns; trust in digital systems; usability and ease of use; risk of misinformation; unequal access to digital services; reduced personal interaction
Organisational issues	Unclear accountability; limited governance frameworks; weak cybersecurity preparedness; insufficient monitoring and validation; poor integration into clinical workflows; need for staff training and human oversight
Resource-related issues	High implementation and infrastructure costs; limited curated datasets; computational requirements; battery and energy constraints in wearable devices; affordability and workforce capacity

The content analysis therefore shifts the interpretation from cybersecurity as a technical problem to cybersecurity as an implementation and governance problem. Secure data governance, structured risk management, interoperable infrastructures, clear organisational responsibilities, and resilience under cyberattack or system failure are closely interconnected. Patient trust and system usability are also relevant because secure digital services need to be accepted and used appropriately.

We then mapped these categories interpretatively to areas relevant to the implementation of the NIS2 Directive, including secure data governance, risk management, interoperability, organisational preparedness, and the resilience of healthcare information systems, as shown in Table 9. The mapping is interpretative rather than direct because most of the selected articles do not explicitly examine the NIS2 Directive or assess compliance with its requirements. Nevertheless, their findings provide a practical rationale for the Directive's emphasis on risk management, organisational preparedness, and resilience.

Table 9. Relevance of the identified issues to areas associated with NIS2 implementation

Category	NIS2 relevance
Technical issues	Secure data governance; risk management; interoperability; system resilience
Patient-related issues	Secure data governance; risk management; organisational preparedness
Organisational issues	Organisational preparedness; risk management; interoperability
Resource-related issues	System resilience; organisational preparedness; risk management

5 Conclusion

This study examined recent research on cybersecurity and information security in healthcare, with particular attention to the protection and secure management of patient data. Based on 1,951 articles published between 2023 and 2025, the findings show that the field is expanding rapidly but remains distributed across several technical and health-related domains. Scientific production increased during the observed period, while the dispersion of articles across many journals reflects the interdisciplinary nature of the topic. The keyword analysis shows that security, blockchain, privacy, cybersecurity, and authentication remain central areas of interest. The emergence of artificial intelligence and federated learning points to new challenges associated with the secure use of increasingly data-intensive technologies. These developments are particularly relevant in healthcare, where digital innovation depends on the availability, quality, and secure exchange of sensitive patient data.

The full-text content analysis of the ten most-cited articles identified four closely connected categories of issues: technical, patient-related, organisational, and resource-related. Although technical safeguards remain essential, their effective implementation also depends on reliable data governance, interoperable and resilient information systems, clear organisational responsibilities, patient trust, and adequate resources. The identified issues are closely related to areas relevant to the implementation of the NIS2 Directive in healthcare, particularly risk management, organisational preparedness, and system resilience.

However, this relationship is interpretative rather than direct, as most articles do not explicitly examine the Directive or assess compliance with its requirements.

This study has several limitations. The analysis was restricted to English-language articles indexed in the Web of Science Core Collection and published between 2023 and 2025. Citation indicators should be interpreted cautiously because more recent articles have had less time to accumulate citations. The content analysis was also limited to the ten most-cited articles, while the mapping to areas relevant to NIS2 implementation reflects an analytical interpretation rather than a regulatory assessment. Future research should examine the implementation of cybersecurity measures in real healthcare settings, with particular attention to organisational readiness, interoperability, resource constraints, the economic implications of cybersecurity investment, and the safe integration of AI-based tools into clinical workflows. As healthcare systems become increasingly dependent on digital infrastructure, cybersecurity should be treated as an integral component of patient safety, organisational resilience, and sustainable healthcare management.

References

Alsyouf, A., Lutfi, A., Alsubahi, N., Alhazmi, F. N., Al-Mugheed, K., Anshasi, R. J., Alharbi, N. I., & Albugami, M. (2023). The use of a technology acceptance model (TAM) to predict patients' usage of a personal health record system: The role of security, privacy, and usability. International Journal of Environmental Research and Public Health, 20(2), 1347. https://doi.org/10.3390/ijerph20021347

Aria, M., & Cuccurullo, C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. Journal of Informetrics, 11(4), 959–975. https://doi.org/10.1016/j.joi.2017.08.007

Carello, M. P., Marchetti Spaccamela, A., Querzoni, L., & Angelini, M. (2023). A systematization of cybersecurity regulations, standards and guidelines for the healthcare sector. arXiv. https://arxiv.org/abs/2304.14955

Chen, S., Fan, S., Qiao, Z., Wu, Z., Lin, B., Li, Z., Riegler, M. A., Wong, M. Y. H., Opheim, A., Korostynska, O., Nielsen, K. M., Glott, T., Martinsen, A. C. T., Telle-Hansen, V. H., & Lim, C. T. (2025). Transforming healthcare: Intelligent wearable sensors empowered by smart materials and artificial intelligence. Advanced Materials, 37, 2500412. https://doi.org/10.1002/adma.202500412

Chen, Y., & Esmaeilzadeh, P. (2024). Generative AI in medical practice: In-depth exploration of privacy and security challenges. Journal of Medical Internet Research, 26, e53008. https://doi.org/10.2196/53008

Chengoden, R., Victor, N., Huynh-The, T., Yenduri, G., Jhaveri, R. H., Alazab, M., Bhattacharya, S., Hegde, P., Maddikunta, P. K. R., & Gadekallu, T. R. (2023). Metaverse for healthcare: A survey on potential applications, challenges and future directions. IEEE Access, 11, 12765–12795. https://doi.org/10.1109/ACCESS.2023.3241628

Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W. M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. Journal of Business Research, 133, 285–296. https://doi.org/10.1016/j.jbusres.2021.04.070

Eggmann, F., Weiger, R., Zitzmann, N. U., & Blatz, M. B. (2023). Implications of large language models such as ChatGPT for dental medicine. Journal of Esthetic and Restorative Dentistry, 35(7), 1098–1102. https://doi.org/10.1111/jerd.13046

European Commission. (2026). NIS2 Directive: Securing network and information systems. Shaping Europe's Digital Future. https://digital-strategy.ec.europa.eu/en/policies/nis2-directive

European Parliament and Council of the European Union. (2025). Regulation (EU) 2025/327 on the European Health Data Space. Official Journal of the European Union, L 2025/327. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32025R0327

Jeyaraman, M., Balaji, S., Jeyaraman, N., & Yadav, S. (2023). Unraveling the ethical enigma: Artificial intelligence in healthcare. Cureus, 15(8), e43262. https://doi.org/10.7759/cureus.43262

Khalid, N., Qayyum, A., Bilal, M., Al-Fuqaha, A., & Qadir, J. (2023). Privacy-preserving artificial intelligence in healthcare: Techniques and applications. Computers in Biology and Medicine, 158, 106848. https://doi.org/10.1016/j.compbiomed.2023.106848

Lampropoulos, K., Zarras, A., Lakka, E., Barmpaki, P., Drakonakis, K., Athanatos, M., Debar, H., Alexopoulos, A., Sotiropoulos, A., Tsakirakis, G., Dimakopoulos, N., Tsolovos, D., Pocs, M., Smyrlis, M., Basdekis, I., Spanoudakis, G., Mihaila, O., Prelipcean, B., Salant, E., ... Khabbaz, M. D. (2023). White paper on cybersecurity in the healthcare sector: The HEIR solution. arXiv. https://arxiv.org/abs/2310.10139

Paul, M., Maglaras, L., Ferrag, M. A., & Almomani, I. (2023). Digitization of healthcare sector: A study on privacy and security concerns. ICT Express, 9, 571–588. https://doi.org/10.1016/j.icte.2023.02.007

Pranckutė, R. (2021). Web of Science (WoS) and Scopus: The titans of bibliographic information in

today's academic world. Publications, 9(1), 12. https://doi.org/10.3390/publications9010012

Wang, D., & Zhang, S. (2024). Large language models in medical and healthcare fields: Applications, advances, and challenges. Artificial Intelligence Review, 57, Article 299. https://doi.org/10.1007/s10462-024-10921-0

Wu, Y., Zhang, L., Berretti, S., & Wan, S. (2023). Medical image encryption by content-aware DNA computing for secure healthcare. IEEE Transactions on Industrial Informatics. https://doi.org/10.1109/TII.2022.3194590

Generative Artificial Intelligence in Business Process Modeling: A Systematic Literature Review of Applications, Opportunities, and Limitations

Matija Šajn
University of Zagreb
Faculty of Organization and Informatics
Department of Information Systems Development
Pavlinska 2, Varaždin, Croatia
matija.sajn@foi.unizg.hr

Abstract. *Generative Artificial Intelligence (GenAI) and Large Language Models (LLMs) are increasingly applied to Business Process Modeling, particularly for Business Process Model and Notation (BPMN). This paper presents a systematic literature review of recent research on GenAI-assisted Business Process Modeling. The review identifies the main application areas, opportunities, limitations, and research gaps. The findings show a shift from automated BPMN generation towards validation, interpretation, and collaborative AI-assisted modeling. However, challenges remain regarding semantic correctness, advanced BPMN constructs, standardized evaluation, and empirical validation. The review provides a conceptual synthesis of current capabilities, challenges, and future research directions.*

Keywords. Generative Artificial Intelligence, Large Language Models, Business Process Modeling, Business Process Model and Notation (BPMN), Process Model Generation, Process Model Validation

1 Introduction

Artificial Intelligence (AI), particularly Generative Artificial Intelligence (GenAI) and Large Language Models (LLMs), have attracted considerable attention in Business Process Management (BPM) due to their ability to understand and generate natural language (Zhao et al., 2023), creating new opportunities to automate tasks traditionally performed by business analysts and process modeling experts. As Filep and Molnár (2025) observe in their structured review, the integration of AI into business process modeling now extends beyond LLMs to include machine learning, natural language processing, and robotic process automation, signaling a broader shift towards intelligent, adaptive process ecosystems. Since business process knowledge is frequently documented as textual process descriptions (Nour Eldin et al., 2024), LLMs have emerged as promising tools for transforming such descriptions into formal BPMN models while also supporting model interpretation, validation, and refinement (Ružička & Štancel, 2026).

Business Process Modeling represents a fundamental activity within BPM, enabling organizations to document, analyze, optimize, and automate their business processes (van der Aalst, 2013). BPMN has become the de facto standard for process modeling. It provides a standardized notation that facilitates communication between business and technical stakeholders while supporting process analysis and automation (Object Management Group, 2014). As emphasized by Mendling et al. (2010) and van der Aalst (2013), process models constitute a central organizational asset that supports process improvement and digital transformation initiatives.

The application of LLMs to Business Process Modeling has evolved considerably during the last few years. Early studies primarily investigated the feasibility of automatically generating BPMN models from textual process descriptions (Ružička & Štancel, 2026; Zang et al., 2026). Subsequent research shifted towards decomposing the modeling task into multiple stages by introducing structured intermediate representations that improve model consistency and facilitate validation (Nour Eldin et al., 2024; Nour Eldin et al., 2025). Building upon these developments, Licardo et al. (2026) proposed an interactive BPMN Assistant that enables users to iteratively create and modify process models through natural language. Safan and Köpke (2025a, 2025b) extended LLM-based modeling towards BPMN collaboration diagrams and introduced additional validation mechanisms.

More recently, Kourani et al. (2025a) demonstrated that LLMs can also support process model comprehension by enabling users to query existing BPMN models using natural language. Kourani et al. (2025b) provided one of the first comprehensive benchmark-based evaluations of LLM capabilities for Business Process Modeling. Chen (2025) extended LLM-based BPMN generation to smart contracts,

translating Solidity code into BPMN models through instruction tuning. This extension shows that LLMs can process structured inputs beyond natural language descriptions. These studies reflect a gradual transition from simple text-to-BPMN generation towards comprehensive AI-assisted support throughout the Business Process Modeling lifecycle.

Despite the growing maturity of the field, several important challenges remain unresolved, particularly regarding semantic consistency, incomplete support for advanced BPMN constructs, and the absence of standardized validation procedures. Semantic consistency between textual descriptions and generated BPMN models remains difficult to achieve (Drakopoulos et al., 2026; Nour Eldin et al., 2025). Support for advanced BPMN constructs is often incomplete (Licardo et al., 2026; Safan & Köpke, 2025b). Standardized validation procedures are largely absent, and empirical evidence under realistic conditions is limited (Kourani et al., 2025b). Consequently, the current research landscape remains fragmented, making it difficult to identify prevailing research directions, methodological trends, and unresolved research challenges.

Motivated by the rapid growth of this research field, this paper presents a systematic literature review of recent studies investigating the application of GenAI to Business Process Modeling. By synthesizing the reviewed literature, the study aims to identify the principal application areas, analyze the reported opportunities and limitations, and highlight the research gaps that should guide future research. Accordingly, this review addresses the following research questions:

RQ1: What are the principal application areas and research trends in GenAI-assisted Business Process Modeling?

RQ2: What capabilities and limitations of GenAI-assisted Business Process Modeling are reported in the existing literature?

RQ3: What research challenges and future research directions emerge from the current body of literature?

The paper is organized as follows. After presenting the applied research methodology, the findings of the systematic literature review are reported and discussed, followed by concluding remarks and directions for future research.

2 Research Methodology

This paper presents a systematic literature review reported in accordance with the PRISMA 2020 guidelines (Page et al., 2021). The review focuses on scientific publications investigating the application of GenAI, particularly LLMs, in Business Process Modeling using BPMN. To identify relevant studies, a structured literature search was conducted on May 25, 2026, using the Scopus and Web of Science (WoS) databases. The search queries combined terms related to GenAI, LLMs, Business Process Modeling and BPMN.

The search returned 159 publications indexed in Scopus and 68 publications indexed in WoS, giving a total of 227 records. After merging the search results, duplicate publications were removed using Digital Object Identifier (DOI), publication titles and bibliographic metadata, reducing the dataset to 157 unique studies.

The search strategy and database selection followed established recommendations for conducting systematic literature reviews in information systems research (vom Brocke et al., 2015).

During the first stage, an initial screening of titles and abstracts was supported by ChatGPT as an auxiliary tool. Each publication was evaluated against four predefined criteria: Generative AI/LLM, Business Process Modeling, BPMN, and Empirical Evaluation. For every criterion, a binary decision (Yes/No) was assigned. The obtained results served exclusively as a preliminary recommendation and were not used to determine the final selection of papers.

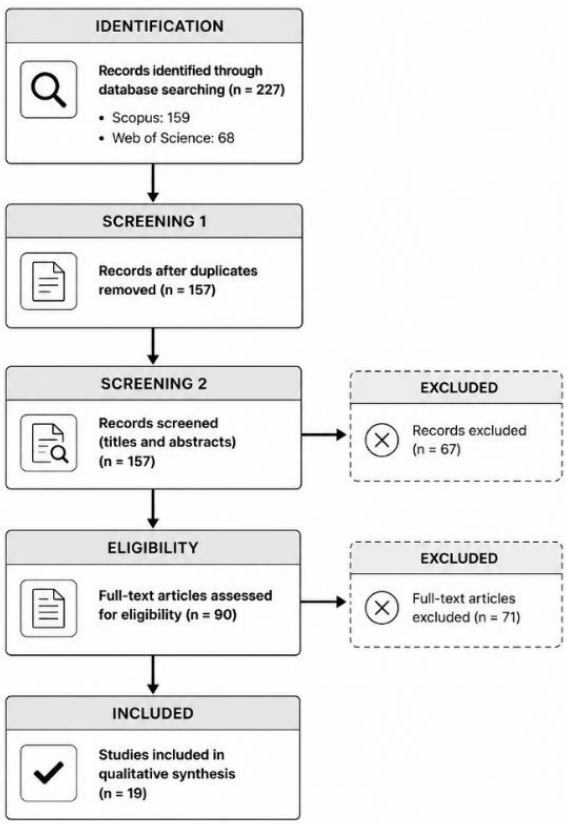


Figure 1. PRISMA flow diagram of the study selection process

The second stage relied entirely on manual assessment. All abstracts were reviewed and evaluated according to five predefined criteria: P1 – Generative AI/LLM, P2 – Business Process Modeling, P3 – BPMN, P4 – Empirical Evaluation, and P5 – Opportunities, Limitations or Challenges. Studies

satisfying P1 and P2 together with at least one of the remaining criteria were considered relevant for further analysis. This procedure resulted in a subset of 90 publications.

An additional qualitative assessment was performed to identify the core literature used in this review. Besides the predefined criteria, each publication was examined with respect to its relevance to the research objectives and its overall contribution to the application of GenAI in Business Process Modeling. Although the preliminary AI-assisted assessment was retained as a consistency check, the final decision was based exclusively on manual evaluation. As a result, 19 studies satisfying at least four of the five predefined criteria were selected for qualitative synthesis.

The complete study selection procedure is presented in the PRISMA flow diagram shown in Fig. 1. Following the concept-centric approach to literature reviews (Webster & Watson, 2002), the review synthesizes existing research by identifying the principal application areas, analyzing the reported opportunities and limitations, and highlighting research gaps that should guide future research.

3 Research Results

The results of the systematic literature review are presented in three parts. First, the geographical distribution of the 19 core papers is analyzed to identify the main research centres contributing to the field. Second, the temporal distribution of the 90 relevant papers is examined to reveal publication trends and the overall growth of research activity. Third, the core papers are classified according to their primary research purpose, providing a structured overview of the main application areas in GenAI-assisted Business Process Modeling.

Fig. 2 presents the geographical distribution of the 19 core papers based on the first author affiliation, grouped into four regional categories. European institutions dominate the sample, accounting for 78.9 % of all papers. The DACH region (Germany, Austria, and Switzerland) accounts for the largest share with 7 papers (36.8 %). The remaining European contributions are spread across seven countries, each represented by a single paper, together accounting for 42.1 %. Outside Europe, Asia contributes three papers (15.8 %), while Africa appears once (5.3 %). This distribution corroborates the leading role of Central European institutions in this domain, which is consistent with the region's established BPM research tradition and its strong presence in process modeling conferences and journals. Although contributions from other regions remain limited, the presence of papers from Asia and Africa indicates a gradual internationalization of the field.

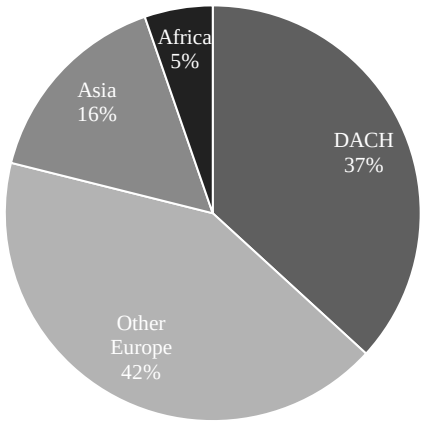


Figure 2. Geographical distribution of core papers by region[1]

Fig. 3 shows the distribution of the 90 relevant papers by publication year. The number of papers rose steadily from 2 in 2022 to 12 in 2023, 25 in 2024, and peaked at 31 in 2025. The year 2026, with 20 papers already identified by May, suggests the upward trend is continuing.

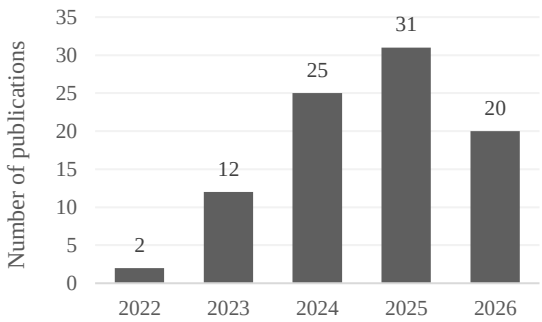


Figure 3. Distribution of relevant papers by publication year

The qualitative analysis of the selected studies identified four predominant research purposes: generation, validation, verification, and interpretation. Generation refers to the automated creation of BPMN models from natural language descriptions (e.g., Nour Eldin et al., 2025; Licardo et al., 2026). Validation encompasses the assessment of model quality, correctness, and semantic alignment (e.g., Drakopoulos et al., 2026; Safan & Köpke, 2025b). Verification involves formal property checking (e.g.,

[1] Based on 19 core papers included in the systematic literature review.

Safan & Köpke, 2025a; Kourani et al., 2024). Interpretation focuses on explaining and querying existing process models using LLMs (e.g., Kourani et al., 2025a). Table 1 maps the 19 core papers according to these categories.

Table 1. Classification of core papers by primary research purpose

Category	No. of papers	Perc.
Generation	12	63.2 %
Validation	4	21.1 %
Verification	2	10.5 %
Interpretation	1	5.3 %

Model generation clearly dominates the current research landscape, representing 12 of the 19 core studies (63.2 %). Validation is addressed in four studies (21.1 %), followed by verification (2 studies, 10.5 %) and interpretation (1 study, 5.3 %). Taken together, these results suggest that although automated BPMN generation remains the primary research focus, increasing attention is being devoted to complementary activities aimed at improving model quality, correctness, and usability.

4 Discussion

The reviewed literature demonstrates significant progress beyond initial experiments with automated BPMN generation. Despite differences in objectives, methodologies, and implementation approaches, several common trends emerge regarding the current capabilities of LLM-based solutions, their limitations, and the research challenges that remain to be addressed.

4.1 Current State of AI-Assisted Business Process Modeling

Current research on GenAI-assisted Business Process Modeling is strongly oriented towards BPMN model generation. However, the reviewed studies also indicate that the field is gradually expanding beyond simple text-to-model transformation towards more comprehensive AI-supported modeling workflows. This evolution is reflected in the variety of approaches proposed for generating, refining, and supporting BPMN models.

Most approaches to BPMN model generation assume that information required to construct a BPMN model can be extracted from natural language process descriptions and represented in a structured form suitable for automated modeling (Nour Eldin et al., 2025; Ružička & Štancel, 2026; Zang et al., 2026). Several studies have demonstrated this concept through prototype systems such as Process Copilot (Caesar et al., 2025), BPMN Assistant (Licardo et al., 2026), Nala2BPMN (Nour Eldin et al., 2025), and BPMN-Chatbot++ (Safan & Köpke, 2025b). These solutions pursue the same objective: reducing the effort required to transform informal business knowledge into formal BPMN models.

An emerging trend in recent research is the growing adoption of multi-stage modeling workflows. Instead of generating BPMN models directly from textual descriptions, several authors introduce structured intermediate representations that separate text interpretation from model construction (Nour Eldin et al., 2024; Licardo et al., 2026; Safan & Köpke, 2025b). Boujmil and Chebil (2025) take a different hybrid approach, combining a traditional Natural Language Processing (NLP) pipeline with GPT-4 through prompt-guided fusion. Their method reduces overgeneration of activities and gateways and improves parallel gateway detection, showing that hybridization can work at the level of model synthesis, not just through intermediate representations. Researchers use intermediate representations, such as JavaScript Object Notation (JSON), formal specifications, and Partially Ordered Workflow Language (POWL) models, to enable partial verification before generating the final BPMN model. This decomposition not only improves model consistency but also enables domain experts to validate intermediate results before completing the modeling process.

The adoption of intermediate representations is motivated not only by technical implementation considerations but also by the need to improve transparency, verification, and user interaction throughout the modeling process. By separating semantic interpretation from BPMN model construction, these approaches reduce the complexity of the generation task and enable incremental verification of process information before the final model is produced. Such decomposition also facilitates user intervention, allowing domain experts to review and refine intermediate results without modifying the complete BPMN model. Consequently, intermediate representations contribute not only to improved model consistency but also to greater transparency and explainability throughout the modeling process (Nour Eldin et al., 2024; Licardo et al., 2026).

LLMs are increasingly viewed as collaborative assistants rather than autonomous model generators. Systems such as BPMN Assistant, BPMN-Chatbot++, and Nala2BPMN support iterative interaction in which users progressively refine process models through natural language instructions. Rather than relying on a single prompt, these approaches adopt iterative prompting strategies that combine user feedback with successive model refinements (Licardo et al., 2026; Safan & Köpke, 2025b). Taken together, these observations suggest that the quality of AI-assisted

Business Process Modeling depends not only on the capabilities of the underlying LLM but also on the design of prompts, the structure of user-system interaction, and the active involvement of domain experts throughout the modeling process. Business Process Modeling therefore evolves from a one-step generation task into a collaborative process in which human expertise and LLM capabilities complement each other.

The growing maturity of LLM-based modeling approaches has also broadened the scope of their application. Recent studies further extend the role of LLMs beyond BPMN model generation by enabling users to interact with existing process models through natural language. Instead of manually analyzing process diagrams, users may query process logic, request explanations of individual activities, or explore relationships between process elements using conversational interfaces. Alharbi et al. (2024) demonstrate this through AIPA, an LLM-based assistant that supports such interactions. Kourani et al. (2025a) extend this line of work with a more comprehensive framework and tool evaluation. Muttillo et al. (2025) explore a different direction, using LLMs to generate synthetic modeling traces and BPMN artifacts for training intelligent modeling assistants in data-scarce environments. While human-generated data remains superior, their results indicate that synthetic data can help address the cold-start problem in process modeling recommenders. These developments suggest that the role of LLMs is gradually extending beyond process generation towards broader support for process understanding and knowledge-intensive modeling activities. At the same time, they also expose several limitations that must be addressed before such approaches can be reliably adopted in practice.

4.2 Challenges and Future Research

Despite the considerable progress achieved in recent years, the reviewed studies consistently indicate that current LLM-based approaches still rely on human supervision or complementary validation mechanisms, suggesting that fully autonomous Business Process Modeling has not yet been achieved. Across the analyzed studies, semantic correctness emerges as the most persistent challenge. A syntactically valid BPMN model does not necessarily preserve the business logic expressed in the original textual description. Drakopoulos et al. (2026) show that this is particularly problematic when processes involve complex decision structures, exception handling, or parallel execution, while Nour Eldin et al. (2025) observe similar issues with incomplete or ambiguous process descriptions. This observation underscores that evaluating generated models solely against BPMN syntax is insufficient and that greater attention should be devoted to semantic fidelity between textual requirements and the resulting process model. A related concern is data privacy, as most LLM-based modeling approaches rely on cloud-based services. Apaydin and Zisgen (2025) address this by proposing a pipeline that uses locally deployed, fine-tuned LLMs to generate process models, demonstrating that smaller open-source models can achieve competitive performance while preserving data confidentiality. Taken together, these findings suggest that future research should explore hybrid deployment strategies that balance model capability with privacy requirements.

Beyond semantic correctness, current limitations also concern both the scope of BPMN support and the evaluation of AI-assisted modeling approaches. While existing solutions successfully generate the core BPMN elements, advanced constructs such as collaboration diagrams, data objects, pools, and lanes remain only partially supported (Licardo et al., 2026; Nour Eldin et al., 2025). In addition, the reviewed studies employ different benchmark datasets, evaluation metrics, experimental settings, and LLM architectures, making a direct comparison of reported results difficult. Real organizational processes frequently depend on these constructs. For processes with complex data-activity relationships, Köpke, Bergenthum, and Kovář (2025) argue that even full BPMN support may be insufficient and propose object-centric modeling as an alternative. Extending LLM capabilities beyond basic process structures should therefore be a priority for future research.

While several studies propose quantitative evaluation procedures, a commonly accepted benchmarking framework has yet to emerge. Establishing standardized datasets and evaluation protocols would improve the reproducibility of future studies and facilitate objective comparison between different modeling approaches.

The reviewed studies consistently indicate that reliable AI-assisted Business Process Modeling cannot yet be achieved through a single prompt or a single LLM invocation. Instead, the most promising approaches combine structured intermediate representations, iterative prompting, and formal validation procedures, all combined with expert supervision. Alayón et al. (2026), for instance, demonstrate this through a two-stage multimodal pipeline where NotebookLM extracts process structure from video, while ChatGPT generates BPMN XML, with domain experts validating the final models. Similarly, Licardo et al. (2026) rely on iterative user feedback and JSON intermediate representations to refine BPMN models. Rather than replacing domain experts, these approaches position LLMs as decision-support tools operating within human-supervised modeling workflows.

The identified research challenges also outline several important directions for future work. In particular, future research should focus on improving semantic fidelity, extending support for advanced BPMN constructs, establishing standardized benchmarking procedures, and developing human-

centered AI-assisted modeling environments. Although empirical evaluation is reported in most reviewed studies, validation is predominantly conducted using prototype implementations, benchmark datasets, controlled experiments, or limited case studies. Evidence from long-term deployment, organizational adoption, and routine use of AI-assisted Business Process Modeling solutions in operational business environments remains largely absent. Consequently, future research should place greater emphasis on longitudinal evaluations and industrial case studies to better understand the practical impact and organizational adoption of these approaches.

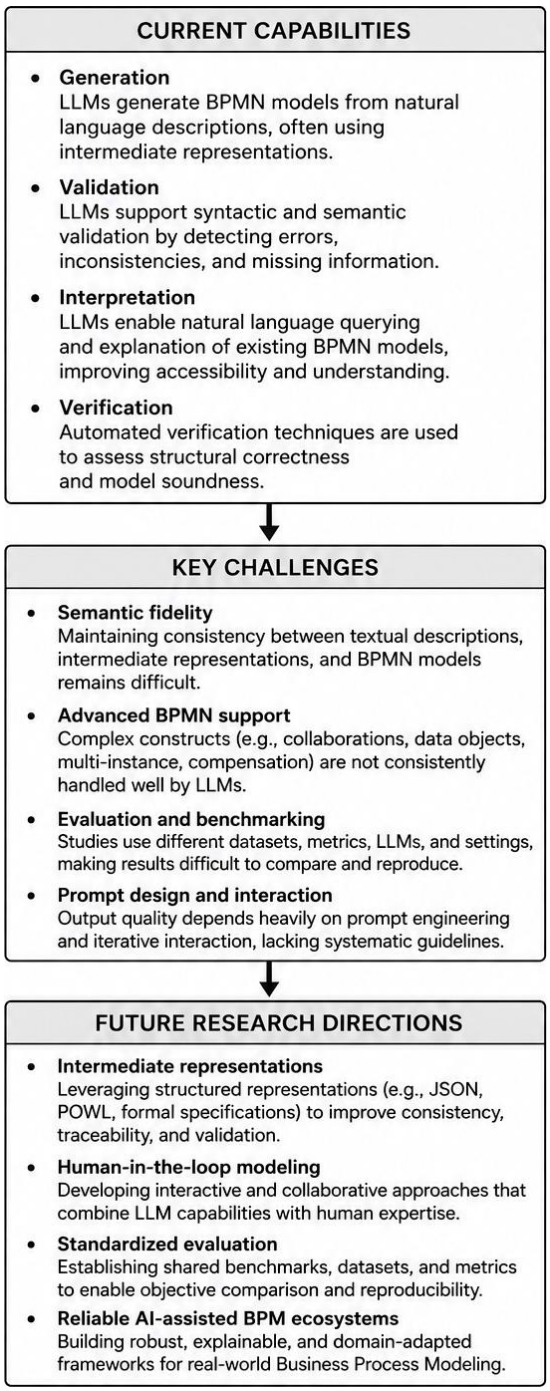


Figure 4. Summary of capabilities, challenges, and future directions

To consolidate these findings, the present review provides a conceptual synthesis of the current state of GenAI-assisted Business Process Modeling, structured around three complementary dimensions: capabilities, challenges, and future research directions. Rather than proposing a prescriptive framework, this synthesis represents an integrated interpretation of the reviewed studies, bringing together fragmented evidence into a coherent perspective. Fig. 4 summarizes this conceptual synthesis by illustrating how the current capabilities of GenAI-assisted Business Process Modeling reveal persistent challenges, which in turn define the principal priorities for future research.

5 Conclusion

This systematic literature review examined how GenAI is being applied to Business Process Modeling and synthesized current evidence on its capabilities, limitations, and future research directions.

The qualitative synthesis reveals that current research remains predominantly focused on automated BPMN model generation, while progressively expanding towards model validation, verification, interpretation, and interactive modeling support. This evolution reflects a broader shift from isolated text-to-model generation towards collaborative AI-assisted modeling approaches that integrate intermediate representations, iterative prompting, validation mechanisms, and human expertise. Overall, these findings demonstrate that the principal application areas of GenAI-assisted Business Process Modeling have expanded well beyond automated model generation, illustrating the rapid evolution and diversification of research in this field. The role of LLMs is therefore evolving from standalone model generators towards intelligent assistants supporting multiple phases of the Business Process Modeling lifecycle.

The reviewed studies demonstrate that recent advances in GenAI have substantially improved the capabilities of AI-assisted Business Process Modeling, particularly in automated model generation, process interpretation, interactive modeling support, and the integration of intermediate representations. Nevertheless, several important challenges remain unresolved. Current evidence consistently highlights issues related to semantic fidelity, limited support for advanced BPMN constructs, the absence of standardized evaluation frameworks, and the need for more comprehensive empirical validation. Addressing these challenges is essential for improving the reliability, transparency, and practical applicability of AI-assisted Business Process Modeling solutions.

The findings should be interpreted in light of several limitations. The review focuses specifically on BPMN and relies on a qualitative, concept-centric synthesis of 19 core papers selected from a broader set of 90 relevant publications. In addition, the literature search was limited to peer-reviewed publications indexed in Scopus and Web of Science at a specific point in time, potentially excluding recent preprints, emerging publication venues, and domain-specific contributions not indexed in these databases. Furthermore, the review did not employ backward or forward snowballing techniques. Future reviews could complement database searches with backward and forward snowballing to identify additional relevant studies. These limitations do not undermine the identified trends but indicate that the results should be interpreted as a structured snapshot of a rapidly evolving research field.

The main contribution of this review is a structured synthesis of current research that integrates the identified application areas, existing limitations, and emerging research directions into a coherent perspective on the evolution of GenAI-assisted Business Process Modeling. The findings suggest that future progress will depend not only on advances in LLM capabilities but also on the development of reliable validation mechanisms, standardized evaluation practices, broader organizational validation, and human-centered AI-assisted modeling frameworks capable of supporting complex real-world business processes.

Acknowledgements

This research has been conducted as part of the wider research in the project Value Perspectives in Digital Transformation Initiatives (VPiDT), which is funded by European Union - NextGenerationEU.

References

Alayón, S., Martín, C., Torres, J., Bacallado, M., Aguilar, R., & Savirón, G. (2026). Transformation of verbal descriptions of process flows into Business Process Modeling and Notation models using multimodal artificial intelligence: Application in justice. Computer Modeling in Engineering & Sciences, 146(2). doi: 10.32604/cmes.2025.073488.

Alharbi, N., Kourani, H., Schuster, D., & van der Aalst, W. M. P. (2024). AIPA: An LLM-Based Assistant for Process Analysis and Comprehension. In 2024 IEEE International Conference on Autonomic Computing (ICAC 2024). IEEE. doi: 10.1109/ICAC64487.2024.10851120.

Apaydin, K., & Zisgen, Y. (2025). Local large language models for business process modeling. In A. Delgado & T. Slaats (Eds.), ICPM 2024 Workshops (Lecture Notes in Business Information Processing, Vol. 533, pp. 605 – 609). Springer. doi: 10.1007/978-3-031-82225-4_44.

Boujmil, I., & Chebil, W. (2025). Improving BPMN model generation from texts through a prompt-guided fusion of NLP and LLM approaches. In Proceedings of the Knowledge Capture Conference 2025 (K-CAP '25). ACM. doi: 10.1145/3731443.3771362.

Caesar, B., Peter, S., Diaz-Tejada, C., Werner, M., Krafft, O., & Mahler, M. (2025). Process Copilot: Speeding up process digitization in the German public sector. In Business Process Management Forum 2025. Springer. doi: 10.1007/978-3-031-96670-5_6.

Chen, L. (2025). Generating structured BPMN models from smart contracts using LLMs. In 2025 IEEE 24th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom 2025) (pp. 3322 – 3327). IEEE. doi: 10.1109/Trustcom66490.2025.00402.

Drakopoulos, P., Malousoudis, P., Nousias, N., Tsakalidis, G., & Vergidis, K. (2026). Do LLMs Speak BPMN? An Evaluation of Their Process Modeling Capabilities Based on Quality Measures. Computation, 14(1), 10. doi: 10.3390/computation14010010.

Filep, R. M., & Molnár, B. (2025). Artificial intelligence in business process modelling: A structured overview of trends, challenges and future research direction. In 2025 IEEE 23rd Jubilee International Symposium on Intelligent Systems and Informatics (SISY 2025). IEEE. doi: 10.1109/SISY67000.2025.11205411.

Köpke, J., Bergenthum, R., & Kovář, J. (2025). KI und BPM als Enabler der Digitalisierung in der Verwaltung. In Joint Proceedings of the STAF 2025 Workshops: OCL, OOPSLE, LLM4SE, ICMM, AgileMDE, AI4DPS, and TTC co-located with the International Conference on Software Technologies: Applications and Foundations (STAF 2025), Koblenz, Germany, June 10 – 13, 2025. CEUR Workshop Proceedings, Vol. 4122, pp. 1 – 6. https://ceur-ws.org/Vol-4122/invited4.pdf

Kourani, H., Berti, A., Hennrich, J., Kratsch, W., Weidlich, R., Li, C.-Y., Arslan, A., van der Aalst, W. M. P., & Schuster, D. (2025a). Leveraging large language models for enhanced process model comprehension. Decision Support Systems, 200, 114563. doi: 10.1016/j.dss.2025.114563.

Kourani, H., Berti, A., Schuster, D., & van der Aalst, W. M. P. (2025b). Evaluating large language

models on business process modeling: Framework, benchmark, and self-improvement analysis. Information Systems, 150, 103360. doi: 10.1016/j.is.2025.103360.

Kourani, H., Berti, A., Schuster, D., & van der Aalst, W. M. P. (2024). Process modeling with large language models. In Enterprise, Business-Process and Information Systems Modeling (Lecture Notes in Business Information Processing, Vol. 511, pp. 229 – 244). Springer. doi: 10.1007/978-3-031-61007-3_18.

Licardo, J. T., Tanković, N., & Etinger, D. (2026). BPMN Assistant: An LLM-Based Approach to Business Process Modeling. Applied Sciences, 16(5), 2213. doi: 10.3390/app16052213.

Mendling, J., Reijers, H. A., & van der Aalst, W. M. P. (2010). Seven process modeling guidelines (7PMG). Information and Software Technology, 52(2), 127 – 136. doi: 10.1016/j.infsof.2009.08.004.

Muttillo, V., Eramo, R., Rubei, R., & Berardinelli, L. (2025). Coupling LLMs and Model-Driven Engineering to Support Synthetic Generation of BPMN Artifacts. In 2025 ACM/IEEE 28th International Conference on Model Driven Engineering Languages and Systems Companion (MODELS-C 2025) (pp. 556 – 565). IEEE. doi: 10.1109/MODELS-C68889.2025.00079.

Nour Eldin, A., Assy, N., Anesini, O., Dalmas, B., & Gaaloul, W. (2024). A Decomposed Hybrid Approach to Business Process Modeling with LLMs. In On the Move to Meaningful Internet Systems: OTM 2024 Conferences (CoopIS 2024) (pp. 243 – 260). Springer. doi: 10.1007/978-3-031-81375-7_14.

Nour Eldin, A., Assy, N., Anesini, O., Dalmas, B., & Gaaloul, W. (2025). Nala2BPMN: Automating BPMN Model Generation with Large Language Models. In Cooperative Information Systems (CoopIS 2024) (Lecture Notes in Computer Science, Vol. 15506, pp. 398 – 404). Springer. doi: 10.1007/978-3-031-81375-7_27.

Object Management Group. (2014). Business Process Model and Notation (BPMN), Version 2.0.2. Object Management Group. https://www.omg.org/spec/BPMN/2.0.2/

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... & Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ, 372, n71. doi: 10.1136/bmj.n71.

Ružička, M., & Štancel, M. (2026). Synthesis of BPMN Models from Text Specification Using GPT Model. Acta Polytechnica Hungarica, 23(2), 47 – 64. doi: 10.12700/APH.23.2.2026.2.3.

Safan, A., & Köpke, J. (2025a). A Framework for LLM-Based Conceptual Modeling: Application to BPMN Collaboration Diagrams. In Companion Proceedings of the 44th International Conference on Conceptual Modeling: Industrial Track, ER Forum, 8th SCME, Doctoral Consortium, Tutorials, Project Exhibitions, Posters and Demos (ER 2025), Poitiers, France, October 20 – 23, 2025. CEUR Workshop Proceedings, Vol. 4099, pp. 177 – 190. https://ceur-ws.org/Vol-4099/forum_paper7.pdf

Safan, A., & Köpke, J. (2025b). BPMN-Chatbot++: LLM-Based Modeling of Collaboration Diagrams with Data. In Proceedings of the Best BPM Dissertation Award, Doctoral Consortium, and Demonstrations & Resources Forum co-located with 23rd International Conference on Business Process Management (BPM 2025), Seville, Spain, August 31 – September 5, 2025. CEUR Workshop Proceedings, Vol. 4032, paper 39. https://ceur-ws.org/Vol-4032/paper-39.pdf

van der Aalst, W. M. P. (2013). Business Process Management: A Comprehensive Survey. ISRN Software Engineering, 2013, 507984. doi: 10.1155/2013/507984.

vom Brocke, J., Simons, A., Riemer, K., Niehaves, B., Plattfaut, R., & Cleven, A. (2015). Standing on the Shoulders of Giants: Challenges and Recommendations of Literature Search in Information Systems Research. Communications of the Association for Information Systems, 37. doi: 10.17705/1CAIS.03709.

Webster, J., & Watson, R. T. (2002). Analyzing the Past to Prepare for the Future: Writing a Literature Review. MIS Quarterly, 26(2), xiii – xxiii. doi: 10.2307/4132319.

Zang, J., Zhang, W., Wang, S., & Li, N. (2026). An LLM-Based Generation and Visualization Method for Cross-Organizational Emergency Response Process Models in Coal–Power Joint Operation Enterprises. In 2026 2nd International Conference on Electrical Automation and Artificial Intelligence (ICEAAI 2026) (pp. 718 – 721). IEEE. doi: 10.1109/ICEAAI68945.2026.11442400.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Tang, J., Liu, Z., Pan, Y., Zhang, W., Zhang, Y., Tian, Q., & Wen, J.-R. (2023). A Survey of Large Language Models. arXiv preprint arXiv:2303.18223. doi: 10.48550/arXiv.2303.18223.

Innovation, Entrepreneurship and Economic Growth

Nikolina Dreven, Lorena Pikl
Generative AI as a Job Resource and Demand: A JD–R Perspective on Work Engagement and Extra-role Behavior

Antonia Kurtović, Nora Miškulin, Vitomir Pavičić
The Invisible Workflow: A Cross-Stakeholder Study of Early-Stage Startup Tracking

Stjepan Mihalic, Tomislav Bakovic, Josip Mihalic
The Critical Steps for Using AI: A Case Study from a Manufacturing Company

Generative AI as a Job Resource and Demand: A JD–R Perspective on Work Engagement and Extra-role Behavior

Nikolina Dreven, Lorena Pikl
University of Zagreb
Faculty of Organization and Informatics
Pavlinska 2, 42 000 Varaždin, Croatia
{ndreven, lpikl}@foi.unizg.hr

Abstract. *This study examines the relationship between GenAI job resources, GenAI job demands, work engagement, and extra-role behavior within the Job Demands–Resources framework. Using survey data from 666 employees in Croatian organizations, a moderated mediation model was tested. The results show that GenAI job resources positively predict extra-role behavior. Although GenAI job resources were not directly related to work engagement, the moderated mediation analysis indicates that work engagement functions as a conditional explanatory mechanism. Specifically, the indirect effect becomes stronger when GenAI job demands are higher, highlighting GenAI's dual role as both a job resource and a job demand.*

Keywords. Generative AI, Job Demands–Resources Model, Work Engagement, Extra-role Behavior

1 Introduction

Generative artificial intelligence (GenAI) is rapidly becoming part of everyday organizational life. Tools based on large language models are increasingly used to search for information, generate and summarize text, support decision-making, and solve complex work-related problems (Brynjolfsson & Mitchell, 2017; Mayer et al., 2025). As a result, GenAI is not only changing how employees perform individual tasks, but also how they experience their work, collaborate with others, and contribute to organizational outcomes (Brynjolfsson & Mitchell, 2017; Karunakaran et al., 2025; Przegalinska et al., 2025).

Research on artificial intelligence (AI) suggests that AI is transforming work, occupations, organizations, and labor markets. Brynjolfsson and Mitchell (2017) argue that advances in machine learning create new opportunities for automation and task redesign, while requiring organizations to reconsider how to combine human and machine capabilities. Similarly, recent integrative research shows that AI may reshape work by complementing employee skills, changing skill requirements, altering access to work opportunities, and creating new forms of workplace inequality (Karunakaran et al., 2025). Evidence from OECD surveys further indicates that both employers and workers are generally positive about the effects of AI on performance and working conditions, although concerns remain regarding job loss, trust, training, and worker consultation (Lane et al., 2023). These findings suggest that the effects of AI at work depend not only on the technology itself, but also on how organizations implement, support, and manage its use.

Although AI has received increasing attention in organizational research, the specific implications of GenAI for employees' motivational and behavioral outcomes remain insufficiently understood. Existing studies indicate that GenAI may support employees by improving access to information, enhancing problem-solving, increasing autonomy, improving work quality, and saving time (Biba & Laumer, 2024; Mas-Machuca et al., 2025). At the same time, GenAI may create new demands by requiring employees to learn new tools, adapt to technological change, and critically evaluate AI-generated outputs (Biba & Laumer, 2024). This dual role suggests that GenAI should not be viewed solely as a productivity-enhancing tool but also as a work-design factor that may simultaneously provide resources and impose demands.

This issue is particularly relevant because employees' responses to AI may shape not only task performance, but also motivational and discretionary behavior at work. Recent research shows that employee–AI collaboration can enhance work engagement by increasing meaningful work and creative self-efficacy, suggesting that AI-enabled work may function as an important motivational resource (L. Sun et al., 2025). However, less is known about whether and how GenAI-related job resources contribute to extra-role behavior, especially through motivational states such as work engagement. To address this gap, the study applies the Job Demands–Resources (JD–R) theory (Bakker et al., 2014; Demerouti et al., 2001) to examine the relationship

between GenAI-related job characteristics, work engagement, and extra-role behavior. Specifically, the study investigates whether GenAI job resources are positively related to extra-role behavior, whether work engagement mediates this relationship, and whether GenAI job demands strengthen the indirect relationship between GenAI job resources and extra-role behavior through work engagement.

The study contributes to the literature in three ways. First, it extends JD–R theory to the emerging context of GenAI use at work by examining GenAI as both a job resource and a job demand. Second, it contributes to research on work engagement by examining engagement as a motivational mechanism through which GenAI job resources may translate into extra-role behavior. Third, it contributes to the growing literature on AI in organizations by shifting attention from broad discussions of productivity and job transformation to employees' perceptions of GenAI and their discretionary behavioral responses at work.

The paper is structured as follows. The next sections present the theoretical framework and develop the hypotheses. The following section describes the research method, sample, and measures. The results section presents the moderated mediation analysis. Finally, the paper discusses theoretical and practical implications, limitations, and directions for future research.

2 Theoretical framework

The JD–R theory elucidates the psychological mechanisms by which job design directly influences work engagement and indirectly affects various individual and organizational outcomes, including job performance (Bakker et al., 2014; Demerouti et al., 2001). The model posits that all job characteristics fall into two broad categories: job demands and job resources (Bakker & Demerouti, 2007). It differentiates between two concurrent psychological processes: the health impairment process and the motivational process. The health impairment process occurs when employees expend sustained physical or psychological effort to manage high job demands, such as time pressure or emotionally taxing work (van Wingerden et al., 2018). Extended exposure to these demands may result in exhaustion and adverse outcomes, with strain serving as a key mediator between job demands and outcomes (Bakker & Demerouti, 2017). Conversely, the motivational process highlights the significance of job resources. Resources such as autonomy, social support, and feedback not only facilitate the attainment of work goals but also fulfill fundamental psychological needs for autonomy, competence, and relatedness (Ryan & Deci, 2000; Schaufeli & Bakker, 2004). Empirical evidence indicates that job resources exert a stronger motivational effect when challenges, as a specific type of job demand, are present (Bakker et al., 2012).

Recent research directly applies the JD–R model to GenAI and conceptualizes it as a source of both job demands and job resources. Particularly relevant is the study by Biba and Laumer (2024), which operationalizes generative artificial intelligence within the JD–R framework as both a job demand and a job resource, and develops measurement scales for both constructs. According to their approach, AI-related demands stem from the need to learn new tools, adapt to technological changes, and invest additional cognitive effort to integrate them effectively into everyday work. At the same time, GenAI can serve as a resource by facilitating information processing, accelerating task completion, improving work quality, and supporting problem-solving.

Empirical evidence demonstrates that when GenAI is perceived as a resource, it is positively correlated with employee work engagement, confirming the activation of the motivational process within the JD–R framework (Biba & Laumer, 2024). Conversely, AI-related demands may constitute an additional burden that, without sufficient resources, can trigger the health impairment process (Bakker & Demerouti, 2017). Qualitative research by Mas-Machuca et al. (2025) further indicates that GenAI can enhance employee autonomy, improve access to information, elevate work quality, and save time, thereby supporting its role as a job resource. However, the same technology may also introduce new demands, such as the necessity for ongoing learning or adaptation, reinforcing its dual role within the JD–R framework. Thus, GenAI should be regarded as a contextual factor that modifies the configuration of job demands and resources in job design.

3 Hypotheses development

Building on the dual conceptualization of GenAI as both a job resource and a job demand, GenAI job resources may contribute to extra-role behavior by increasing employees' capacity to perform beyond formal role requirements. Extra-role behavior refers to discretionary behavior that goes beyond formal job requirements and supports organizational functioning (Organ, 2015; Organ et al., 2006). Examples of such behavior include helping a new employee adjust to the job, sharing the workload with a colleague, and training colleagues to use new equipment (Podsakoff et al., 2000). Extra-role behavior may also involve refraining from unnecessarily emphasizing workplace problems, avoiding blame toward colleagues, tolerating organizational difficulties, and preventing tensions and conflicts with coworkers during times of crisis (Polat, 2009).

By helping employees access information, analyze problems, generate ideas, and complete tasks more efficiently, GenAI may increase their capacity to help colleagues, take initiative, and contribute beyond their formal role. This assumption is consistent with

research showing that job resources and favorable job characteristics are related to organizational citizenship behavior and extra-role performance (Fan et al., 2025; Mihalca et al., 2024; Nguyen et al., 2025). It is also supported by studies showing that job crafting and employee–AI collaboration can foster innovative, proactive, and extra-role behaviors by strengthening employees' fit, motivation, and capacity to take initiative (Ok & Lim, 2022; C. Sun et al., 2025). Therefore, employees who perceive GenAI as a job resource may be more likely to engage in discretionary behaviors that support colleagues and the organization.

H1: Generative AI job resources are positively related to extra-role behavior.

Work engagement is a central motivational mechanism in the JD–R framework and is defined as a positive, fulfilling work-related state characterized by vigor, dedication, and absorption (González-Romá et al., 2006; Schaufeli et al., 2002). Similarly, Byrne (2015) describes work engagement as a moment-to-moment motivational state in which employees are psychologically present, focused, and aligned with work and organizational goals. The same author further emphasizes that work engagement involves investing oneself in work and performing with authenticity, passion, persistence, and energy. Prior studies show that job resources foster work engagement by facilitating goal achievement, supporting personal growth, and satisfying basic psychological needs (Bakker & Demerouti, 2007; Schaufeli & Bakker, 2004). More recent research also confirms the importance of job resources for work engagement across sectors and over time (Hakanen et al., 2024; Van den Broeck et al., 2017; Xanthopoulou et al., 2009).

In the GenAI context, resources such as improved information processing, problem-solving support, and creative idea generation may strengthen employees' sense of competence and effectiveness, thereby increasing engagement. Engaged employees are more likely to show extra-role behavior because they have higher levels of energy, involvement, and persistence. Empirical evidence supports this connection: work engagement has been associated with contextual performance and extra-role behavior and has been shown to mediate the relationship between job characteristics and extra-role outcomes (Christian et al., 2011; Demerouti et al., 2015; Ok & Lim, 2022; Sulea et al., 2012).

H2: Work engagement mediates the relationship between generative AI job resources and extra-role behavior.

Although GenAI can provide job resources, its use may also create job demands. Employees may need to learn new tools, evaluate AI-generated content, and continuously adapt to technological changes, which can increase cognitive effort and skill requirements (Biba & Laumer, 2024). It is important to distinguish between GenAI-related demands as potential challenge and hindrance demands. Demands may support work engagement when they are perceived as meaningful and manageable, but may undermine work engagement when they are excessive, unclear, or poorly supported (Crawford et al., 2010). In line with the JD–R framework, job resources may become especially motivational under demanding conditions (Bakker & Demerouti, 2007, 2017). Therefore, this study argues that GenAI resources are more likely to activate work engagement when GenAI demands create a meaningful need for these resources and help employees cope with AI-enabled work requirements. According to JD–R theory, job demands may contribute to strain when employees lack sufficient resources, but they may also strengthen the motivational role of job resources under demanding conditions (Bakker & Demerouti, 2017; Demerouti et al., 2001).

Bakker et al. (2007) showed that job resources boost work engagement, particularly when job demands are high. In the GenAI context, AI-related resources may be especially valuable when employees face higher demands for learning, adaptation, and skill acquisition. Consequently, GenAI job demands are expected to strengthen the indirect relationship between GenAI job resources and extra-role behavior through work engagement.

H3: Generative AI job demands moderate the indirect relationship between GenAI job resources and extra-role behavior through work engagement, such that the indirect effect is stronger when GenAI job demands are higher.

The conceptual research model (Fig. 1) illustrates the proposed relationships among GenAI job resources, GenAI job demands, work engagement, and extra-role behavior. GenAI job resources are expected to have a direct positive effect on extra-role behavior, while work engagement mediates this relationship. GenAI job demands are proposed to moderate the relationship between GenAI job resources and work engagement, thereby strengthening the indirect effect of GenAI job resources on extra-role behavior through work engagement.

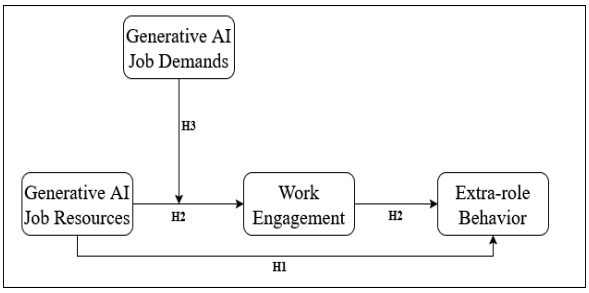


Figure 1. Conceptual research model and hypotheses

4 Method

4.1 Sample

A quantitative online survey questionnaire was used in the study, and primary data were collected from October to December 2025. The data were gathered from 37 Croatian organizations employing at least 20 staff members. The participating organizations came from a broad range of organizational and industry contexts. Based on the organizational information provided by respondents, the sample included employees from private companies, public institutions, and publicly owned organizations. The represented sectors included telecommunications and information technology, retail, manufacturing, food and beverage production, transport and logistics, infrastructure, energy and utilities, financial and insurance services, tourism and hospitality, education, and healthcare. Although the study did not examine industry-level differences, this diversity indicates that the findings are based on respondents from varied organizational settings rather than from a single sector.

Prior to data collection, the study protocol and survey questionnaire were approved by the institutional Ethics Committee. Participation in the study was voluntary and anonymous, and respondents were informed about the purpose of the research before completing the questionnaire.

The sample consisted of 666 respondents. Most participants were 35 years or younger (37.1%) or between 36 and 45 years of age (33.3%), and 56.6% were female. Regarding education, the majority had completed secondary education (36.6%) or held a university degree (29.4%).

In terms of organizational tenure, 40.7% had worked in their organization for 10 years or more, while 27.3% had less than three years of tenure. Most respondents were executive/operational employees (44.3%) or professional administrative staff (32.6%), with fewer participants in lower and middle management (20.9%) and top management (2.3%).

4.2 Measures

The study examined job characteristics related to the use of GenAI applications at work. Job characteristics were measured using the scale developed by Biba and Laumer (2024), who conceptualize GenAI as both a job demand and a job resource within the Job Demands–Resources (JD–R) framework. Their instrument builds on the Work Design Questionnaire (WDQ), originally developed by Morgeson and Humphrey (2006) and adapts it to the GenAI context to capture AI-related cognitive enhancements and skill requirements. Specifically, the scale includes items reflecting AI-related demand (acquiring skills) and AI-related resources (information processing and problem-solving). Although the scales adopted from Biba and Laumer (2024) are relatively recent, they demonstrated satisfactory psychometric properties in the present study. Nevertheless, additional validation across different contexts and samples is warranted to further establish their generalizability.

Work engagement was measured using the Utrecht Work Engagement Scale (UWES) developed by Schaufeli et al. (2002). The scale captures the three core dimensions of engagement: vigor, dedication, and absorption. Extra-role behavior was measured using the Organizational Citizenship Behavior Scale (3 items) developed by Motowidlo (2000). All items were measured using a 5-point Likert scale (1 = strongly disagree; 5 = strongly agree).

All measurement items were administered in Croatian. The original English items were translated into Croatian by the research team and reviewed to ensure conceptual equivalence, clarity, and suitability for the Croatian organizational context. Where necessary, minor wording adjustments were made to preserve the original meaning while making them understandable to Croatian respondents. The GenAI job resources and GenAI job demands scales were adapted to refer specifically to the use of GenAI applications in the workplace.

Table 1 presents the descriptive statistics for all items (N = 666). AI-related resource items show moderate mean values (M = 2.944–3.153), while AI-related demand items are slightly lower (M = 2.808–2.977), indicating moderate perceptions of GenAI as both a resource and a demand. Work engagement items show moderate to relatively high mean values (M = 3.069–4.033), indicating generally positive engagement among respondents. Extra-role behavior items show relatively high mean values (M = 4.030–4.198), indicating that respondents generally report frequently engaging in discretionary behaviors such as taking initiative and investing additional effort beyond formal job requirements.

Table 1. Descriptive statistics of items

Items	M	SD
Utilizing GenAI apps enhances my ability to interact with vast amounts of information.	3.096	1.136
Utilizing GenAI apps enhances my ability to reflect more deeply and think critically.	2.944	1.123
Utilizing GenAI apps enhances my ability to consider multiple perspectives and ideas.	3.153	1.141
Utilizing GenAI apps enhances my ability to analyze and evaluate information.	3.086	1.120
Utilizing GenAI apps enhances my ability to find solutions for problems without clear or definitive answers.	3.066	1.114
Using GenAI apps enhances my creativity and ability to explore innovative ideas.	3.050	1.137

Using GenAI apps enhances my ability to address problems I have not encountered before.	3.062	1.138
Utilizing GenAI apps enhances my ability to develop novel approaches to problems.	3.132	1.125
Utilizing GenAI apps requires a variety of skills.	2.977	1.114
Using GenAI apps requires me to draw on a variety of skills.	2.919	1.098
Utilizing GenAI apps requires me to use several high-level skills.	2.808	1.083
Using GenAI apps requires several skills.	2.856	1.097
When I get up in the morning, I feel like going to work.	3.326	1.098
My job inspires me.	3.396	1.059
At my job, I feel strong and vigorous.	3.368	1.056
At my work, I feel bursting with energy.	3.069	1.061
I am enthusiastic about my job.	3.414	1.039
I am proud of the work that I do.	3.842	1.032
I feel happy when I am working intensely.	3.707	1.066
I get carried away when I am working.	3.751	1.010
I am immersed in my work.	4.033	.974
I do tasks beyond those required without being asked.	4.030	1.006
I put in additional effort at work to complete tasks successfully when necessary.	4.198	0.863
I put forth additional effort in my job to make the business a success.	4.107	0.948

4.3 Data analysis

Common method bias was addressed using both procedural and statistical remedies. Procedurally, participation was voluntary and anonymous, respondents were informed about the purpose of the study, and no personal identifiers were collected. The questionnaire also included clear instructions, and respondents were assured that there were no right or wrong answers. Statistically, Harman's single-factor test was conducted by entering all measurement items into an unrotated exploratory factor analysis and extracting a single factor.

Before testing the hypothesized relationships, the measurement model was assessed. A confirmatory factor analysis (CFA) was conducted in JASP to examine the distinctiveness of the four constructs: GenAI job demands, GenAI job resources, work engagement, and extra-role behavior. In addition, reliability, convergent validity, and discriminant validity were assessed using SmartPLS. Internal consistency reliability was evaluated using Cronbach's alpha and composite reliability (CR), convergent validity was assessed using the average variance extracted (AVE), and discriminant validity was examined using the heterotrait–monotrait ratio of correlations (HTMT).

The hypotheses were tested using the PROCESS macro for SPSS (Hayes, 2018), Model 7. A moderated mediation model was estimated, with GenAI job resources as the independent variable, work engagement as the mediator, extra-role behavior as the dependent variable, and GenAI job demands as the moderator of the relationship between GenAI job resources and work engagement. The analysis was conducted using 5,000 bootstrap samples and 95% confidence intervals. The index of moderated mediation was used to determine whether the indirect effect of GenAI job resources on extra-role behavior through work engagement varied significantly across levels of GenAI job demands.

5 Results

The four-factor CFA model demonstrated an acceptable fit to the data, χ^2 (246) = 1104.700, $p < .001$, $\chi^2/df = 4.49$, CFI = .946, TLI = .939, RMSEA = .073, and SRMR = .043. Although the chi-square test was significant, this result was expected given the large sample size. Overall, the CFA results support the adequacy of the four-factor measurement model.

Reliability and convergent validity results are presented in Table 2. Internal consistency reliability was supported, as Cronbach's alpha values ranged from .886 to .963, while CR values ranged from .929 to .972, exceeding the recommended threshold of .70. Convergent validity was also supported, as all AVE values ranged from .648 to .897 and were above the recommended threshold of .50. All HTMT values were below the conservative threshold of .85, with the highest value observed between GenAI job resources and GenAI job demands (HTMT = .681). These results indicate satisfactory reliability, convergent validity, and discriminant validity of the measurement model.

Table 2. Reliability, convergent validity, and discriminant validity

Construct	Cronbach's α	CR	AVE
GenAI job resources	.962	.969	.817
GenAI job demands	.967	.972	.897
Work engagement	.932	.943	.648
Extra-role behavior	.882	.929	.815

Note. CR = composite reliability; AVE = average variance extracted.

Harman's single-factor test was conducted to assess potential common method bias. All measurement items were entered into an unrotated exploratory factor analysis, and a single factor was extracted. The results showed that the single factor explained 41.55% of the total variance, which is below the commonly used threshold of 50%. This suggests that common method bias is unlikely to be a major concern in this study.

Table 3 presents the results of the moderated mediation analysis. GenAI job resources had a positive and significant direct effect on extra-role behavior (B = .061, SE = .029, p = .032), supporting H1. This suggests that employees who perceive GenAI as a useful job resource are more likely to engage in behaviors that go beyond their formal job requirements. The regression model predicting work engagement was statistically significant, F (3, 662) = 19.72, p < .001, explaining 8.2% of the variance in work engagement (R^2 = .082).

The direct relationship between GenAI job resources and work engagement was not significant (B = .041, SE = .073, p = .574). However, work engagement was strongly and positively related to extra-role behavior (B = .504, SE = .035, p < .001). In addition, GenAI job demands were negatively associated with work engagement (B = −.165, SE = .084, p = .049). In contrast, the interaction between GenAI job resources and GenAI job demands was significant (B = .061, SE = .023, p = .008). The interaction term explained an additional 1.0% of the variance in work engagement beyond the main effects (ΔR^2 = .010, F = 6.99, p = .008), indicating a modest but statistically significant moderation effect. This indicates that GenAI job demands moderate the relationship between GenAI job resources and work engagement, such that GenAI job resources become more positively associated with work engagement as GenAI job demands increase. The regression model predicting extra-role behavior was also statistically significant, F (2, 663) = 126.06, p < .001, explaining 27.6% of the variance in extra-role behavior (R^2 = .276).

Table 3. Results of moderated mediation analysis

	Work engagement	Extra-role behavior
GenAI-JR	.041 (.073)	.061* (.029)
GenAI-JD	−.165* (.084)	—
GenAI-JR × GenAI-JD	.061** (.023)	—
Work engagement	—	.504*** (.035)

Note. Unstandardized coefficients are reported, with standard errors shown in parentheses. p < .05, ** p < .01, *** p < .001.

Table 4 shows that the index of moderated mediation was statistically significant (Index = .031, Boot SE = .014, 95% CI [.004, .057]), as the confidence interval did not include zero. The results do not support H2 in its simple mediation form. GenAI job resources were not significantly related to work engagement, indicating that work engagement does not function as a general mediator between GenAI job resources and extra-role behavior. However, the significant interaction between GenAI job resources and GenAI job demands shows that the relationship between GenAI job resources and work engagement depends on the level of GenAI job demands. Furthermore, the significant index of moderated mediation indicates that the indirect effect of GenAI job resources on extra-role behavior through work engagement is conditional on GenAI job demands. Therefore, work engagement should be interpreted as a conditional explanatory mechanism rather than as a universal mediator. Accordingly, H2 is not supported in its simple form, while H3 is supported.

Table 4. Index of moderated mediation

	Index	Boot SE	Boot LLCI	Boot ULCI
GenAI-JD	.031	.014	.004	.057

6 Discussion, contribution and limitations

The findings support the dual role of GenAI within the JD–R framework. GenAI job resources were positively associated with extra-role behavior, suggesting that when employees perceive GenAI as useful for information processing, problem-solving, and idea generation, they are more likely to engage in discretionary behaviors beyond formal job requirements. This finding is consistent with JD–R research showing that job resources and favorable job characteristics can support positive performance-related outcomes (Demerouti et al., 2015; Mihalca et al., 2024; Schaufeli & Bakker, 2004), as well as with studies linking job resources and new ways of working to organizational citizenship and extra-role behavior (Fan et al., 2025; Nguyen et al., 2025). The results also align with emerging AI-related research suggesting that employee–AI collaboration may foster proactive behavior by reducing workload and strengthening employees' capacity to take initiative (C. Sun et al., 2025).

One of the most interesting findings of this study is that GenAI job resources were not directly related to work engagement. This suggests that the mere availability or perceived usefulness of GenAI tools may not be sufficient to motivate employees. GenAI resources may not automatically increase vigor, dedication, and absorption unless the work context creates a meaningful need for these resources. When employees face higher GenAI-related demands, such

as the need to learn new skills, evaluate AI-generated outputs, or adapt to AI-enabled work processes, GenAI resources may become more salient and motivationally relevant. This finding refines the JD–R perspective by showing that the motivational role of GenAI resources depends on the demand context in which they are embedded.

The study contributes to the literature in three ways. First, it extends the JD–R model to the emerging context of GenAI use at work by examining GenAI simultaneously as a job resource and a job demand, following the approach proposed by Biba and Laumer (2024). Second, it contributes to research on work engagement by showing that engagement functions as a conditional explanatory mechanism rather than a universal mediator. Although GenAI job resources were not directly related to work engagement, the indirect effect on extra-role behavior became stronger when GenAI job demands were higher. This supports the JD–R assumption that job resources may become more salient and motivational under demanding work conditions (Bakker et al., 2007, 2014). Third, the study contributes to research on AI in organizations by shifting attention from broad discussions of productivity and automation to employees' perceptions of GenAI and their discretionary behavioral responses at work.

From a managerial perspective, the findings suggest that organizations should avoid treating GenAI implementation as a purely technical process. Introducing GenAI tools is unlikely to enhance work engagement on its own. Managers should ensure that employees understand how GenAI can support their work, provide sufficient training, and create opportunities for employees to use GenAI in tasks where it has clear practical value. At the same time, organizations should be careful not to interpret higher GenAI-related demands as automatically beneficial. Demands may become motivational only when they are manageable and accompanied by adequate resources, support, and skill development. Without such support, GenAI demands may function as hindrance demands, reducing work engagement.

Several limitations should be noted. First, the study used cross-sectional survey data, which limits causal interpretation of the relationships among GenAI job resources, work engagement, and extra-role behavior. Future studies should use longitudinal or experimental designs to examine how employees' perceptions of GenAI change over time. Second, the data were collected from Croatian organizations, which may limit the generalizability of the findings to other cultural or organizational contexts. Third, all variables were measured using self-reported data collected from the same respondents, which may increase the risk of common method bias. Although procedural remedies and Harman's single-factor test were used to assess and reduce this concern, common method bias cannot be fully excluded. Future research should use multi-source designs, particularly by collecting supervisor or peer ratings of extra-role behavior, as such ratings would provide a more objective assessment of discretionary employee behavior. Nevertheless, the high reliability values for GenAI job resources and GenAI job demands suggest that future studies could further examine whether shorter versions of these scales may reduce potential item redundancy while preserving content validity.

References

Bakker, A. B., & Demerouti, E. (2007). The Job Demands-Resources model: State of the art. *Journal of Managerial Psychology*, *22*(3), 309–328.

Bakker, A. B., & Demerouti, E. (2017). Job demands–resources theory: Taking stock and looking forward. *Journal of Occupational Health Psychology*, *22*(3), 273–285. https://doi.org/10.1037/ocp0000056

Bakker, A. B., Demerouti, E., & Sanz-Vergel, A. I. (2014). Burnout and Work Engagement: The JD–R Approach. *Annual Review of Organizational Psychology and Organizational Behavior*, *1*(1), 389–411. https://doi.org/10.1146/annurev-orgpsych-031413-091235

Bakker, A. B., Hakanen, J. J., Demerouti, E., & Xanthopoulou, D. (2007). Job resources boost work engagement, particularly when job demands are high. *Journal of Educational Psychology*, *99*(2), 274–284. https://doi.org/10.1037/0022-0663.99.2.274

Bakker, A. B., Tims, M., & Derks, D. (2012). Proactive personality and job performance: The role of job crafting and work engagement. *Human Relations*, *65*(10), 1359–1378. https://doi.org/10.1177/0018726712453471

Biba, S., & Laumer, S. (2024). Generative AI as Job Demand and Resource: First Results of an Empirical Study. *Proceedings of the Twenty-Ninth DIGIT Workshop*.

Brynjolfsson, E., & Mitchell, T. (2017). What can machine learning do? Workforce implications. *Science*, *358*(6370), 1530–1534. https://doi.org/10.1126/science.aap8062

Byrne, Z. S. (2015). *Understanding employee engagement: Theory, research, and practice*. Routledge.

Christian, M. S., Garza, A. S., & Slaughter, J. E. (2011). Work Engagement: A Quantitative Review and Test of Its Relations with Task and Contextual Performance. *Personnel Psychology*, *64*(1), 89–136. https://doi.org/10.1111/j.1744-6570.2010.01203.x

Crawford, E. R., LePine, J. A., & Rich, B. L. (2010). Linking job demands and resources to employee engagement and burnout: A theoretical extension and meta-analytic test. *Journal of Applied Psychology*, *95*(5), 834–848. https://doi.org/10.1037/a0019364

Demerouti, E., Bakker, A. B., & Gevers, J. M. P. (2015). Job crafting and extra-role behavior: The role of work engagement and flourishing. *Journal of Vocational Behavior*, *91*, 87–96. https://doi.org/10.1016/j.jvb.2015.09.001

Demerouti, E., Bakker, A. B., Nachreiner, F., & Schaufeli, W. B. (2001). The Job Demands-Resources Model of Burnout. *Journal of Applied Psychology*, *86*(3), 499–512. https://doi.org/10.1037//0021-9010.86.3.499

Fan, L., Yang, J., Hou, Q., Feng, S., Chen, A., & Han, L. (2025). Impact of job resources on organizational citizenship behavior among primary care staff in Guangzhou, China: A cross-sectional study. *Frontiers in Public Health*, *13*, 1580148. https://doi.org/10.3389/fpubh.2025.1580148

González-Romá, V., Schaufeli, W. B., Bakker, A. B., & Lloret, S. (2006). Burnout and work engagement: Independent factors or opposite poles? *Journal of Vocational Behavior*, *68*(1), 165–174. https://doi.org/10.1016/j.jvb.2005.01.003

Hakanen, J. J., Bakker, A. B., & Turunen, J. (2024). The relative importance of various job resources for work engagement: A concurrent and follow-up dominance analysis. *BRQ Business Research Quarterly*, *27*(3), 227–243. https://doi.org/10.1177/23409444211012419

Hayes, A. F. (2018). *Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach* (Second edition). The Guilford Press.

Karunakaran, A., Lebovitz, S., Narayanan, D., & Rahman, H. A. (2025). Artificial Intelligence at Work: An Integrative Perspective on the Impact of AI on Workplace Inequality. *Academy of Management Annals*, *19*(2), 693–735. https://doi.org/10.5465/annals.2023.0230

Lane, M., Williams, M., & Broecke, S. (2023). *The impact of AI on the workplace: Main findings from the OECD AI surveys of employers and workers* (OECD Social, Employment and Migration Working Papers No. 288; OECD Social, Employment and Migration Working Papers, Vol. 288). https://doi.org/10.1787/ea0a0fe1-en

Mas-Machuca, M., Akhmedova, A., & Marimon, F. (2025). Generati/ve AI and Workplace Productivity: A Qualitative Study in Spain. *Proceedings of Central European Conference on Information and Intelligent Systems*, 625–631.

Mayer, H., Yee, L., Chui, M., & Roberts, R. (2025). *Superagency in the workplace: Empowering people to unlock AI's full potential*. McKinsey & Company. https://www.mckinsey.com/capabilities/tech-and-ai/our-insights/superagency-in-the-workplace-empowering-people-to-unlock-ais-full-potential-at-work

Mihalca, L., Ratiu, L., Helm, C., Brendea, G., & Metz, D. (2024). The relationship of job characteristics with in-role and extra-role performance: The mediating effect of job crafting. *Baltic Journal of Management*, *19*(2), 163–182. https://doi.org/10.1108/BJM-05-2023-0191

Morgeson, F. P., & Humphrey, S. E. (2006). The Work Design Questionnaire (WDQ): Developing and validating a comprehensive measure for assessing job design and the nature of work. *Journal of Applied Psychology*, *91*(6), 1321–1339. https://doi.org/10.1037/0021-9010.91.6.1321

Motowidlo, S. J. (2000). Some Basic Issues Related to Contextual Performance and Organizational Citizenship Behavior in Human Resource Management. *Human Resource Management Review*, *10*(1), 115–126. https://doi.org/10.1016/S1053-4822(99)00042-X

Nguyen, P. N.-D., Giang, T. T., Duong, H. C., & Nguyen, H. H. (2025). Empowering Workplaces: How New Ways of Working Foster Organizational Citizenship Behaviors? *Sage Open*, *15*(4), 21582440251393887. https://doi.org/10.1177/21582440251393887

Ok, C. M., & Lim, S. (Edward). (2022). Job crafting to innovative and extra-role behaviors: A serial mediation through fit perceptions and work engagement. *International Journal of Hospitality Management*, *106*, 103288. https://doi.org/10.1016/j.ijhm.2022.103288

Organ, D. W. (2015). Organizational Citizenship Behavior. In *International Encyclopedia of the Social & Behavioral Sciences: Second Edition*. https://doi.org/10.1016/B978-0-08-097086-8.22031-X

Organ, D. W., Podsakoff, P. M., & MacKenzie, S. B. (2006). *Organizational Citizenship Behavior: Its Nature, Antecedents, and Consequences*. SAGE Publications.

Podsakoff, P. M., MacKenzie, S. B., Paine, J. B., & Bachrach, D. G. (2000). Organizational Citizenship Behaviors: A Critical Review of the Theoretical and Empirical Literature and Suggestions for Future Research. *Journal of Management*, *26*(3).

Polat, S. (2009). Organizational citizenship behavior (OCB) display levels of the teachers at secondary schools according to the perceptions of the school administrators. *Procedia - Social and Behavioral*

Sciences, *1*(1), 1591–1596. https://doi.org/10.1016/j.sbspro.2009.01.280

Przegalinska, A., Triantoro, T., Kovbasiuk, A., Ciechanowski, L., Freeman, R. B., & Sowa, K. (2025). Collaborative AI in the workplace: Enhancing organizational performance through resource-based and task-technology fit perspectives. *International Journal of Information Management*, *81*, 102853. https://doi.org/10.1016/j.ijinfomgt.2024.102853

Ryan, R. M., & Deci, E. L. (2000). Self-Determination Theory and the Facilitation of Intrinsic Motivation, Social Development, and Well-Being. *American Psychologist*, *55*(1), 68–78.

Schaufeli, W. B., & Bakker, A. B. (2004). Job demands, job resources, and their relationship with burnout and engagement: A multi-sample study. *Journal of Organizational Behavior*, *25*(3), 293–315. https://doi.org/10.1002/job.248

Schaufeli, W., Salanova, M., González-Romá, V., & Bakker, A. (2002). The Measurement of Engagement and Burnout: A Two Sample Confirmatory Factor Analytic Approach. *Journal of Happiness Studies*, *3*(1), 71–92. https://doi.org/10.1023/A:1015630930326

Sulea, C., Virga, D., Maricutoiu, L. P., Schaufeli, W., Zaborila Dumitru, C., & Sava, F. A. (2012). Work engagement as mediator between job characteristics and positive and negative extra-role behaviors. *Career Development International*, *17*(3), 188–207. https://doi.org/10.1108/13620431211241054

Sun, C., Zhao, X., Guo, B., & Chen, N. (2025). Will Employee–AI Collaboration Enhance Employees' Proactive Behavior? A Study Based on the Conservation of Resources Theory. *Behavioral Sciences*, *15*(5), 648. https://doi.org/10.3390/bs15050648

Sun, L., Hu, R., & Su, H. (2025). Unlocking human potential in the AI Age: How employee-AI collaboration transforms work engagement through dual psychological pathways. *Frontiers in Psychology*, *16*, 1705671. https://doi.org/10.3389/fpsyg.2025.1705671

Van den Broeck, A., Elst, T. V., Baillien, E., Sercu, M., Schouteden, M., De Witte, H., & Godderis, L. (2017). Job Demands, Job Resources, Burnout, Work Engagement, and Their Relationships: An Analysis Across Sectors. *Journal of Occupational & Environmental Medicine*, *59*(4), 369–376. https://doi.org/10.1097/JOM.0000000000000964

van Wingerden, J., Derks, D., & Bakker, A. B. (2018). Facilitating interns' performance: The role of job resources, basic need satisfaction and work engagement. *Career Development International*, *23*(4), 382–396. https://doi.org/10.1108/CDI-12-2017-0237

Xanthopoulou, D., Bakker, A. B., Demerouti, E., & Schaufeli, W. B. (2009). Reciprocal relationships between job resources, personal resources, and work engagement. *Journal of Vocational Behavior*, *74*(3), 235–244. https://doi.org/10.1016/j.jvb.2008.11.003

Acknowledgments

This work was supported by European Union funds through the NextGenerationEU instrument under the project number 2025-657.

Acknowledgment of the use of artificial intelligence (AI)

AI tools were used only for language-related assistance, specifically for translation and for checking grammar, spelling, and style. The authors take full responsibility for the content, analysis, interpretation of results, and conclusions presented in the paper.

The Invisible Workflow: A Cross-Stakeholder Study of Early-Stage Startup Tracking

Antonia Kurtović, Nora Miškulin, Vitomir Pavičić
RIT Croatia
Damira Tomljanovića Gavrana 15
10 000 Zagreb, Croatia
{ak4438, nm1128, vp4596}@g.rit.edu

***Abstract**. This research investigated early-stage startup tracking within the ecosystem. It's based on 57 interviews with different stakeholders (founders, incubators, mentors, investors) to get a cross-stakeholder view. The aim was to identify inefficiencies in current support systems, to help founders navigate their startup journeys more effectively, and investors, mentors, and incubators to make more informed decisions. Findings showed founders rely on fragmented workflows and optimistic self-reporting, incubators face manual tracking burdens, while mentors and investors rely on intuition. This cross-stakeholder analysis revealed a system-wide lack of shared visibility into startup progress. Recommendations include standardized reporting, ecosystem coordination, and AI-powered progress tracking.*

Keywords. early-stage startups, progress tracking, startup ecosystem, startups, incubators, mentorship, investors, qualitative research

1 Introduction

With the everyday emergence of new digital tools, it is becoming easier and easier for entrepreneurs to go from an abstract thought within one's mind to a first prototype, or even a product that can be sold to thousands of customers. There are many tools out there that can help out with ideation, market validation, customer research, prototyping, and other common phases of startup development that founders go through. Actually, it seems like, nowadays, a person does not even need to have the initial idea, even that can be brainstormed together with digital assistants powered by artificial intelligence (AI). Of course, while on one side, these tools are decreasing the time, money, and perhaps even effort required to launch a new startup, on the other, that means that the competition is becoming fiercer than ever (Information Age, 2022). Products are not launched anymore within months, but within weeks or sometimes even hours. While in this research we will not go into determining whether an ocean of copy-cats and ChatGPT wrappers can be called a true innovation, one thing definitely stands clear – digital tools powered by artificial intelligence are reshaping how founders are building. From automatized workflows, over vibecoding, to AI-cofounders, today's entrepreneurs are living in an exciting era. But in this startup world that is defined by speed, uncertainty, and constant change, it seems that systems designed to track and support founders' progress haven't kept pace. One of the common places where founders seek support and often kick-start their journeys are startup incubators (European Commission, 2018). There, they are part of pre-incubation, incubation or acceleration programs (depending on the startup stage). These programs vary in length, but almost always include workshops, mentorship support, and, in some cases, funding or certain perks (like office space or digital tool subscriptions). While all incubators and mentors within their networks do their best to support the new batches of founders, and try to stand out from each other by emphasizing the scope of their support, the systems they use to track, communicate, and validate progress remain surprisingly fragmented and informal. This issue spreads to investors and venture capital funds as well. Although, on average, a bit more technically advanced than incubators (because they are more likely to use internal software for manual progress tracking), investors actually have no way to see what is actually happening inside a startup on a weekly basis (Tykova et al., 2017). Startups seem like black holes – no one knows what is happening within them, sometimes not even the founders themselves. Has the time come for the support given to the founders to become as innovative as the founders themselves? And is our initial hypothesis (of the startup ecosystem draining into fragmented and manual tools) even correct? We are soon to find out.

But before we bother ourselves with the research and results, we should ask one fundamental question – why does this matter? Startup stakeholders across the ecosystem rely on each other. Founders need good resources, tools, mentorship, and funds to increase their chances of success. Incubators, mentors, and investors need to supply these at the right amount and at the right point in time in order to do their part of the deal. But their information flows seem to be misaligned, inconsistent, and divided by tools not built for early-stage startup realities. While in today's digital world everyone seems to celebrate data-driven decision-making, the startup ecosystem has not followed the trend, yet. On one hand, early-stage

founders rely on numerous scattered tools and disconnected platforms to stay on top of their progress. On the other, organizations supporting them rely on manual reporting frameworks and forced, generic milestones, while still frequently lacking the visibility into founder workflows. All sides need transparency to offer their best, but rarely have reliable insight into the startup progress. While the startup ecosystem is supposed to be working as a team towards the same goal, it currently looks more like one player does not want to pass to another, while the third one is running into the offside on purpose. So, while AI and automation are reshaping workflow management, early-stage startup operations remain untouched by these shifts. Is there a lot of room for improvement in the ecosystem? Perhaps, but right now it is time to stop with the existential questions, and look at what the research says.

This study examined how early-stage founders track progress, store knowledge, and communicate with stakeholders. These three factors seem unrelated at first, but they all actually depend on each other - you need to store your startup information in order to track progress, and then communicate it with stakeholders. The study also investigated how incubators, mentors, and investors evaluate progress and maintain visibility into founders' work. While existing academic and industry research tends to focus on success factors, or founder psychology, this research dives into the day-to-day operational infrastructure that underpins early-stage startup development. By gathering perspectives across all four stakeholders (mentors, incubators, investors, and founders), the research gives a complete view into the current state of the (primarily) Croatian startup ecosystem. The central hypothesis states that the current ecosystem relies on informal ways (such as live conversations) or on fragmented, manual tools to track startup progress, which leads to inefficiencies, misalignment, and missed opportunities for all sides. This gap should not be just viewed as operational slack, but as a direct threat to support quality and program outcomes. Some of the questions that the study aims to answer are as follows: How do early-stage startups currently track and manage their progress? How do ecosystem stakeholders attempt to monitor and support this progress? Where do mismatches and blind spots occur? While the focus is on offering the current cross-stakeholder view primarily on the Croatian startup ecosystem, the research can also highlight the opportunities for improvement for startup ecosystems around the world. Rather than focusing on theory alone, the study is based on the lived realities (interviews) of 57 ecosystem participants (mostly Croatian, but international as well). The findings invite the ecosystem to reconsider how progress is measured, supported, and understood.

2 Literature Review

Before we dive into the results, we will cover some basic terms required for proper understanding of the paper, and look at what the existing literature states about the startup ecosystem.

For the start, let's define what a startup even is. According to Schuh et al. (2022), a startup is a newly founded venture that is still in the process of validating its business model, developing its product, and trying to scale it. Startups often operate with limited resources, high uncertainty, and mostly depend on the capabilities of the founding team. When it comes to phases of startup development, there are multiple different terms used to describe the startup life-cycle. Some of them are more popular in the investment world (such as pre-seed, seed, etc.), while some are more used in startup programs, such as ideation, MVP, and so on (Latitud, 2023). For the purpose of this paper, we just need to know what are some widely agreed-upon characteristics of early-stage startups, since those founders were interviewed. In the early-stage, founders have most likely moved beyond a pure idea or a prototype, but have not yet reached consistent growth or organizational maturity. If a basic product and initial revenue exist, these startups tend to secure their first funding from investors. They are iterating their products, developing the required technology, and trying to figure out how to effectively sell their products. In this phase, progress tracking is challenging because traditional performance metrics (such as revenue or profit) are either not applicable yet or unreliable, so key stakeholders focus on appropriate learning-oriented metrics, like validated experiments, customer discovery outcomes, and similar (Shepherd and Gruber, 2021). In addition, fragmented data systems, limited reporting capabilities, and pressure to put the product on market as soon as possible lead to incomplete and inconsistent operational and learning data, which tends to be embedded in founders' tacit knowledge rather than a structured dashboard (RevelX, 2025).

Startup incubators are organizations which offer programs that help founders develop those early-stage startups into viable businesses by connecting them with a large mentor network, and giving them office space, training, education, and/or business support services (Duquesne University, 2020). The difference between an incubator and an accelerator should be that the accelerator also tends to invest a small amount of seed capital in startups in exchange for equity (Tasic et al., 2015). These organizations usually offer one (or more) of the following programs: pre-incubation, incubation and acceleration. Pre-incubation programs usually last a couple of weeks, focusing on founders that are in the idea stage and just starting out. Incubation programs often last for several months and target founders who have already built their first prototypes or basic products and are now figuring out how to put them on the market. Acceleration programs are made for

founders who have already launched their products on the markets and have early traction. These are usually short, but focused on intensive growth. Each program tailors its support for the specific phase a founder is in (European Commission and OECD, 2019). However, sometimes founders themselves do not know which kind of support they need. This points to a structural gap in how support is matched to startup needs (Clayton, 2024). Another key ecosystem player is venture capital funds (investors). A venture capital fund is a way to pool investments. It is generally set up as a limited partnership, with limited partners putting in most of the money and general partners managing the money by choosing, overseeing, and pulling out of investments in private companies with fast growth. Limited partners, who are usually institutional or wealthy investors, tend to be passive and rely on fund managers to identify opportunities and make investment decisions on their behalf (Topinka and Kerr, 2001). After the VC fund makes an investment, founders have an ongoing obligation to report their progress. This usually takes the form of periodic reports.

The final term that will be explained briefly is artificial intelligence. According to the OECD (2024), AI can be defined as "a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments." Nowadays, many digital tools are powered by AI, which means that those applications are using such systems to perform certain tasks (like pattern recognition, personalization, etc.).

Now, the existing literature explains startup progress tracking as the selection and use of quantitative and qualitative indicators, such as number of users, revenue, etc. (Kamulegeya et al., 2018). These metrics should make early-stage performance observable both to founders and other stakeholders, but empirical studies show wide variation and point out the lack of "best practice" when it comes to what startup founders should actually measure (Ben Saad, 2024). Due to the lack of standardized requirements, research shows that founders tend to assemble their own workflows to manage tasks, data, and decisions (ESMT Berlin, 2017). This then leads to heterogeneous reporting packages that stakeholders get from the founders, as some founders collect rich financial and operational details, while others send software-generated tables. On the other hand, investors encourage founders to use milestones or KPI dashboards to ensure relevant updates, but, unfortunately, those dashboards are usually constrained by inconsistent definitions, infrequent reporting, and the difficulty of capturing learning and experimentation with simple numerical indicators. This makes it hard for the stakeholders to assess the startup's actual operational state.

3 Methodology

This research was designed as a qualitative, interview-based exploratory study that aimed to uncover how early-stage startup activity is currently tracked, supported, and evaluated across all startup ecosystem stakeholders. To keep the consistency, but also enable participants to elaborate on their experiences, a total of 57 semi-structured interviews were conducted. Startup ecosystem stakeholders include startup founders, startup incubator representatives, startup mentors and VC investors. The cross-stakeholder comparison was purposely used to obtain the full picture of the ecosystem. Geographic focus was primarily on Croatia (49 participants), but some of the participants were from Serbia (5), Montenegro (1), Austria (1), and the Czech Republic (1), which adds a Central/Eastern European context to the study. Interviews were conducted in English and/or Croatian. Participants were selected based on their active involvement in the early-stage ecosystem. Their participation was voluntary, they were all informed of the research purpose, and they provided consent for recording. However, data was anonymized and aggregated to protect confidentiality due to the small ecosystem size. Interviews were conducted between April and June 2025. The duration of interviews ranged between 20 minutes and 1 hour, depending on the participant's depth of answers. The majority of interviews were conducted online. Questions were tailored for each stakeholder group, but all interviewees were asked the same questions within the same group (startup incubator representatives 11 questions, mentors 9, VC investors 9, early-stage founders 12, idea-stage founders 16). Questions followed pre-defined thematic categories (exact question set for each group can be found in the Appendix). That is why the answers were then also analyzed according to the themes and cross-comparison. Each interview was recorded and then transcribed, with a clean summary made for each participant. Interviews that were conducted in Croatian were translated into English before the analysis. All transcriptions were then manually reviewed multiple times by the authors, while extracting and grouping insights into positive and negative patterns. The process was conducted iteratively to match patterns thematically across all stakeholder groups. At the end, AI assistance was employed to cross-check the identified patterns and flag any overlooked connections, with all final interpretations reviewed and confirmed by the full research team. To further strengthen analytical reliability, findings were additionally reviewed by two experienced academic researchers before settling on the final version.

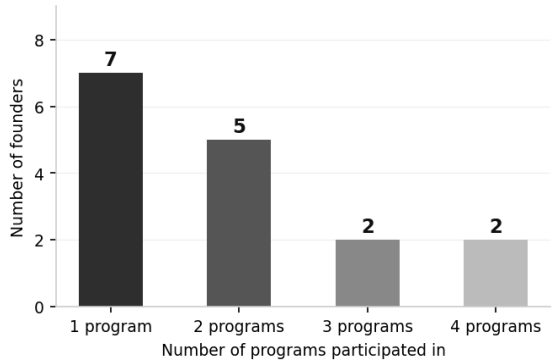
In total, 11 incubator representatives were interviewed. Themes covered included: challenges in working with startups, communication and engagement, tracking startup progress and performance, and efficiency and operational improvements. Incubators were included in the study because they stand as the primary pillar for

the founder support in the ecosystem. Along with incubator representatives, 19 startup mentors participated. Themes focused on: introduction and mentorship process, challenges in mentorship, communication and engagement, time, energy, and tools, and measuring impact. Mentors were included because they serve as an intersection point between incubators and founders. Also, 4 investors were included in the study. Themes explored with them were: investment process and criteria, tracking and monitoring startup progress, challenges in evaluating startups, communication and transparency, and ecosystem gaps and frictions. They were included in the study because their impact is clear – they support founders with funding, but in order to do that, they need data. Finally, there were 23 founders (or founding teams) included. They were separated into 2 groups. The first group consisted of 16 early-stage founders. Themes covered: expectations and motivation for joining (in the context of incubator programs), challenges and frustrations, support, communication and transparency, and startup development. In this group, the founders offered a reflective perspective as they were describing their experiences with programs, tracking, and data storage. The second group focused on 7 idea-stage founders. Themes included: pain points, goal-setting and progress tracking, workflow organization, validation and readiness, learning and resource-seeking, collaboration, and gamification. These founders focused on sharing their current operational reality.

4 Results

4.1 Startup Founders (Group 1)

Questions for this group were directed primarily towards founders' experiences in startup incubator programs and how founders themselves track their own progress and organize their internal data. 9 out of 16 founders (around 56%) even participated in multiple incubators (2 to 4 programs each). Complete distribution of number of programs attended per founder is shown on Graph 1.



Graph 1. Number of Programs Attended per Founder (Group 1 founders, n=16)

4.1.1 Positive Patterns

One of the main reasons why founders join incubators is the networking opportunities, including access to mentors and other experienced founders. A few participants pointed out excellent individual mentors, describing them as key contributors and reasons why the program paid off. Several also highlighted valuable education in fields such as business fundamentals, pitching, deep tech guidance, or industry insights. This was especially useful to founders coming from a technical background. Along with networking and education, some founders stated that incubators contributed to their internal structure, accountability, and a sense of community. They say that multiple programs required frequent communication, which mostly happened either via Slack, WhatsApp or email. Transparency around tasks and expectations was highly important and appreciated among founders, as well as the availability of mentors for ad-hoc sessions. A few founders appreciated honest feedback, even when it was critical. In some of the programs, founders were required to follow formal reporting structures like bi-weekly reports, mandatory tasks, and gamified point systems (utilized through platforms such as Slack or Trello). Founders did acknowledge that these structured reporting approaches sometimes contributed to maintaining accountability. They also said that these incubator tasks occasionally led to the creation of useful business documents (for example, market research summaries). Some founders benefited from clear tasks and program expectations, while community aspects (such as peer support, shared experiences, and group sessions) repeatedly emerged as a positive motivator.

When it comes to internal progress tracking, there is no standardized way that founders approach it, but they all rely on their self-made systems. The main tools that the founders pointed out were Notion, Trello, ClickUp, Jira, and Asana. In more structured teams, these tools are being used for weekly planning or sprint cycles, while some of the founders still rather use physical notebooks or basic task lists. Teams that do internal KPI-setting, often tie them to those weekly or sprint cycles. For storing insights gathered from experiments and interviews, many founders use a mix of Google Drive/Docs, email and spreadsheets. What was observed among the majority of founders is strong self-initiative, which includes running pilot tests, conducting interviews, and/or building prototypes independently. Several teams interviewed have already achieved meaningful early results, such as securing pilot companies, completing MVPs, winning pitch competitions, and opening laboratories (applicable to deep tech teams). Founders demonstrated rapid learning cycles, during which they pivoted ideas based on customer interviews or market research (at least 5 founders explicitly mentioned pivoting at some point).

4.1.2 Negative Patterns

What can be seen, is that the founders' experiences vary based on the program they were a part of. For some of the programs, founders reported a lack of structure and inconsistency, which were tied to disorganized schedules, last-minute changes, inconsistent mentor availability, and misalignment between content and startup stage. The lack of personalization was obvious, as several founders described particular incubators as too theoretical or general (lack of content relevant to the founders' industry or stage). For example, early-stage founders were forced to work on financial projections or execute repetitive tasks with minimal added value. Founders also reported unclear expectations, over-engineered platforms nobody finds useful, forms and tasks that were never checked, and mentors who were difficult to reach or inconsistent. Several founders also experienced mentor mismatch because of random assignments, or lack of mentor's domain expertise. Some incubators even gave tasks that had no follow-up, which created a feeling of administrative burden. Repeated information requests and too many mandatory sessions also contributed to administrative overload. Many founders pointed out a lack of active tracking by incubators. Most founders communicated with incubators on a weekly or bi-weekly basis, but some reported month-long gaps. They also stated that reporting felt repetitive or superficial, progress wasn't consistently reviewed or discussed, and, while certain platforms existed, they weren't used. Also, if founders were part of multiple programs, they were required to use different disconnected tools, causing organizational fragmentation.

Internally, many founders struggle with fragmented documentation, which is scattered across Google Drive (most common), local files, Notion (second most common), Discord, GitHub, and other platforms. Nearly all founders referenced using at least 3 to 5 tools simultaneously just for internal organization. There is limited centralization and reliance on multiple free-tier tools due to budget restrictions. Also, incubators using different tools led to a frequent need to "rewrite the same information" for different programs. Founders pointed out customer interviews as one of the hardest early tasks (initiating conversations, sourcing interviewees, interpreting feedback). Several teams also mentioned team composition issues (a lot of early-stage teams lack certain roles, have overlapping team skills or unclear responsibility division). Also, a major barrier for founders is balancing external commitments, such as university or job, with startup work. The majority of founders also faced self-doubt, emotional roller-coasters, difficulty setting realistic milestones, and constant shift of priorities. Some of the mistakes they highlighted include: higher focus on development instead of customers, delayed sales outreach, lack of documented processes and struggle with legal and administrative burdens. There is a recurring pattern of founders feeling overwhelmed by the size of their to-do lists and the number of possible directions they can pursue.

4.2 Startup Founders (Group 2)

Questions for this group were targeted primarily at founders' current operational processes and struggles. These founders were actively developing their ideas for a shorter period of time than founders from the first group.

4.2.1 Positive Patterns

During the answer analysis, it was noticed that all the founders used some sort of validation methods for their ideas. Most of them used informal validation methods, such as focus groups, conversations with friends and peers, social sentiment (e.g. community posts, public reactions), and mentor or professor guidance. While such methods are not the most reliable methods for gathering objective insights, they are better than avoiding any sort of validation prior to the development phase. Only one team used structured validation methods, such as recorded interview logs and step-by-step problem-definition loops (a subset of design thinking methods). However, several teams received early validation, including positive focus group responses and clear industry feedback (according to the founders' opinions). When it comes to internal organization, all founders relied on 3+ tools simultaneously for progress tracking, communication, and documentation. Most of the tools repeated like in the first group: Google Drive, Slack, Discord, WhatsApp, Trello, ClickUp, Jira, Notion, Figma, etc. For team communication, WhatsApp was the most used channel. To ensure accountability, some teams focused on maintaining weekly or almost-weekly meeting routines. These meetings also contributed to strong team commitment, along with approaching decision-making with equal voting. Teams felt especially motivated when powered by small wins, positive feedback, external validation, and peer recognition, all of which are tied to external motivation. For learning, founders showed a preference for practical, hands-on learning methods: trial-and-error approach, YouTube, Google, ChatGPT, podcasts, startup founder stories, experiential learning over structured tutorials, visual whiteboarding tools, and overall, "learning by doing" rather than planning. Also, founders enjoy gamification elements when they are applied to progress tracking, breaking large goals into smaller tasks, offering recognition or status, and allowing friendly competition or benchmarking. As examples, some founders pointed out Duolingo and Waze-inspired gamification. Also, early pivots seemed common among the teams.

4.2.2 Negative Patterns

On the other hand, most teams highlighted severe issues with balancing startup work and academic or job

responsibilities, which often lead to inconsistent progress, missed deadlines, irregular meetings, and delayed development. Also, although they agreed on the use of tools for organization, they were not always used because the team would not keep them regularly updated. Nearly all founders reported scattered documents across multiple platforms, causing them not to have a single source of truth, ideas being stored across different places, and workflows becoming messy or confusing. Several teams expressed desire for a more centralized place with better structure, described as a "digital notebook." Furthermore, one common struggle observed was lack of clear priorities. These teams had difficulties deciding what to do next, choosing between features, balancing programming and business tasks, and overall, felt overwhelmed by too many possible directions. Founders also addressed lack of certain skills, such as weak sales skills, difficulty with financial planning, poor public speaking, lack of technical clarity, or limited knowledge of how to execute a go-to-market strategy or determine pricing. This led to the founders feeling "stuck", and unsure how to take next steps. Additionally, teams struggled with internal coordination because of uneven availability across members, asynchronous work misalignment, difficulty assigning roles, and conflicts in decision-making. Overthinking, cognitive overload, confusion about workflows, and low motivation during complex tasks were also stated by the founders.

4.3 Startup Incubator Representatives

These questions focused on examining common challenges in working with startups and how effectively incubators track their progress.

4.3.1 Positive Patterns

Most incubators use multiphase programs (usually divided into 2 to 3 different cycles), and are trying to set milestone tracking (either through monthly evaluations or demo-day performances). Incubator coordinators directly communicate with founders and mentors by using Slack (used by 4+ incubators), WhatsApp, Discord, email, calls or face-to-face communication. Incubators focus on maintaining high-touch communication, which offers quick access to staff. Teams often show a strong personal connection with incubator staff, described as "family-like closeness." Also, incubators maintain large mentor networks (sometimes with even 50+ mentors that come from different backgrounds). Mentorship is based on regular monthly meetings and specialized expertise. Several incubators use structured reporting: bi-weekly reports, quarterly or biannual feedback forms, or tracking KPIs program-wide. On top of that, some incubators add on mentor validation to verify a startup's progress, and many incubators state that they observe visible progress of teams throughout and post their programs. Tools used for internal operations include: Monday.com, Trello, Slack, OneDrive, Excel, Google Drive and Miro.

4.3.2 Negative Patterns

Some incubators reported a noticeable drop in the number of applications, tied to demographic declines (specific for certain cities), lower-quality idea submissions, and a smaller number of young people entering programs. When it comes to commonly faced issues with founders, incubators point out: lack of proactivity (founders wait and react to instructions rather than initiating), low commitment (due to full-time jobs or studies), and mindset gaps (difficulty shifting from technical to business thinking, hesitation around market validation, and inconsistent follow-through on tasks). For communication specific issues, incubators face uneven participation within communication channels, the need for constant reminders, and variation in communication styles (different mediums). On top of that, global programs can face cultural barriers, language differences, and uneven engagement across regions (remote models). Also, progress tracking limitations can be observed. Many incubators lack a centralized progress-tracking platform because they rely on manual tracking (explicitly mentioned by 3 incubators), "databases" that are rarely used, and self-reported progress by founders (for which they note can be overly optimistic). There is no use of advanced analytics. Also, one incubator acknowledges a "drop-off" after program completion as a major issue. Furthermore, incubators face several constraints: they depend on small staff size, limited funding for tools and platforms (and overall), difficulty scouting mentors in niche verticals such as biotech or medtech, and have ongoing challenges in coordinating large networks of mentors. On the ecosystem level, they noted small domestic markets, lack of cooperation between different ecosystem players, fragmented support landscape, regulatory complexity in international expansion, and general lack of experienced founders. Moreover, incubators report a gradual drop-off in attendance at the program's activities, especially after initial enthusiasm fades, the official part of the program ends, or when founders struggle to balance work and program demands. Finally, several incubators expressed a wish for better digitalization, but currently lack budget, integrated platforms and automated analysis to do so. One incubator noted actively creating an AI-powered progress tracking method because manual methods cannot be scaled.

4.4 Startup Mentors

Mentors were asked about common challenges in mentorship, how they navigate mentoring in multiple incubators, and how they measure the impact of their mentorship.

4.4.1 Positive Patterns

Nearly all mentors stated that mentorship should be highly personalized and adaptive. When talking with founders, they try to adapt to the founder's personality, startup stage, and their current needs. To evaluate the state of the startups, mentors rely on deep questioning to reveal gaps, pitch-based dissection, business model deep dives, and technical feasibility reviews. In order to provide good advice, mentors leverage their industry experience, international market exposure, fundraising and scaling insights, corporate and startup background, real examples, and case-based guidance. Many mentors also state that their purpose is to anchor the founder in reality, which can include simplification of ideas, identification of risks, and pointing out blind spots. Some of the common areas that mentors frequently help founders with include: pitch refinement, identification of the problem that the startup is trying to solve, improvement of narrative clarity, prioritization of features, understanding of monetization and preparation for market entry. The most used communication channel between mentors and mentees is WhatsApp (due to speed), but email, phone calls, and face-to-face communication are common as well. Mentors tend to detect a founder's potential by looking at coachability, responsiveness, energy, willingness to learn and execution speed. Overall, engagement consistency is repeated as a key indicator of future success. Mentors evaluate the success of their mentorship based on several factors: level of founder gratitude, improvement in founder's pitch or clarity, implementation of advice, achievement of funding, startup's survival beyond first year, winning pitch competitions, and long-term follow-up stories from founders.

4.4.2 Negative Patterns

The most common challenge that mentors face when working with startups is founder commitment issues. They say that a lot of founders are not working on their startups full-time. On top of that, they are slow in communication, disappearing for weeks, face difficulty balancing jobs, studies, and startup, rely on "vibes" rather than structured planning, and showcase declining engagement after the initial enthusiasm. Multiple mentors also reported that founders misrepresent their progress. According to mentors, startups frequently exaggerate their progress, avoid mentioning setbacks, and blend aspirational progress with actual progress. Mentors often need to "read between the lines" to understand the whole picture. Some of the common founder weaknesses observed by mentors include: poor competition analysis, weak problem-solution definition, incorrect valuations, poor marketing setup, and weak financial planning. Overall, lack of business understanding. From a psychology perspective, mentors often encounter challenges with founders who show: ego issues that block learning, emotional volatility, overconfidence in early ideas, indecision, shifting priorities, and resistance to validation of assumptions. Mentors also come across multiple communication frustrations: founders struggling to simplify complex ideas, withholding or distorting information, or having misaligned expectations between tech and business roles. Furthermore, mentors reported several ecosystem-level gaps: fragmentation, limited accumulated experience, poor failure culture, and inefficient mentor-startup matching. On a personal level, mentors consistently cite time as a limitation. They experience scheduling difficulties (especially if they are mentors in several incubators). They also prefer short and efficient follow-up sessions (around 30 minutes), and state that long-term and structured mentoring is rare simply due to the amount of time they can dedicate. When it comes to progress tracking, almost none of the mentors use tools to track founder progress. Instead, they rely on intuition, sporadic check-ins, WhatsApp messages, and founder-submitted documents. Finally, mentors identified commercialization, finding market fit, and scaling as the hardest startup phases.

4.5 Investors

Finally, investors were asked about how they track startup progress and cooperate with incubators, what they look for in founders, and what are some ecosystem gaps they noticed.

4.5.1 Positive Patterns

What can be observed from the investors' answers is that a startup's evaluation follows a predictable structure. First, there is an intro screening call through which they assess scalability, founder quality, and long-term potential. Then there is a deep-dive call if signs are promising. After that, investors tend to consult with external experts (especially if they lack the domain expertise within the fund). Finally, the investment committee meets for formal approval, which will be followed by a term sheet. Investors really look for capable teams primarily – founders that possess qualities such as integrity, vision, passion, complementarity within the team, and who are tied to a real problem worth solving. Besides due diligence, investors also rely on gut-feel assessment. They ask themselves would that be a person they would like to work with for the next 10 years. Post-investment, investors expect clear and consistent tracking. It can be in the form of monthly or quarterly reporting, monthly update calls, or through CRM systems used for milestone tracking. Investors focus on metrics such as revenue growth, user growth, retention, and similar, but also on how capable founders are at acquiring new talent for the team. Investors do think they should provide ongoing mentorship and support, including: strategic introductions, market expansion advice, and deep operational and fundraising guidance. Investors can attend board meetings, act as observers, or maintain advisory relationships. For communication, investors appreciate structured, concise updates that

are transparent. At the ecosystem level, they believe Croatia is a growing ecosystem with potential. They noted strong community and support initiatives, and view certain incubators as valuable channels for early exposure.

4.5.2 Negative Patterns

One thing in particular that investors dislike is dishonesty. They state that founders sometimes exaggerate their progress, stay in "selling mode" post-investment (instead of being realistic), present unrealistic valuations, hide weak metrics, provide fake term sheets or misrepresent advisors. Investors point out that tracking in the pre-seed stage is rather difficult because there are no reliable metrics to follow. They heavily rely on founder communication since tools like CRMs work only after some traction exists. Validation stage progress is especially hard to put into numbers. Also, they noted that there are no standardized reporting formats across different funds or incubators. Moreover, investors face these common challenges with founders in this region: local-first thinking (low ambition to go global), lack of commitment (part-time founders), ignoring investor advice, lack of go-to-market clarity, weak understanding of the competitive landscape, and struggles with shifting from technical to business mindset. Ecosystem-wide, investors report a weak entrepreneurial mindset (they think this should start being fostered in schools), fragmented ecosystem (lack of cross-institution collaboration), lack of geographic decentralization (in Croatia, the startup ecosystem is primarily based in Zagreb), scarcity of late-seed and series A funding in the region, and cultural barriers to risk-taking and failure. When it comes to tools used by investors, some of them build their own internal progress tracking tools because they think that existing platforms are too expensive, poorly suited for early-stage tracking, and/or reliant on generic data sources (in the case of sourcing platforms). However, investors also rely on CRM tools, Google Sheets, and monthly verbal updates.

5 Discussion

From the gathered data, it can be seen that founders operate in highly fragmented and improvised workflows. They rely on scattered tools and documentation that is saved everywhere. This pushes founders toward reactive behaviors, where they are focused on resolving "today's fire", instead of following a structured process. To be able to face those environmental threats and stress, founders need to possess high self-efficacy, according to Baum and Locke (2004). Also, there are misalignments between incubators' and founders' expectations. Founders would like a more tailored approach, while incubators want more proactive engagement from founders. Also, throughout the early-stage, founders deal with overwhelm, uncertainty, and pivoting. This aligns with the research conducted by Aryadita et al. (2023), who mention that founders frequently face the uncertainty during the startup development process because of many setbacks and failures, which then leads to frustration, and sometimes even giving up on the startup altogether. On top of that, founders tend to misjudge their stage of development, which then adds to confusion about priorities, which then leads to difficulty with setting clear milestones. Founders' reliance on external validation through mentors, incubators and overall "small wins" shows how fragile early-stage progress tracking actually is. Incubators often have small internal teams, and rely on manual and inconsistent reporting systems, which creates an additional operational burden. Although incubators do their best to support founders holistically, they do not have the infrastructure to simply do so. They cannot track learning, execution, and consistency in a measurable way. Unfortunately, their visibility depends entirely on the founder's self-reporting. And the engagement drop-off among startups cannot be viewed as random – it reflects unclear expectations and the absence of tools that would keep the momentum going. Moreover, there are no standardized incubator tools, so each incubator uses a combination of its own communication channels, tasks, and formats, which adds administrative burden to founders who are part of multiple programs. From the mentor's point of view, there is no use of systematic tracking tools whatsoever, although mentors themselves state that founders can exaggerate progress. Also, the overall impact of mentorship is sometimes reduced due to inconsistent matching processes, which leads to mentors being underutilized or mismatched. Finally, it is the investors who face the greatest uncertainty in the early-stage because initial metrics are usually unreliable, which puts the biggest emphasis on founder honesty and clarity. Investors have to heavily rely on the founder as a person because the early-stage ecosystem does not provide trustworthy indicators.

When all four perspectives are combined, the dominant theme emerges – there is a lack of shared visibility into startup progress. Each group sees only a partial, inconsistent piece. These groups sometimes experience misaligned expectations, especially because there is no infrastructure to keep them all "in one place". There is no standardized way to track startup progress across the ecosystem, as the whole ecosystem relies on manual and fragmented systems. While mutual trust is an essential part of any relationship, these misalignments can actually lead to a trust deficit – if founders overstate their progress and cannot themselves figure out where or when they need help, other stakeholders essentially cannot help them properly. One thing is for sure, and that is that the whole ecosystem is working hard – with a lot of grit and passion. However, nobody has a real picture of the progress, which reduces success predictability. If every stakeholder is doing their part individually without

having the full picture, can a startup ecosystem then truly be called an ecosystem?

6 Recommendations

This research identified several operational gaps within the startup ecosystem, and the following are potential ways to address these gaps: standardization of progress tracking across the ecosystem (both standardized reporting format and adoption of common terminology for metrics, validation, and traction), implementation of centralized or integrated tracking systems (to reduce fragmentation, support cross-program continuity, and enable real-time progress tracking), strengthening founder education in operational skills (founders skill gap in certain areas was observed by multiple stakeholders, also incubators do not commonly offer education on internal organization, management, milestone-setting, etc.), prioritization of mentor-startup matching quality (by developing systematic pairing processes and reducing mentor underutilization by making their expertise more visible), and overall improvement of ecosystem coordination (including fostering cross-institution collaboration). Moreover, with the emergence of AI, ecosystem stakeholders should look into how AI platforms could lead to improvements in all mentioned areas. There is already existing research which shows that startup progress can be analyzed through data patterns rather than subjective reporting, and that AI models can sometimes detect patterns founders and mentors cannot (Li et al., 2024).

Now, this research could be extended as well. Future research should focus on expanding stakeholder diversity (including stakeholders from different countries rather than primarily focusing on one ecosystem), incorporating quantitative elements (validating patterns not just qualitatively, but developing surveys or tracking behavioral changes over time), and, in general, expanding the number of investors interviewed.

7 Conclusion

Anyone who has tried developing a startup knows that the process itself is messy. It involves a lot of self-doubt, pivoting, and uncertainty. And yes, these will probably forever be associated with startups. If everyone knew for sure how to build a multi-million-worth business, entrepreneurs would not be seen as the odd ones in society. However, what contributes to the messiness of this process is reliance on fragmented tools and large invisibility into startup progress across the ecosystem. Communication and reporting are not based on data, but on intuition, scattered tools, and optimistic self-assessments. Although ecosystem stakeholders work hard, their clarity, consistency, and accountability are limited in the current ecosystem. There is a lack of shared standards, and that can lead to duplicated effort and unclear expectations. The cross-stakeholder analysis shows that early-stage tracking needs integration, transparency, and perhaps even real-time visibility. And there are many opportunities that could lead to improved frameworks, especially as AI-driven progress analysis has come into the story in recent times. Overall, this study highlights that ecosystems should match how early founders actually work, and not how everyone just assumes they work.

References

Aryadita, H., Sukoco, B. M., & Lyver, M. (2023). Founders and the success of start-ups: An integrative review. *Cogent Business & Management*, *10*(3). https://doi.org/10.1080/23311975.2023.2284451

Baum, J. R., & Locke, E. A. (2004). The relationship of entrepreneurial traits, skill, and motivation to subsequent venture growth. *Journal of Applied Psychology*, 89(4), 587–598. https://doi.org/10.1037/0021-9010.89.4.587

Ben Saad, F. (2024). How to measure the success of technology-based start-ups – A comprehensive overview of the perspectives of academics & practitioners. *Junior Management Science*, 9(1), 1306–1340. https://doi.org/10.5282/jums/v9i1pp1306-1340

Clayton, P. (2024). Mentored without incubation: Start-up survival, funding, and the role of entrepreneurial support organization services. *Research Policy*, *53*(4), 104975. https://doi.org/10.1016/j.respol.2024.104975

Duquesne University Small Business Development Center. (2020). *What is a startup incubator and how does it work?* https://www.sbdc.duq.edu/Blog-Item-What-is-a-Startup-Incubator

ESMT Berlin. (2017, November 7). *What information do startups provide to their venture capital investors?* ESMT Knowledge Hub. https://esmt.berlin/knowledge/research-insights/what-information-do-startups-provide-their-venture-capital-investors

European Commission. (2018). *Policy brief on incubators and accelerators that support inclusive entrepreneurship*. Publications Office of the European Union.

European Commission and OECD. (2019). *Policy brief on business incubators and accelerators*. OECD Publishing.

Information Age. (2019, July 12). *Technology is lowering barriers to entry, but increasing barriers to exit*. https://www.information-age.com/golden-

age-of-startups-technology-lowering-barriers-to-entry-increasing-barriers-to-exit

Kamulegeya, G., Mugwanya, R., & Hebig, R. (2018). Measurements in the early stage software start-ups: A multiple case study in a nascent ecosystem. *Foundations of Computing and Decision Sciences*, 43(4), 283–308. https://doi.org/10.1515/fcds-2018-0014

Latitud. (2023). *7 stages of a startup: From ideation/MVP to growth/maturity*. https://www.latitud.com/blog/stages-of-a-startup

Li, Y., Zadehnoori, I., Johwar, A., Wise, S., Laplume, A., & Zihayat, M. (2024). Learning from yesterday: Predicting early-stage startup success for accelerators through content and cohort dynamics. *Journal of Business Venturing Insights*, 22, e00490. https://doi.org/10.1016/j.jbvi.2024.e00490

OECD. (2024). *Explanatory memorandum on the updated OECD definition of an AI system* (OECD Artificial Intelligence Papers, No. 8). OECD Publishing. https://doi.org/10.1787/623da898-en

RevelX. (2025, January 20). *Measuring innovation: A guide to building investment confidence*. https://www.revelx.co/blog/measuring-innovation-a-guide-to-building-investment-confidence/

Schuh, G., Studerus, B., & Hämmerle, C. (2022). Development of a life cycle model for deep tech startups. *Journal of Production Systems and Logistics*. https://d-nb.info/1253580243/34

Shepherd, D. A., & Gruber, M. (2021). The lean startup framework: Closing the academic–practitioner divide. *Entrepreneurship Theory and Practice*, *45*(5), 967–998. https://doi.org/10.1177/1042258719899415

Tasic, I., Montoro-Sánchez, Á., & Cano, M.-D. (2015). Startup accelerators: An overview of the current state of the acceleration phenomenon. In *Proceedings of the XVIII AECA Congress*. Asociación Española de Contabilidad y Administración de Empresas.

Topinka, J. E., & Kerr, C. (2001). Liabilities of general partners of PE/venture capital funds: Old legal theories, new business reality. VCExperts. https://www.vcexperts.com/buzz_articles/297

Tykova, T., Buhrer, S., Radicic, D., & Uyarra, E. (2017). Startup support programs: An overview of the literature and their performance. European Commission.

Appendix

Questions for startup incubator representatives:

Challenges in Working with Startups

1. Can you walk us through your typical process of working with startups within your programs? Can you walk us through the whole program? What does it contain?
2. What are the most common challenges you face with startups during your programs? How do you typically handle these challenges? What are the common challenges you face as an incubator?
3. Despite mentorship, what are the main reasons startups struggle? How do you identify which startups need additional support?
4. In your experience, which areas of startup development founders find most challenging, and why?

Communication and Engagement

5. How do you currently facilitate communication between startups and your team?
6. What frustrations do you experience when communicating with startup founders and mentors?

Tracking Startup Progress and Performance

7. How do you track the progress of startups?
8. What tools or resources (if any) do you use to track startup progress effectively?
9. What patterns, if any, have you noticed in how startups communicate their progress? How do you factor that into your work with them?
10. How does the way startups report their progress affect your work with them?

Efficiency and Operational Improvements

11. How do you currently track and manage data related to your startups?

Questions for startup mentors:

Introduction and Mentorship Process

1. Can you walk me through your typical process when working with a startup, from the first meeting to ongoing mentorship?
2. Do you mentor startups across multiple incubators or programs? If so, how do you balance and manage those responsibilities?

Challenges in Mentorship

3. What are the biggest challenges you face when mentoring startups?

Communication and Engagement

4. What are your preferred channels of communication with startups, and why?
5. What are the biggest communication challenges you face when working with startup founders?
6. In your experience, how well do startup progress updates reflect where a startup actually stands? What influences your view?

Time, Energy and Tools

7. What aspects of working with startups take up the most time or energy?

8. What tools or resources (if any) do you use to track startup progress and mentor effectively?

Measuring Impact

9. How do you measure the success of your mentorship? What outcomes matter most to you?

Questions for VC investors:

Investment Process and Criteria

1. Can you walk me through your typical process when evaluating early-stage startups for investment?
2. When evaluating a startup for investment or monitoring an existing portfolio company, what indicators do you look for to determine whether they're on track and moving in the right direction?

Tracking and Monitoring Startup Progress

3. How do you currently track the progress of your portfolio startups?
4. What tools (if any) do you currently use to track milestones or outcomes across your investments?

Challenges in Evaluating Startups

5. What are the most common challenges you face when assessing early-stage startups?
6. How would you assess the accuracy of progress reporting among startups you have encountered? What factors, if any, influence how easily you can verify a startup's reported progress?

Communication and Transparency

7. What's the most frustrating part of communication between investors and startups?
8. How would you assess the transparency of reporting from incubators you've worked with?

Ecosystem Gaps and Frictions

9. How would you describe the quality of interaction and information flow between incubators, mentors, and startups?

Questions for early-stage founders (1st group):

Expectations and Motivation for Joining

1. Why did you join the program, and what were your expectations? Have those expectations been met so far? If not, why?

Challenges and Frustrations

2. What has been the biggest challenge or frustration you've faced during the incubator's program?
3. Have you ever struggled to meet an incubator's milestone or deadline? What happened, and how was it handled?

Support, Communication and Transparency

4. How often do you communicate with mentors or incubator staff? Why/Why not more?
5. What's been the most frustrating part of communicating with the incubator team?
6. How would you assess the transparency of the incubator regarding expectations, milestones, and program success criteria?
7. How do you currently track and report your startup's progress to the incubator?

Startup Development

8. What are the biggest challenges you have faced in the early-stage of your startup development?
9. Can you walk us through how the early-stage looked like for you? Tell us your journey – specifically, how did you get started'
10. How do you set milestones as a team and make sure to meet them? How do you stay motivated?
11. How do you track your own progress?
12. How do you organize your data (for example, customer interviews, prototype testing data, etc.)?

Questions for idea-stage founders (2nd group):

Pain Points

1. What pats of the startup development process do you find most confusing or difficult to navigate?

Goal-Setting and Progress Tracking

2. How do you currently track your progress (e.g. tasks, milestones, KPIs)? What tools do you use?
3. How often do you reflect on your progress – and what usually triggers that reflection?

Workflow Organization

4. What does a typical week look like for your team? What takes up most of your time?
5. Where do your ideas and documents currently live (Notion, notes app, whiteboard, brain fog)? Is there one place where you keep everything, or is it scattered across tools and spaces?
6. What's the messiest part of your current workflow?

Validation and Readiness

7. Walk me through how you validate your idea today – step by step.
8. How do you document or reflect on learnings from customer interviews, testing, or market feedback?

Learning and Resource-Seeking

9. What's been your main way of learning how to build a startup so far?
10. When you reach a new stage in your startup, how do you decide what to focus on or where to turn for guidance?
11. What entrepreneurial skills are you currently trying to develop?

12. How do you prefer to learn – through videos, articles, exercise, feedback, or something else?

Collaboration

13. How do you share progress, problems, and decisions with your team members?

Gamification

14. Have you used any gamified tools before? What did you like or dislike about them?
15. How do you feel about receiving recognition for your achievements?
16. How do you feel about friendly competition or benchmarking against other startups?

The Critical Steps for Using AI: A Case Study from a Manufacturing Company

Stjepan Mihalić, Josip Mihalić
Jedinstvo Krapina d.o.o.
Mihaljekov Jarek 33
49000 Krapina, Croatia
{sm.jr,jmihalic}@jedinstvo.com

Tomislav Baković
University of Zagreb
Faculty of Economics & Business
Trg J. F. Kennedyja 6
10000 Zagreb, Croatia
tbakovic@efzg.hr

Abstract. *AI adoption among Central and Eastern European manufacturers remains low. This paper reports the first phase of a case study in a Croatian mid-cap manufacturer of modular buildings. Drawing on recent literature, quality-management documentation and fifteen semi-structured interviews covering ten organisational roles, it maps where AI can support company processes and proposes a phased implementation framework that begins with governance and low-risk office applications. A TOE-based survey of 63 of 82 employees shows high intention to use AI but weak organisational readiness; trust and perceived market pressure are the strongest predictors of intention (R^2 = .70).*

Keywords. artificial intelligence, business value, technology adoption, TOE framework, modular construction, manufacturing mid-cap

1 Introduction

Over the past few years AI has crossed from pilot experiments into everyday use. Its effect on manufacturing performance is now measurable rather than speculative, yet adoption across Central and Eastern Europe has not kept pace. Drawing on the European Manufacturing Survey, Kovic et al. (2024) report that only 18.4% of manufacturers in Croatia, Slovenia and Slovakia run AI software in even one production area. What sets the adopters apart, they find, is readiness for Industry 4.0 rather than company size or position in the supply chain. More recent figures show movement but a persistent gap: by 2025 about 15% of enterprises in Croatia used at least one AI technology, up from about 12% a year earlier yet still below the EU-27 average of roughly 20% (Eurostat, 2025), and a 2025 survey of more than 3,200 SMEs across eleven CEE countries found that, while some three-quarters report using AI, only about a quarter do so at scale (AI Chamber, 2025). Construction tells a similar story. Reviews of modular and prefabricated building argue that AI and robotics are needed to unlock the full advantage of industrialised production (Pan et al., 2022), and the arrival of generative AI and large language models (LLMs) has added a class of applications that companies can take up without heavy upfront investment (Ghimire et al., 2024; Ma et al., 2025).

The case examined here is one such company. It is a Croatian mid-cap manufacturer in the north-west of the country, with several decades of tradition in steel structures and modular buildings, and it now produces modular residential, educational, hotel and office units (E-houses among them), mostly for export to Germany and other west European countries. Its integrated management system is certified to ISO 9001, ISO 14001 and ISO 45001, and since 2025 it has been introducing lean management: 5S, the Last Planner System, Kaizen and visual management. Management is now weighing how to bring AI into its processes in a systematic way.

Three questions guide the study. The first, RQ1, asks which business processes of a modular construction manufacturer hold the most potential for AI support. The second, RQ2, asks how ready employees are for that step, read through the technology-organisation-environment (TOE) lens. The third, RQ3, asks what implementation path suits a manufacturing mid-cap that is adopting AI in the middle of a lean transformation: in what order, starting when, governed how, and with how much outside help.

The rest of the paper is organised as follows. Section 2 sets out what AI promises companies and why they adopt it. Section 3 turns to the more specific literature on AI in modular construction and on adoption in manufacturing SMEs. Section 4 describes the company and where it stands today, and Section 5 sets out the research design. Section 6 puts forward the preliminary framework, Section 7 presents and discusses the survey results, and Section 8 concludes.

2 The promise of AI for companies

The case for adopting AI rests less on the technology itself than on the value it is expected to create. Reviewing the literature on AI and business value, Enholm et al. (2022) sort that value into two kinds. The first is automational: AI takes over work that people would otherwise do. The second is informational and transformational: it sharpens decisions, products and ways of working. The same review adds a caution that runs through the rest of this paper, namely that the value never appears on its own. It is realised only when AI is matched with the right data, skills and organisational routines.

The first and most tangible promise is productivity. McKinsey's analysis of generative AI puts the prize at USD 2.6-4.4 trillion a year across the activities it examined, and about three-quarters of that value falls in a handful of functions: marketing and sales, customer operations, software engineering, and research and development (Chui et al., 2023). Field evidence is beginning to catch up with the projections. Studying 5,172 customer-support agents, Brynjolfsson et al. (2025) found that access to a generative-AI assistant raised the number of issues resolved per hour by about 15%, and the largest gains went to the least experienced staff, who in effect absorbed the know-how built into the tool. Qualitative work points the same way: in a study of Spanish firms, Mas-Machuca et al. (2025) describe how generative AI moves effort away from routine drafting and searching toward judgement and client-facing work, as long as employees are given time and clear guidance. For a manufacturer the equivalent gains sit not on the line but in the office and engineering work around it, from drafting and translation to searching documentation and preparing quotations. Recent evidence from smaller manufacturers suggests these office-side efficiencies are where AI pays off first (Turcic et al., 2025).

Quality and consistency are a second reason. An AI system applies the same logic every time and can screen far more cases than a person, so it cuts the errors and rework that build up in manual, repetitive work. In construction this is clearest in computer-vision inspection, where the technology already rivals manual checking on well-defined defects (Yevu et al., 2025). A third reason is better decisions, and better use of what a firm already knows. From 152 cognitive-technology projects, Davenport and Ronanki (2018) separate three things AI does for companies: it automates back-office processes, it draws insight out of data, and it engages with customers and employees. The firms that gain most, they observe, treat AI as a way to augment people rather than replace them, and they get there through incremental projects rather than 'moon shots'. A related benefit is the ability to hold on to institutional knowledge. A searchable, AI-supported knowledge base keeps the experience of departing staff available to those who remain, which for many firms is a quiet but real source of value.

None of this follows automatically, which is the fourth and most important point. Mikalef and Gupta (2021) show empirically that performance comes not from buying AI but from an AI capability, a bundle of data, technical skill and organisational resources such as coordination and a supportive culture, which in turn raises organisational creativity and firm performance. How the technology lands inside a firm matters as much as the technology itself. Tracing the arrival of generative AI through organisational networks, Olmedo Mejia and Cvetkoska (2025) find that value appears only as roles, routines and responsibilities are renegotiated around the tool rather than bolted onto it. This fits the manufacturing evidence that readiness for Industry 4.0, not firm size, separates adopters from the rest (Kovic et al., 2024), and the broader finding that digital and innovation capabilities are the main drivers of AI adoption among European SMEs (Arroyabe et al., 2024).

These mechanisms take a particular shape in manufacturing and construction. Within manufacturing, real AI use concentrates in quality control, production-process management and maintenance. There, vision systems and predictive models cut defects and downtime, while the office and engineering functions gain most from language and search tools (Kovic et al., 2024). In construction the promise is twofold. On the production side it is machine vision and analytics; on the information side it is turning scattered technical documentation into something a system can read and answer questions about. Recent work shows large language models moving beyond classic business intelligence to process and summarise unstructured documents directly (Krcmar et al., 2025), precisely the capability a documentation-heavy modular manufacturer needs. The promise of AI, then, is not generic. It is strongest where a firm's particular work is bound up with information.

Against these promises stands a counterweight the same literature keeps raising. The gains do not arrive on their own. Beyond the technical and data barriers there is a human one: working with AI can itself become a source of stress and quiet resistance, which a recent bibliometric review flags as a fast-growing theme under the label of technostress (Zron, 2025). The practical implication is that the case for AI has to be made to a company's employees and not only to its management, and that adoption is as much a matter of preparation, of time, training and trust, as of technology. It is partly for this reason that the present study sets its analysis of processes alongside a direct measurement of how ready the workforce actually is.

A final reason is competitive, and it is as much about risk as opportunity. As AI spreads through an industry, early adopters accumulate data, skills and routines that later entrants struggle to copy. Those who wait, meanwhile, meet customers and partners who

increasingly take digital ways of working for granted. For an export-oriented manufacturer that pressure is immediate, and it turns AI from a discretionary investment into a condition for staying competitive. Taken together, these five reasons, productivity, quality, better decisions, capability-building and competitiveness, explain why companies are turning to AI. They also frame the question this paper sets out to answer: not whether the promise is real, but how a manufacturing mid-cap can realise it given its own processes and its own state of readiness.

3 Theoretical background

3.1 AI in modular and prefabricated construction

Work on AI in modular construction has expanded quickly. In a systematic review of 97 articles, Pan et al. (2022) organise the field along a concept-methodology-value scheme that runs from design through manufacturing and transport to on-site assembly, and they argue for introducing AI in stages along the value chain rather than as a single leap. Liu et al. (2024) survey its use in production, operations and logistics, and single out predictive scheduling and the optimisation of assembly sequences as the most promising avenues. Wang et al. (2025) sort AI-based robots in industrialised building manufacturing into four capability groups, namely cognition, communication, control, and collaboration and coordination, and map them onto five factory stages that include cutting, welding and forming. Their advice is to de-risk such investment with digital twins and to lean on transfer learning where volumes are low and variety is high. Quality control is where the evidence is firmest. Convolutional neural networks dominate, with artificial neural networks next, and productivity and safety are the recurring problem areas (Yevu et al., 2025), which is why machine vision is the most mature and least risky place to begin. A PRISMA review of generative design and optimisation in prefabrication maps four research trends against twelve obstacles, of which algorithmic complexity and data interoperability are the most serious (Rangasamy & Yang, 2025). In a more applied vein, Heredia Alvaro and Gonzalez Barreda (2025) show how a retrieval-augmented generation (RAG) system built on LLMs turns manufacturing documentation into working support for quality control.

3.2 AI adoption in manufacturing SMEs

A second strand treats adoption as an organisational problem. Within manufacturing, real use clusters in quality control, production-process management and maintenance, and every step up in Industry 4.0 readiness raises the likelihood of using AI by roughly 21% (Kovic et al., 2024). Sanchez et al. (2025) combine the TOE and diffusion-of-innovations frameworks, list ten obstacles facing SMEs, from access to data and a shortage of skills to cultural resistance, and lay out a six-phase roadmap. On a far larger sample of 12,108 European SMEs, digital capability turns out to be the main driver of adoption, amplified by innovation capability (Arroyabe et al., 2024). German data tell a similar story of early-stage adoption and a gap between how relevant firms believe AI to be and how much they actually use it (Ulrich et al., 2021). In construction specifically, the strongest pull for LLMs comes from the company level, from partnerships, in-house research capacity and staff training, while hallucinations, data security and the complexity of the domain hold use back (Ghimire et al., 2024; Ma et al., 2025). And, as upper-echelons research keeps showing, it is the engagement of senior managers that finally decides whether AI integration succeeds (Bevilacqua et al., 2025).

3.3 Lean management as an organisational foundation for AI

Lean management and digital technology are increasingly read as two sides of the same coin. Industry 4.0 tools reinforce lean practices and the reverse holds too (Buntak et al., 2019), and lean maturity is one of the building blocks in the Croatian model of the innovative smart enterprise (Veza et al., 2016). A lean transformation steadies processes, brings waste into view and grows a culture of continuous improvement (Womack & Jones, 1996), the very conditions the adoption literature ties to lower organisational barriers to AI. The influence runs the other way as well: AI can sharpen lean practice, whether by feeding predictions into the Last Planner System or by flagging the anomalies that trigger andon signals.

4 Case company and current state

In EU terms the case company is a mid-cap. It has outgrown the SME thresholds of 250 employees and EUR 50 million in turnover, yet it stays below the large-enterprise scale that, in EU usage, reaches up to about 3,000 employees (European Commission, 2025). Larger than an SME, it nonetheless shares many of the resource and skills constraints documented for smaller manufacturers, so the SME adoption literature remains the closest evidence base. The company works with modern CNC and robotic equipment for cutting, bending and welding steel, and with software for optimising sheet and plate yield. Its processes are laid down in an integrated quality manual and a set of procedures that cover sales and quotation, contract review, technical documentation, supplier approval, procurement and incoming inspection, production planning, in-process and final inspection, and internal

audit. Digitalisation, however, is patchy. Design and engineering rely on several CAD and 3D packages; administration and procurement run on an ERP system whose limitations are well known internally; and the shop floor still works largely on paper, supported by a single mobile application for quality.

As part of the lean rollout, fifteen semi-structured interviews covering ten roles were conducted in May 2026, from the general manager down to a line worker. Five problems came up again and again. Production is not planned at company level, so work stays reactive. Waiting is the largest single waste, with an estimated 30-35% of hours non-productive and about 30% of line tasks started without all of their preconditions in place. Technical and technological documentation arrives late or after the fact, and the resulting rework costs roughly EUR 100-1,000 a case. Material is ordered by hand and sometimes twice. And what the management system prescribes and what actually happens line up only about 49-60% of the time. These five problems are the backdrop against which the AI use cases are judged.

5 Research design

The work follows an exploratory single-case design (Cooper & Schindler, 2014) in three phases. The first analysed documents, including the quality manual, the procedures and the lean working instructions, alongside the fifteen interviews just described. The second mapped the AI use cases found in the literature onto the company's own processes, which is what produced the preliminary framework in Section 6. The third phase was an anonymous online survey of every employee who uses a computer and e-mail regularly, across seven departments: design and engineering; production preparation and technology; commercial and procurement; accounting; production; project management; and quality control. It ran in June 2026 on SurveyMonkey. Invitations went to 82 addresses, and 63 employees completed the questionnaire in full, a response rate of 76.8%. The instrument has 27 statements answered on a seven-point Likert scale (1 = strongly disagree, 4 = neither agree nor disagree, 7 = strongly agree). These are grouped into six blocks, namely usefulness and ease of use (B1), trust in AI (B2), organisational readiness (B3), the environment of market and regulation (B4), readiness and intention to use (B5), and the current state of work (B6). The blocks correspond in turn to the technological, organisational and environmental sides of the TOE framework (Tornatzky & Fleischer, 1990; Baker, 2012), with intention and current state added. The trust block was broken down into general trust, perceived reliability and accuracy, willingness to rely on the tools, and concern about data security. Six open questions followed, on tasks suited to AI, fears, time losses, training needs and ideas for the respondent's own department. Reliability by block, measured with Cronbach's alpha, runs from 0.76 to 0.92, and the open answers were coded thematically (Saldana, 2013). Participation was voluntary and anonymous, consent was implied by completing the questionnaire, and the answers were analysed only in aggregate, in line with the GDPR.

The interviews warrant fuller description. The ten roles span all five hierarchical levels of the company, from the general manager (strategic level), through tactical roles in engineering, production preparation, procurement, quality and production management, to operational staff, including a shop-floor line worker; roles held by several people were interviewed in parallel, which is how ten roles yield fifteen interviews. The interviews belong to the company's lean initiative: they were conducted in May 2026 for the lean rollout, not for AI specifically, and this study draws on them as qualitative evidence about the state of the processes. Participants were selected purposively so that every core process was represented. All interviews followed a common semi-structured guide covering the respondent's process, the wastes and bottlenecks encountered, current tool and documentation use, and expectations of the lean transformation. Transcripts were coded against a predefined lean framework, built from the five lean principles (Womack & Jones, 1996), the eight wastes (TIMWOODS) and a catalogue of fourteen lean tools, using a three-point presence scale (1 = explicitly mentioned; 0.5 = implicit or possible; 0 = not mentioned), and verbatim quotations of up to thirty words were retained for interpretive validity. The five recurring problems reported in Section 4 emerged from this coding; employee attitudes to AI were measured separately, by the survey described next.

The survey items likewise have a documented origin. The usefulness, ease-of-use and intention statements (B1, B5) were adapted from the technology-acceptance tradition (Davis, 1989; Venkatesh & Bala, 2008); the organisational and environmental statements (B3, B4) operationalise the TOE framework (Tornatzky & Fleischer, 1990; Baker, 2012); and the trust and data-security statements (B2) reflect the concerns documented for LLM adoption in construction (Ghimire et al., 2024; Ma et al., 2025). Item wording was reviewed with company management and refined through two internal iterations before distribution. In addition to descriptive statistics, block scores were formed as the mean of a respondent's items in each block, and the relationships among them were examined with Pearson correlations, an ordinary least squares regression on the intention block, and Kruskal-Wallis tests across levels of prior AI experience.

6 Preliminary implementation framework

6.1 Process-level AI use cases

Table 1 ties the company's processes and their documented problems to specific AI applications and the benefits each would bring. It combines the literature in Section 3 with the internal analysis in Section 4.

6.2 Departments, benefits and problems addressed

The first wave should take in the departments whose work is mostly about information: commercial and procurement, design and engineering, production preparation and technology, project management, accounting and quality management, with top management as the sponsor. The shop floor comes in only indirectly at this stage, through AI-generated visual work instructions and enlarged views of drawings, well before any direct automation. What the company stands to gain is time back on repetitive and administrative work, fewer reworks caused by documentation, less waiting once planning improves, quicker quotation, knowledge kept in a searchable base rather than in people's heads, and steadier quality. Each of these answers directly to one of the five problems set out in Section 4.

Table 1. Mapping of business processes to AI applications and expected benefits

Process	Documented problem	AI application	Expected benefit	Phase
Sales and quotation	Manual processing of inquiries	LLM-based extraction of requirements; draft offers	Faster, more consistent quotations	1
Design and engineering	Late or retroactive documentation; rework	Documentation completeness checks; generative and parametric design	Fewer reworks; faster design variants	1; 3
Production planning	No company-level plan; waiting	AI-supported scheduling extending the Last Planner System	Less waiting; higher plan reliability (PPC)	2
Procurement and warehouse	Manual and duplicate ordering	Reorder-point logic and duplicate detection over ERP data	Fewer stockouts and surpluses	1-2
Quality control	Manual inspection; no root-cause analysis	Machine-vision inspection; RAG-based defect diagnostics	Fewer defects and complaints	2; 1
Management and QM system	Administrative burden; audit preparation	RAG knowledge base over the quality manual and procedures	Lower administrative load; better compliance	0-1

6.3 Timing of adoption

Two arguments about timing follow from this. The first is that readiness is built by doing, through training, in-house capacity and partnerships (Ma et al., 2025; Sanchez et al., 2025), which argues for starting now with low-risk office and documentation tasks that do not hinge on the quality of production data. The second is that capital-heavy applications, such as robotics, automated guided vehicles and vision systems wired into the lines, assume stable processes and trustworthy data, and so belong after the lean work has settled and data interoperability has improved (Wang et al., 2025; Rangasamy & Yang, 2025). The path proposed here therefore begins at once with two phases. Phase 0 is governance: usage rules that satisfy the GDPR and the EU AI Act, an internal owner, and basic training. Phase 1 is an office AI assistant and a RAG knowledge base. Phase 2, a machine-vision pilot in quality control together with AI-supported planning, falls within six to twelve months, and Phase 3, generative design and robotics de-risked with digital twins, sits one to two years out. Given the low regional adoption rate and export customers who expect ever more digital maturity, putting off the first phases would simply throw away a cheap chance to learn.

6.4 Internal ownership and the role of external consultants

On outside expertise, the evidence points to a hybrid arrangement. A shortage of skills is one of the most consistently reported barriers (Sanchez et al., 2025; Wang et al., 2025), and manufacturers of this size seldom have the in-house ability to build a RAG system or a machine-vision application. Bringing in specialists for those well-defined technical pilots, and for the first round of training, therefore makes sense. Ownership of the initiative is another matter, and it has to stay inside the company. The interviews showed the parallel lean effort losing momentum whenever no one

internal owned it, and the adoption literature is consistent that management engagement is what carries these projects (Bevilacqua et al., 2025). Handing AI adoption over wholesale would risk a poor fit with the processes and a lasting dependency. The model that fits is therefore an internal owner of the AI initiative, tied into the lean governance structure and backed by external consultants only for specified technical deliverables.

Set against existing roadmaps, the proposed phasing is consistent with the staged, value-chain sequencing advocated for modular construction (Pan et al., 2022) and with the phased SME roadmap of Sanchez et al. (2025), while differing in two respects: it adds an explicit governance phase before any pilot, and it ties the sequence to the company's lean transformation and to measured employee readiness rather than to technology maturity alone. For now it remains a proposal rather than a tested framework: it has not been implemented or benchmarked against alternatives, and its value can be confirmed only by the planned Phase 0-1 pilots and a follow-up readiness measurement.

7 Survey results and discussion

7.1 Sample and demographics

Sixty-three employees completed the survey, spread across every department. The largest groups were design and engineering (19) and commercial and procurement (14), followed by project management (8), production preparation and technology (7), quality control (7), production (5) and accounting (3). By role, constructors and technologists (20) and managers (17) were the most numerous, alongside clerical staff and other professions. The sample is even across age, with most respondents between 30 and 49 (39), and tenure runs the full range, from under two years to more than twenty. These are people who live at a computer: 40 are at one all day, and another 18 for several hours. Experience with AI is already common. Sixteen use it regularly, 35 now and then, and only five never. Willingness to spend time learning is modest but real, with most prepared to give between half an hour and two hours a week, and only four unable to free up any time at all. Because participation required regular computer and e-mail use, the sample represents office and engineering staff rather than the shop floor, and the computer-use item reflects screen time during the working day, not shift length. What respondents mean by AI here is generative AI: the questionnaire introduced it with the examples of ChatGPT, Copilot and Claude, and these are also the tools they report using (Section 7.4).

7.2 Readiness across the TOE blocks

Table 2 gives the mean level of agreement and the reliability of each block. The strongest signal is intention. The readiness-and-intention block scores highest (M = 5.32), 86% of respondents sit above the midpoint, and the leading statements are the intention to use the tools (M = 5.59) and the willingness to try them (M = 5.54). Employees also see AI as something they can pick up, since the statement on ease of learning is the highest in the technology block (M = 5.40), even if they are more guarded about how far it will speed up their work or improve its quality (M = 4.49 and 4.35). Trust sits in the middle, at M = 4.49 across the positively worded statements. People are readier to lean on the tools (M = 4.89) than to credit them with consistent reliability (M = 4.13), and worry about the security and confidentiality of data is spread right across the scale (M = 4.11; SD = 1.77).

Table 2. Employee readiness by block (seven-point Likert scale)

Block	Content	Items	M (SD)	alpha
B1	Usefulness and ease of use (technology)	4	4.76 (1.23)	0.80
B2	Trust in AI (positively worded items)	4	4.49 (1.34)	0.92
B3	Organisational readiness	8	4.14 (1.09)	0.87
B4	Environment - market pressure	3	5.19 (1.11)	0.76
B5	Readiness and intention to use	4	5.32 (1.37)	0.92
B6	Current state (waiting; repetitive work)	2	4.95 (1.49)	-

Note. M = mean (1-7), SD = standard deviation, alpha = Cronbach's alpha. The data-security concern item (B2; M = 4.11) and the regulatory-knowledge item (B4; M = 3.90) are reported separately owing to their opposite direction and distinct content, respectively.

Table 3. Pearson correlations among the block scores

	B1	B2	B3	B4	B5	B6
B1 Usefulness and ease of use	1					
B2 Trust in AI	.82	1				
B3 Organisational readiness	.53	.47	1			
B4 Market pressure	.66	.64	.36	1		
B5 Intention to use	.74	.78	.48	.71	1	
B6 Current state	.41	.30	.14	.42	.44	1

Note. N = 63 (B6: N = 62, pairwise). Coefficients above .25 are significant at p < .05 and above .33 at p < .01. B2 comprises the positively worded trust statements; B4 the three market-pressure statements.

Pressure from outside is felt clearly enough. Respondents agree that customers and partners expect steadily more digitalisation (M = 5.65) and that AI would make the company more competitive (M = 5.06), but knowledge of the legal framework, the GDPR and the EU AI Act, trails behind (M = 3.90; SD = 1.97), which points to a real gap in regulatory literacy. The lowest scores belong to organisational readiness (M = 4.14). Inside that block, management support (M = 5.47) and the expectation of training (M = 5.02) are high, but a clear vision (M = 3.89), an open culture (M = 3.61), time to learn (M = 3.32) and the state of the data and IT base (M = 3.02) all fall below the midpoint. The current-state statements bear out the waiting problem from the interviews. Respondents agree strongly that they often wait for information or documentation (M = 5.34), and rather less that repetitive tasks eat their time (M = 4.56).

The same shape shows up throughout. At the level of the individual, readiness, openness and a sense that the tools are easy to use are all high, while the organisational footings of time, the data base, a clear vision and the culture lag behind. That is exactly the case for the order proposed in Section 6: begin with low-risk office and documentation work that does not depend on the quality of production data, put governance and training around it, and hold the more expensive applications until data and processes have been brought into shape. Strong management support sits next to a weak score for clarity of vision, and the combination says something specific, that the moment has come to write that vision down and say it out loud to the staff.

One caveat about scope follows from the instrument itself. Because the questionnaire introduced AI through generative-AI examples, the readiness and intention measured here concern generative-AI and LLM-based office tools, the technology of Phases 0 and 1, rather than the machine-vision and predictive-scheduling applications of Phases 2 and 3, which are a different technology class and will warrant their own readiness check before rollout. The alignment is deliberate: the framework begins exactly where measured readiness is highest.

7.3 Relationships among the constructs

Table 3 reports the correlations among the block scores. The intention to use AI (B5) is strongly associated with usefulness and ease of use (r = .74), trust (r = .78) and market pressure (r = .71; all p < .001), and moderately with organisational readiness (r = .48) and the current-state pressures of waiting and repetitive work (r = .44). Concern for data security is negatively related to trust (r = -.49, p < .001) and to intention (r = -.32, p = .010). Prior experience with AI goes hand in hand with usefulness (r = .48), trust (r = .49) and intention (r = .40; all p ≤ .001), whereas age is unrelated to intention (r = -.06, n.s.): readiness is broadly distributed across age groups, not a property of the young.

A regression of intention on the four TOE blocks explains 70% of its variance (R^2 = .70, adjusted R^2 = .68, F(4, 58) = 34.1, p < .001). Only trust (β = .42, p = .002) and market pressure (β = .31, p = .002) are significant independent predictors; usefulness adds little once trust is controlled, reflecting how tightly the two are entwined (r = .82). Intention, usefulness and trust all differ significantly by prior AI experience (Kruskal-Wallis, all p ≤ .002), while organisational readiness does not (p = .13), as expected, since it describes the company rather than the person. The pattern mirrors the wider evidence: experience-driven acceptance echoes the learning effects reported for generative-AI assistants (Brynjolfsson et al., 2025), and the primacy of trust matches the barriers reported for LLMs in construction (Ghimire et al., 2024; Ma et al., 2025). For the framework this means the first lever is trust, protected by governance, data-security rules and honest training, and the second is experience itself, which the low-risk Phase 1 pilots are designed to supply.

7.4 Qualitative findings

The open questions, answered by between 42 and 53 people, gave a concrete picture that lines up with the numbers (Table 4). Asked which tasks suit AI best, respondents named translation and foreign-language

correspondence, drafting and tidying up e-mails and letters, building material specifications by pulling data out of CAD and Revit models, searching technical data, standards and catalogues, and analysing prices and offers. These are almost exactly the information-heavy departments marked out for the first wave in Section 6.2, and the use cases in Table 1.

Fears are mild. About half of the respondents say outright that they have none. Among the rest, two worries dominate: whether the output is accurate and reliable, which they feel must always be checked, and the security and confidentiality of company data. A few warn about leaning on the tools too heavily and letting their own judgement slide, and a handful make the point that the company's systems and data have to be put in order before AI can do anything useful, which is the same thing the low score for data readiness was saying. On where time goes, the answers echo the interviews: documentation that is late or incomplete, waiting for materials and information, a poor flow of information, and stock figures in the system that do not match the shelf. The training people ask for is hands-on, meaning short workshops built around real examples and tied to the tools they already use (CAD, Excel, Opera), help with writing prompts, and clear rules on what may and may not be typed into a tool. Their own suggestions most often describe a system that would join up the production plan, the state of the documentation and the availability of resources and raise a flag when something is missing, together with a searchable store of past projects and solutions. These are, in other words, the two things the framework already rests on, AI-supported planning and a RAG knowledge base. A good number already use these tools at work. ChatGPT comes up most often (32 mentions), then Claude (12) and Gemini (8), with the odd mention of Copilot, NotebookLM and Mistral, while eight respondents use none. That this is already happening, unmanaged, is one more reason to put usage rules and governance in place sooner rather than later.

Table 4. Thematic summary of the open-ended responses

Open-ended question	Most frequent themes
Tasks suited to AI	Translation and foreign-language correspondence; drafting and revising e-mails and letters; material specifications from CAD/Revit; searching standards and technical data; price and offer analysis
Fears and concerns	No fear (about half); accuracy and reliability of outputs; security and confidentiality of data; over-reliance and decline in critical judgement
Time losses	Late and incomplete documentation; waiting for materials and information; poor information flow; inaccurate stock levels
Training needed	Practical, example-driven workshops; integration with existing tools (CAD, Excel, Opera); prompt-writing instruction; data-protection rules
Application proposals	Linking the production plan, documentation status and resources with alerts; a searchable knowledge base from the project archive
Tools in use	ChatGPT (32), Claude (12), Gemini (8); isolated Copilot, NotebookLM, Mistral; 8 respondents use none

7.5 Limitations and future work

The limits of the study follow from its design: a single case, self-reported answers, and a single point in time at one company. Any generalisation beyond comparable manufacturing mid-caps therefore needs care. Anonymity, useful as it is, also meant that non-response could not be examined below the level of the department, and the sample is too small to break down reliably into subgroups. A few of the review articles the synthesis leans on were available only as abstracts. Because the evidence comes from one organisation, the findings generalise analytically rather than statistically (Yin, 2018): they transfer most readily to manufacturing mid-caps of comparable size, export orientation and lean maturity. The correlational and regression results rest on 63 respondents and should be read as exploratory; the production subgroup in particular is too small to support departmental conclusions. The next steps are to put Phase 0 and Phase 1 into practice and then to measure readiness again after the first pilots, which would move the framework from a preliminary proposal to something tested against evidence.

8 Conclusion

This paper set recent work on AI in modular construction and on AI adoption in manufacturing SMEs against the documented process problems of one Croatian mid-cap manufacturer of modular buildings. The reading of the evidence is threefold. The nearest gains lie not in expensive automation but in office, documentation and quality work. Adoption should move in phases kept in step with the company's lean transformation, starting straight away with governance and a few low-risk pilots. And the most realistic way to govern it is a hybrid one, with internal ownership and outside expertise brought in only where it is

genuinely needed. Within the limits of a single case study, the survey backs this up. Employees are willing and open, but they rate the organisational footings, time, the data base, a clear vision and the culture, as the weakest link, and that is precisely what gives the governance framework, the training and the modest first pilots their justification. The regression points the same way: trust and perceived market pressure are the main levers of intention, so the immediate task is to protect trust through governance, data-security rules and honest communication about what the tools can and cannot do, while hands-on experience builds the rest.

Acknowledgments

The authors thank the employees and the management of the case company for taking part in the interviews and the employee survey, and the anonymous reviewers for their constructive comments.

References

AI Chamber. (2025). How do SMEs in CEE find their way in the world of AI? AI Chamber. https://aichamber.eu/report/how-do-smes-in-cee-find-their-way-in-the-world-of-ai/

Arroyabe, M. F., Arranz, C. F. A., Fernandez de Arroyabe, I. F., & Fernandez de Arroyabe, J. C. (2024). Analyzing AI adoption in European SMEs: A study of digital capabilities, innovation, and external environment. Technology in Society, 79, 102733. https://doi.org/10.1016/j.techsoc.2024.102733

Baker, J. (2012). The technology-organization-environment framework. In Y. K. Dwivedi, M. R. Wade, & S. L. Schneberger (Eds.), Information systems theory: Explaining and predicting our digital society (Vol. 1, pp. 231-245). Springer. https://doi.org/10.1007/978-1-4419-6108-2_12

Bevilacqua, S., Masarova, J., Perotti, F. A., & Ferraris, A. (2025). Enhancing top managers' leadership with artificial intelligence: Insights from a systematic literature review. Review of Managerial Science. https://doi.org/10.1007/s11846-025-00836-7

Brynjolfsson, E., Li, D., & Raymond, L. R. (2025). Generative AI at work. The Quarterly Journal of Economics, 140(2), 889-942.

Buntak, K., Kovacic, M., & Mutavdzija, M. (2019). Utjecaj sustava Industrije 4.0 na lean menadzment. In Proceedings of the 19th Croatian Quality Conference. Croatian Quality Society, Vodice, Croatia.

Chui, M., Hazan, E., Roberts, R., Singla, A., Smaje, K., Sukharevsky, A., Yee, L., & Zemmel, R. (2023). The economic potential of generative AI: The next productivity frontier. McKinsey & Company. https://www.mckinsey.com/capabilities/tech-and-ai/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier

Cooper, D. R., & Schindler, P. S. (2014). Business research methods (12th ed.). McGraw-Hill Education.

Davenport, T. H., & Ronanki, R. (2018). Artificial intelligence for the real world. Harvard Business Review, 96(1), 108-116.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly, 13(3), 319-340. https://doi.org/10.2307/249008

Enholm, I. M., Papagiannidis, E., Mikalef, P., & Krogstie, J. (2022). Artificial intelligence and business value: A literature review. Information Systems Frontiers, 24(5), 1709-1734. https://doi.org/10.1007/s10796-021-10186-w

European Commission. (2025). Commission Recommendation of 21 May 2025 on the definition of small mid-cap enterprises (C(2025) 3500). https://eur-lex.europa.eu/eli/reco/2025/1099/oj

Eurostat. (2025). Use of artificial intelligence in enterprises. Eurostat Statistics Explained. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Use_of_artificial_intelligence_in_enterprises

Ghimire, P., Kim, K., & Acharya, M. (2024). Opportunities and challenges of generative AI in construction industry: Focusing on adoption of text-based models. Buildings, 14(1), 220. https://doi.org/10.3390/buildings14010220

Heredia Alvaro, J. A., & Gonzalez Barreda, J. (2025). An advanced retrieval-augmented generation system for manufacturing quality control. Advanced Engineering Informatics, 64, 103007. https://doi.org/10.1016/j.aei.2024.103007

Kovic, K., Tominc, P., Prester, J., & Palcic, I. (2024). Artificial intelligence software adoption in manufacturing companies. Applied Sciences, 14(16), 6959. https://doi.org/10.3390/app14166959

Krcmar, T., Sabanovic, D., Svarcmajer, M., Kohler, M., Lukic, I., & Carevic, M. (2025). Beyond traditional BI: Large language models as game-changers in document processing. In Central European Conference on Information and Intelligent Systems (CECIIS 2025). Faculty of Organization and Informatics, University of Zagreb.

http://archive.ceciis.foi.hr/public/conferences/2025/Proceedings/S7/1.pdf

Liu, Q., Ma, Y., Chen, L., Pedrycz, W., Skibniewski, M. J., & Chen, Z.-S. (2024). Artificial intelligence for production, operations and logistics management in modular construction industry: A systematic literature review. Information Fusion, 109, 102423. https://doi.org/10.1016/j.inffus.2024.102423

Ma, L., Zhao, X., Jiang, R., Wu, C., Liao, L., Yang, Z., & Tan, J. (2025). Adopting large language models in the construction industry: Drivers, barriers, and strategic implications from China. Buildings, 15(23), 4296. https://doi.org/10.3390/buildings15234296

Mas-Machuca, M., Marimon, F., & Akhmedova, A. (2025). Generative AI and workplace productivity: A qualitative study in Spain. In Central European Conference on Information and Intelligent Systems (CECIIS 2025). Faculty of Organization and Informatics, University of Zagreb. http://archive.ceciis.foi.hr/public/conferences/2025/Proceedings/S9/1.pdf

Mikalef, P., & Gupta, M. (2021). Artificial intelligence capability: Conceptualization, measurement calibration, and empirical study on its impact on organizational creativity and firm performance. Information & Management, 58(3), 103434. https://doi.org/10.1016/j.im.2021.103434

Olmedo Mejia, D. F., & Cvetkoska, V. (2025). From human-centered to hybrid: Mapping the emergence of GenAI in organizational actor-networks. In Central European Conference on Information and Intelligent Systems (CECIIS 2025). Faculty of Organization and Informatics, University of Zagreb. http://archive.ceciis.foi.hr/public/conferences/2025/Proceedings/S5/4.pdf

Pan, M., Yang, Y., Zheng, Z., & Pan, W. (2022). Artificial intelligence and robotics for prefabricated and modular construction: A systematic literature review. Journal of Construction Engineering and Management, 148(9). https://doi.org/10.1061/(ASCE)CO.1943-7862.0002324

Rangasamy, V., & Yang, J.-B. (2025). AI-driven generative design and optimization in prefabricated construction. Automation in Construction, 177, 106350. https://doi.org/10.1016/j.autcon.2025.106350

Saldana, J. (2013). The coding manual for qualitative researchers (2nd ed.). SAGE Publications.

Sanchez, E., Calderon, R., & Herrera, F. (2025). Artificial intelligence adoption in SMEs: Survey based on TOE-DOI framework, primary methodology and challenges. Applied Sciences, 15(12), 6465. https://doi.org/10.3390/app15126465

Tornatzky, L. G., & Fleischer, M. (1990). The processes of technological innovation. Lexington Books.

Turcic, I., Stjepic, A.-M., & Juric, T. (2025). Leveling up process efficiency: AI as a game-changer for SMEs. In Central European Conference on Information and Intelligent Systems (CECIIS 2025). Faculty of Organization and Informatics, University of Zagreb. http://archive.ceciis.foi.hr/public/conferences/2025/Proceedings/S6/9.pdf

Ulrich, P., Frank, V., & Kratt, M. (2021). Adoption of artificial intelligence technologies in German SMEs - Results from an empirical study. PACIS 2021 Proceedings, 189.

Venkatesh, V., & Bala, H. (2008). Technology Acceptance Model 3 and a research agenda on interventions. Decision Sciences, 39(2), 273-315. https://doi.org/10.1111/j.1540-5915.2008.00192.x

Veza, I., Mladineo, M., & Gjeldum, N. (2016). Selection of the basic lean tools for development of Croatian model of innovative smart enterprise. Tehnicki vjesnik, 23(5), 1317-1324. https://doi.org/10.17559/TV-20160202120909

Wang, M., Cai, J., Hu, D., Hu, Y., Han, Z., & Li, S. (2025). AI-based robots in industrialized building manufacturing. Frontiers of Engineering Management, 12(1), 59-85. https://doi.org/10.1007/s42524-025-4099-x

Womack, J. P., & Jones, D. T. (1996). Lean thinking: Banish waste and create wealth in your corporation. Simon & Schuster.

Yevu, S., Blay, K., Ayinla, K., & Hadjidemetriou, G. (2025). Artificial intelligence in offsite and modular construction research. Automation in Construction, 171.

Yin, R. K. (2018). Case study research and applications: Design and methods (6th ed.). SAGE Publications.

Zron, A. (2025). Impact of technostress on the adoption of artificial intelligence: A bibliometric analysis. In Central European Conference on Information and Intelligent Systems (CECIIS 2025). Faculty of Organization and Informatics, University of Zagreb. http://archive.ceciis.foi.hr/public/conferences/2025/Proceedings/S5/1.pdf

Intelligent Systems

Ayşenur Elif Abanoz, Sevgi Koyuncu Tunç, Ece Çağlayan, Şemsettin Ekin Işık, Lütfiye Dila Çalık, Ahmet Burak Albayrak, Murat Tut, Utku Eyerci, Ayberk Özel, İlke Durmaz, İbrahim Sefa Sağlam
A Systematic Mapping of Paradigms, Devices, and Reproducibility Practices

Luka Lukač, Damjan Strnad
The Impact of Quantisation on Performance of Generative Adversarial Networks

Marko Horvat, Sven Leo Kupres, Marko Delimar
Assessing the importance of knowledge graphs for improving LLM answer accuracy using data augmentation

Lazerta Rista
Language-Aware Evaluation of AI-Driven Search: A Benchmark Study of Asymmetry in a Low-Resource Language

Dino Vlahek, Livija Ružić
Machine Learning for Subject Classification of Books Using Open Bibliographic Metadata

Antonia Šarčević, Alan Jović, Jakša Vukojević, Nikolina Frid
Cross-Method Analysis of Explainability in EEG-Based Major Depressive Disorder Classification

Darko Andročec
Vibe Coding and Web Application Security: A Twin-Prompt Study

Oliver Lohaj, Vladyslav Stakhov
Development of a Browser Extension for Toxic Content Detection Using an NLP Model

Kolok Peter
Correlation-Based Reduction of Environmental Input Variables for Efficient GRU-Based Prediction in IoT Systems

Hannah Hansa, Hanlie Smuts
Responsible AI Usage to Improve Knowledge Management For Sustainability Outcomes

Hanlie Smuts, Blossom Alloziem, Catherine Dyason, Keiara Govender
Preserving Human Agency in AI-Driven Interactions: The Role of User Experience Design

Dominik Čambál, Monika Herchlová, Juraj Ďuďák
A Comparative Study of SNN Learning Paradigms in Hybrid Graph-Spiking Neural Networks for Biomedical Data Analysis

Georgi Trajkov, Marko Grobelnik, Dunja Mladenić
From News Text to Causal Infographics via SemanticTriplets: Making an Opaque Image Pipeline Inspectable

Marko Horvat, Sven Leo Kupres
Evaluating AI-Generated Code: A Taxonomy and Comparative Catalog of Vibe-Coding Benchmarks

Goran Bubaš, Andreja Kovačić
Using Mind Maps to Investigate the Typology of Students' Use of Large Language Models: A Preliminary Study

David Slavik, Tim Juić
A Hybrid Document Reading, Few-Shot Classification, and LLM Extraction Pipeline for Accounting Document Processing

EEG-Based Brain-Computer Interfaces (2011-2025): A Systematic Mapping of Paradigms, Devices, and Reproducibility Practices

Sevgi Koyuncu Tunç, Ahmet Burak Albayrak, Lütfiye Dila Çalık, İlke Durmaz, Utku Eyerci, Şemsettin Ekin Işık, Ayberk Özel, İbrahim Sefa Sağlam, Murat Tut, Ayşenur Elif Abanoz

Department of Software Engineering, Faculty of Engineering, Çankaya University, Ankara
Yukarıyurtçu Mahallesi, Mimar Sinan Caddesi No:4, 06815 Etimesgut, Ankara, Türkiye
sevgik@cankaya.edu.tr,
{burak_albayrak0, ekinisik4}@icloud.com,
{dilaacalik, ilkedurmaz4, utkueyercii, ayberksu60, saglamibrahimsefa, tutmurattut, ayselifabanoz}@gmail.com

Ece Çağlayan

Department of Health Informatics, Graduate School of Informatics, Middle East Technical University (METU), Ankara, Türkiye
Üniversiteler Mahallesi, Dumlupınar Bulvarı No:1, 06800 Çankaya, Ankara, Türkiye
cece@metu.edu.tr

Abstract. *Reproducibility deficits, hardware-paradigm misalignment, and fragmented benchmarking constrain EEG-based brain-computer interface research. We systematically mapped 265 publications from Scopus and citation snowballing (2011-2025), coded across six dimensions by five reviewers (Cohen's κ = 0.83). Speller systems (38.3%) and command recognition (33.1%) dominated, while motor imagery accounted for 8.1%. Emotiv devices appeared in 41 of 52 P300 speller studies. Deep learning rose sharply after 2017, but transfer learning appeared in only 1.5%. Author-deposited data occurred in 5.3%, benchmark reuse in 8.7%, and verified code in 5.7%; 85.7% shared neither code nor data, and only 3.4% acknowledged reproducibility as a limitation.*

Keywords. EEG, brain-computer interface, systematic mapping, PRISMA, P300, SSVEP, motor imagery, open science, reproducibility, dataset availability

1 Introduction

EEG-based brain-computer interfaces decode neural activity recorded from the scalp into communication or control signals without muscular involvement (Farwell & Donchin, 1988; Wolpaw et al., 2002). Their appeal lies in non-invasiveness, millisecond temporal resolution, and falling hardware costs. Since Farwell and Donchin's P300 speller paradigm in 1988, the field has diversified to include steady-state visual evoked potential (SSVEP) systems, motor imagery (MI) interfaces, and mental command recognition, with growing interest in rehabilitation and affective computing. Affordable consumer headsets and off-the shelf machine learning libraries have substantially lowered the barrier to entry, driving a publication surge visible in our corpus: 9 papers in 2011 versus 33 in 2023.

Despite this growth, high within-subject accuracy in offline settings has not translated into deployable, generalisable systems. This gap is not unique to BCI research: applied intelligent systems more broadly face the same compounding problems of studies that cannot be replicated because neither data nor code are shared, design choices driven by hardware availability rather than task requirements, and progress that is hard to measure because each group benchmarks against private datasets. EEG-BCI is a particularly clear example, old enough to have produced several hundred papers and diverse enough to span paradigms and hardware grades, yet not previously examined for all three problems at once. We present a systematic mapping of the 2011-2025 EEG-BCI literature to characterise what the community has studied, how, and with what resources, and to identify the structural barriers that constrain progress. Five research questions guide the analysis:

RQ1: Which application domains and paradigms dominate the literature?

RQ2: Which signal processing and classification methods are most prevalent, and how has this changed over time?

RQ3: What EEG hardware is used, and how does device grade relate to paradigm choice?

RQ4: What is the state of open-science compliance in terms of dataset and code sharing?

RQ5: What methodological limitations recur, and how do they constrain the field structurally?

Prior EEG-BCI reviews have each addressed a slice of these questions. Nicolas-Alonso and Gomez-Gil

(2012) surveyed paradigms and signal processing methods up to 2012. Lotte et al. (2018) focused on classification algorithms over the following decade and noted that transfer learning remained underused despite its theoretical promise. Rashid et al. (2020) provided a broader landscape review but did not systematically examine hardware selection or open science practices. Each of these is a valuable contribution, but none treats hardware, methodology, and reproducibility jointly, and none attempts to measure whether the outputs of the field are actually reusable. More recent work has examined individual facets of these questions in isolation. Chevallier et al. (2024) built a large open benchmark (MOABB) that re implements and compares machine-learning pipelines across public EEG datasets, addressing reproducibility at the level of algorithm evaluation. Saibene et al. (2024) systematically reviewed deep learning for motor-imagery decoding across 394 papers, documenting the post-2017 shift toward deep architectures and the reproducibility practices used within that subfield. On the hardware side, Zhang et al. (2023) surveyed the proliferation of wearable EEG BCI devices for medical applications, and Lee et al. (2024) empirically evaluated consumer-grade wireless headsets against a research-grade reference and reported measurable signal-quality gaps. Each of these advances one dimension (benchmark reproducibility, deep-learning methodology, or hardware capability), but none maps paradigms, devices, methods, and open science practice jointly across the full corpus, which is the structural question this study addresses.

This study makes three contributions. First, it combines paradigm distribution, hardware selection, signal processing trends, and open-science compliance in a single coded dataset, so their interactions become visible. Second, it separates studies that deposited original data to a public archive from those that merely reused pre-existing benchmarks such as BCI Competition or PhysioNet: conflating the two overstates data sharing, since 23 papers reuse benchmarks but only 14 deposit new data, an author deposit rate of 5.3%. Third, it reads these findings as an instance of the broader reproducibility problem in intelligent systems research rather than a peculiarity of BCI.

2 Methodology

2.1 Search Strategy and Screening

The literature search was conducted in October 2025 using the Scopus database. The Boolean query combined EEG and BCI terminology with signal processing and hardware keywords: TITLE-ABS KEY(("electroencephalography" OR "EEG") AND ("brain computer interface" OR "BCI" OR "brain machine interface" OR "BMI") AND ("communication" OR "speller" OR "mental command" OR "command recognition" OR "assistive communication") AND ("signal processing" OR "feature extraction" OR "classification" OR "machine learning" OR "deep learning" OR "canonical correlation analysis" OR "CCA" OR "SSVEP" OR "P300" OR "motor imagery") AND ("headset" OR "OpenBCI" OR "Emotiv" OR "Muse" OR "wearable" OR "portable")) AND PUBYEAR > 2010 AND PUBYEAR < 2026 AND (LIMIT-TO (SRCTYPE,"p") OR LIMIT-TO (SRCTYPE,"j")) AND (LIMIT-TO (DOCTYPE,"cp") OR LIMIT-TO (DOCTYPE,"ar")) AND (LIMIT-TO (LANGUAGE,"English") OR LIMIT-TO (LANGUAGE,"Turkish"))

Table 1. Screening summary (PRISMA 2020)

Stage	Action	Records
Database search	Scopus Boolean query (October 2025)	254
Supplementary	Citation snowballing + open repository search	20
Total identified	Candidate pool before screening	274
Deduplication	Exact duplicates removed	-2
Exclusion	Non-EEG modality studies	-2
Exclusion	Conceptual or theoretical, no empirical validation	-2
Exclusion	Editorials and review-only publications	-3
Final corpus	Full-text analysis	265

Scopus was chosen as the primary database because it provides broad coverage of both engineering and computer science venues relevant to this mapping. We acknowledge that this choice has consequences for the composition of the corpus. Web of Science also indexes relevant interdisciplinary journals that may not be fully represented in Scopus. IEEE Xplore indexes a substantial body of biomedical engineering and signal processing work that does not always appear in Scopus, and PubMed covers clinical and rehabilitation-oriented BCI research that is similarly underrepresented. Studies in these venues tend to focus on motor imagery, high-density EEG, and clinical populations. The relative scarcity of motor imagery papers in our results (8.1% of empirical studies) should therefore be interpreted with this search boundary in mind: it may partly reflect a genuine paradigm distribution within Scopus-indexed venues rather than the field as a whole. The same caveat applies to the hardware and open science distributions reported in the Results section.

To partially compensate for this limitation, we supplemented the database search with citation snowballing from the retrieved papers and manual searches of four open data repositories: PhysioNet, OpenNeuro, BCI Competition, and Kaggle. This added 20 further studies, giving 274 candidates in total.

Screening followed PRISMA 2020 guidelines (Page et al., 2021). Two papers were removed as exact duplicates. Six further papers were excluded on the following grounds: two used a non-EEG primary modality, two were conceptual or theoretical with no

empirical validation, and two were editorials or review-only publications. The final corpus of 265 papers was retained for full-text analysis. Table 1 summarises each screening stage.

The corpus contains 248 empirical studies and 17 secondary-literature items, including review and survey papers. The 17 secondary items were retained to support background synthesis but are excluded from all quantitative counts in the Results section. Unless otherwise stated, percentages in the Results are computed over the empirical subset of 248 papers. Open-science and hardware counts use the full corpus of 265 papers, since every publication regardless of type makes an implicit reproducibility decision. Inclusion required: (i) EEG as the primary acquisition modality; (ii) description of a signal processing or classification pipeline; and (iii) empirical results in a BCI communication or control context.

2.2 Data Extraction and Inter-Rater Agreement

Five coders independently extracted data from 52-54 papers each. Each paper was coded across six dimensions: (1) primary application domain and paradigm; (2) EEG hardware and electrode type; (3) preprocessing, feature extraction, and classifier; (4) dataset availability, scored positive only when a verifiable URL, DOI, or named repository was cited (generic "available upon request" statements were not counted); (5) source code availability, scored positive only when a live, version-controlled repository link was provided; and (6) self-reported limitations, extracted verbatim and grouped into thematic categories. A shared codebook specifying decision rules and worked examples was used throughout. To estimate inter-rater reliability, a stratified random sample of 30 papers (11.3% of the corpus) was coded independently by a second coder. Disagreements were resolved by discussion. Cohen's kappa across all six dimensions on this overlap set reached 0.83 (substantial agreement, following the convention of Landis and Koch as cited in (Petersen et al., 2015)), with per-dimension values ranging from 0.78 (application domain) to 0.91 (code availability). Of the 248 empirical papers, 155 (62.5%) reported a quantitative accuracy metric (e.g., classification accuracy, AUC, or F1-score). The remaining 93 papers reported only qualitative outcomes or no classification accuracy metric and were excluded from that analysis. For consistency and comparability across studies, only classification accuracy values were included in the descriptive analysis presented in Section 4.4.

3 Results

3.1 Publication Trend

The 265 papers span 2011 to 2025. Annual volume grew from 9 papers in 2011 to a peak of 33 in 2023, with 29 papers in 2024 and 12 in 2025. The 2025 count is a lower bound due to indexing lag at the time of search and should not be interpreted as a downturn in publication activity. A sustained increase from 2018 onward accounts for 64.5% of the corpus (171 papers), coinciding with the appearance of deep learning methods in the EEG-BCI literature (Section 3.3) and the proliferation of low-cost consumer headsets. Figure 1 illustrates the annual publication trend across the corpus.

3.2 Application Domains and Paradigms (RQ1)

Table 2 shows paradigm frequencies for the 248 empirical papers. Categories are multi-label: a hybrid P300/SSVEP study is counted in both sub-rows, so percentages sum to more than 100%. Speller-type systems (95 papers, 38.3%) and command recognition (82 papers, 33.1%) together account for roughly 70% of empirical output. Motor imagery, despite its theoretical centrality, is the primary focus of only 20 papers (8.1%). Rehabilitation and neurofeedback account for 15 papers (6.0%), emotion recognition for 7 (2.8%), and drone or UAV control for 6 (2.4%), confirming that clinically and ecologically motivated applications remain marginal.

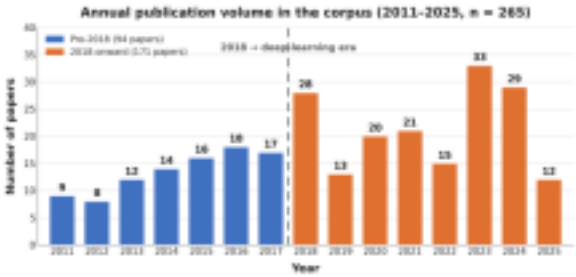


Figure 1. Annual publication volume in the corpus (2011-2025, n = 265). The dashed line marks 2018, after which deep learning methods appear in the corpus.

Table 2. Paradigm and application distribution (empirical papers, n = 248; multi-label; % of n = 248)

Application / Paradigm	n	%	Primary Signal Feature
Speller systems (total)	95	38.3%	ERP / frequency-tagged response
P300 speller	52	21.0%	N200/P300 complex at Pz
SSVEP speller	5	2.0%	Frequency-tagged occipital response
Command recognition	82	33.1%	Band power, spectral features
Binary control (yes/no, left/right)	47	19.0%	ERP or SSVEP frequency
Object / cursor control	43	17.3%	MI event-related desynchronisation
Motor imagery (primary focus)	20	8.1%	mu/beta ERD/ERS at C3-Cz-C4

Application / Paradigm	n	%	Primary Signal Feature
Wheelchair / mobility	18	7.3%	SSVEP or MI commands
Hybrid paradigms	12	4.8%	Multi-modal feature fusion
Rehabilitation / neurofeedback	15	6.0%	MI + closed-loop feedback
Emotion recognition	7	2.8%	Frontal alpha asymmetry, gamma
Drone / UAV control	6	2.4%	SSVEP / mental commands

3.3 Signal Processing and Classification (RQ2)

ERP-based P300 pipelines appear in 86 papers (32.5%), consistent with speller dominance. SSVEP with canonical correlation analysis (CCA) appears in 56 papers (21.1%). SVM (47, 17.7%) and LDA (44, 16.6%) remain the dominant classifiers (Lotte et al., 2018). CNN and other deep learning architectures, including EEGNet (Lawhern et al., 2018), appear in 51 papers (19.2%). Of these, 2 predate 2018 (2016 and 2017), while 49 are from 2018 onward, confirming a sharp adoption boundary at that year. Transfer learning or domain adaptation appears in only 4 papers (1.5%), well below the 4.1% identified by Lotte et al. in 2018 (Lotte et al., 2018). CSP-based spatial filtering, a well-established MI technique (Ang et al., 2012), appears in 14 papers (5.3%). Table 3 summarises the signal processing and classification methods identified in the corpus.

Table 3. Signal processing and classification methods (n = 265; multi-label; % of n = 265)

Method / Classifier	n	%	Paradigm Association
ERP / P300 pipeline	86	32.5%	P300 speller, ERP-based BCI
SSVEP + CCA pipeline	56	21.1%	SSVEP frequency decoding
SVM (linear, RBF, polynomial)	47	17.7%	MI, binary, command recognition
LDA / Fisher discriminant	44	16.6%	MI, P300, general classification
CNN / deep learning (EEGNet, LSTM)	51	19.2%	Cross-paradigm (MI, emotion), post-2017 primarily
FFT / power spectral density	29	10.9%	Preprocessing, SSVEP band power
CSP (common spatial patterns)	14	5.3%	Motor imagery spatial filtering
Wavelet transform (DWT/MODWT)	14	5.3%	Feature extraction, denoising
Transfer learning / domain adapt.	4	1.5%	Cross-subject generalisation

3.4 Hardware Utilisation (RQ3)

Table 4 presents device frequency across all 265 papers. Emotiv devices appear in 92 papers (34.7%), making them the dominant hardware platform by a wide margin. Including OpenBCI (23), NeuroSky (9), and Muse (8), consumer-grade devices appear in 132 studies (49.8%). Research-grade systems, namely g.tec (24), Neuroscan (13), and BrainProducts (4), collectively account for 41 papers (15.4%). The remainder use custom amplifiers, unspecified laboratory equipment, or benchmark datasets without recording hardware specification.

The hardware distribution has a direct methodological consequence. Table 5 presents a device-by-paradigm cross-tabulation derived from the coded dataset. Emotiv devices appear in 41 of the 52 coded P300 speller studies and in 29 of 82 command recognition studies. Section 4.2 examines why this concentration represents a mismatch between device capability and paradigm requirements rather than a design choice supported by the literature.

Table 4. EEG hardware platforms identified in the corpus (n = 265)

Device	n	Grade	Specifications (channels, sampling rate, electrode)	Suitable For
Emotiv EPOC / EPOC+ / Insight	92	Consumer	5-14 ch, 128-256 Hz, wet (saline felt)	Command, binary, SSVEP (speller: limited)
OpenBCI Cyton + Ultracortex	23	Consumer	8-16 ch, 250 Hz, wet/dry	SSVEP, motor imagery, P300 speller
g.tec (Nautilus, Unicorn, etc.)	24	Research	8-64 ch, 250-500 Hz, wet/hybrid	All paradigms; clinical and research grade
Neuroscan SynAmps	13	Research	32-64 ch, 1000 Hz, wet Ag/AgCl	Clinical ERP, source imaging
NeuroSky MindWave	9	Consumer	1 ch, 512 Hz, dry	Attention, binary (single-channel)
Muse / Muse 2	8	Consumer	4 ch, 256 Hz, dry	Introductory / attention paradigms
BrainProducts actiCHamp	4	Research	32-160 ch, 1000 Hz, active-wet	High-density ERP, source analysis
BCI2000 (software platform)	5	Platform	Device-agnostic framework (underlying device varies)	Framework; underlying device varies

Table 5. Device-by-paradigm cross-tabulation (columns cover the seven most frequent paradigm categories)

Device	P300 Speller	SSVEP	Command Recog.	Binary	Object Control	Motor Imagery	Rehab / NFB
Emotiv	41	4	29	19	12	6	7
OpenBCI	6	6	10	2	3	4	6
g.tec	14	2	6	1	4	0	1
Neuroscan	7	5	3	2	2	2	1
NeuroSky	2	1	3	3	1	0	1
Muse	0	0	1	0	0	0	0
BrainProducts	1	1	1	1	1	0	0
BCI2000	3	0	1	1	0	1	0

Each cell is the number of studies using that device for that paradigm. Because coding is multi-label, a single study may appear under more than one device or paradigm, so cell counts are not mutually exclusive. Shaded cells highlight the concentration of Emotiv consumer headsets in P300 speller and command-recognition studies.

3.5 Open-Science Compliance (RQ4)

Table 6 presents the open-science audit. Code was counted as available only when a live, version-controlled repository URL was verifiable at time of search; 15 papers (5.7%) meet this criterion. Dataset availability was divided into two categories: author deposited data (original data deposited by the paper's authors to GitHub, Zenodo, Figshare, OpenNeuro, OSF, IEEE Dataport, or Kaggle; n = 14, 5.3%) and benchmark reuse (papers that used a pre-existing public dataset such as BCI Competition II/III/IV or PhysioNet eegmmidb without depositing new data; n = 23, 8.7%). Only 14 papers (5.3%) share both verifiable code and some form of public data. The remaining 227 papers (85.7%) share neither.

Table 6. Open-science compliance (n = 265; % of n = 265)

Metric	n	%	Notes
Code: verified repository URL	15	5.7%	GitHub / GitLab / OSF link confirmed live at search date
Data: author deposited original data	14	5.3%	Deposited to Zenodo, Figshare, OpenNeuro, OSF, IEEE Dataport, or Kaggle by the authors
Data: benchmark reuse only	23	8.7%	Paper used BCI Competition II/III/IV, PhysioNet eegmmidb, or similar pre-existing repository
Both code and some public data	14	5.3%	Minimum artefacts for independent replication present
Neither code nor any public data	227	85.7 %	Independent replication not possible from published materials

The gap between objective non-sharing prevalence (85.7%) and self-reported acknowledgement is notable. Only 9 papers (3.4%) mention reproducibility or open-science as a limitation, indicating that the community treats non-sharing as a norm rather than a methodological deficit.

3.6 Self-Reported Limitations (RQ5)

Table 7 summarises limitation themes coded from the 265 papers. Signal quality issues (70 papers, 26.4%) are the most frequently cited limitation, followed by accuracy or performance ceiling (53, 20.0%) and real time or latency constraints (53, 20.0%). Cross-subject generalisation failure is cited in 31 papers (11.7%), and small or non-diverse participant cohorts in 35 (13.2%). Notably, 21 papers (7.9%) report no limitations at all, which is implausible for any empirical EEG study. Reproducibility or open-science deficit is cited in only 9 papers (3.4%), contrasting sharply with the 85.7% non-sharing rate reported in Section 3.5.

Table 7. Self-reported limitation themes (multi label; n = 265; % of n = 265)

Limitation Theme	n	%	Implication
Signal noise / artefact contamination	70	26.4 %	Exacerbated by dry consumer electrodes and limited channel counts
Accuracy / performance ceiling	53	20.0 %	Lab-only gains not demonstrated to hold in deployment conditions
Real-time / latency constraints	53	20.0 %	Most studies are offline; closed-loop validation is rare
Small / non-diverse participant cohorts	35	13.2 %	Healthy young adults only; clinical populations underrepresented
Lack of cross subject generalisation	31	11.7 %	Transfer learning underused (4 papers, 1.5%)
No limitations stated	21	7.9%	Implausible for empirical EEG studies
Reproducibility / open-science deficit	9	3.4%	Critically under-reported relative to actual non-sharing rate (85.7%)

4 Discussion

4.1 Paradigm Concentration and Application Gap

The paradigm distribution has remained stable over fourteen years. P300 and SSVEP speller systems together account for roughly 40% of empirical output, consistent with Nicolas-Alonso and Gomez-Gil (2012). This stability reflects path dependency rather than deliberate prioritisation: established paradigms attract further work by lowering the cost of entry through shared benchmarks and familiar pipelines. Rehabilitation, hybrid closed-loop systems, and emotion recognition collectively account for fewer than 15% of empirical papers despite having clearer clinical rationale. Motor imagery at 8.1% deserves separate attention given its theoretical centrality. Two explanations are plausible. MI research may concentrate in IEEE TNSRE and PubMed-indexed clinical venues not captured by our Scopus search. Alternatively, the dominance of Emotiv devices in this corpus creates a selection pressure against MI, which requires stable electrode placement at C3, Cz, and C4 that consumer headsets cannot reliably provide. Table 5 is consistent with the second explanation: MI appears in only 6 of 92 Emotiv studies. The 8.1% figure should be treated as a lower bound rather than an accurate representation of MI's role in the broader field. Transfer learning presents a related puzzle. Deep learning adoption grew sharply after 2017, yet transfer learning appears in only 1.5% of studies, lower than the 4.1% Lotte et al. identified in 2018. This is most plausibly a downstream consequence of the reproducibility problem: transfer learning requires multi-subject data collected under comparable conditions, and in a field where 85.7% of studies use private datasets, those preconditions rarely exist.

4.2 Hardware-Paradigm Misalignment

Emotiv devices appear in 41 of 52 coded P300 speller studies. This matters because P300 detection depends on a low-amplitude ERP component at parieto-occipital electrodes, and Emotiv's limited channel count and fixed placement provide restricted coverage of precisely that region. A practical alignment rule emerges from the data: SSVEP and command recognition paradigms are tolerant of consumer-grade hardware because their signal features do not require dense spatial coverage, whereas P300 spellers and motor imagery tasks have specific electrode placement requirements that consumer devices often cannot meet (Arpaia et al., 2022). Hardware availability rather than paradigm suitability appears to drive device selection across a substantial portion of the corpus, introducing avoidable performance constraints that are rarely disclosed.

4.3 The Structural Reproducibility Problem

The reproducibility picture from Section 3.5 is the paper's central structural finding. With most studies sharing neither code nor data, and benchmark reuse counted separately from genuine author deposits, the field generates 30 or more papers per year that are nearly all built on private datasets compared only against other private datasets within the same study. The result is a literature where reported accuracy figures cannot be aggregated and models cannot be validated across laboratories.

That only 9 papers cite reproducibility as a self-reported limitation despite an 85.7% non-sharing rate indicates that non-sharing is treated as a norm. These findings suggest that improving reproducibility requires structural interventions beyond individual researcher practices. Such interventions may include journal data availability policies, funder mandates for persistent archiving, and community agreement on a maintained shared repository for new EEG-BCI datasets. The infrastructure to implement these changes already exists.

4.4 Accuracy Reporting and Ecological Validity

Across the 155 empirical papers reporting a quantitative accuracy, the per-paper median accuracy was 90.9% and the mean is 87.6%. The large majority of these use offline, within-session, within-subject evaluation, the most favourable possible conditions for classification performance. Twenty-six papers report accuracy at or above 99%, nearly all from single subject offline evaluations. Cross-session or cross subject validation is rarely reported. High median accuracy therefore coexists with minimal cross-subject validation and no shared benchmark that would allow direct comparison across groups. Whether reported gains accumulate into field-level progress is a question that current reporting norms do not allow us to answer.

5 Conclusion

This systematic mapping of 265 EEG-BCI papers from 2011 to 2025 produces five findings mapped onto the research questions. The paradigm landscape (RQ1) is concentrated around P300 and SSVEP spellers and has not substantially diversified over fourteen years, leaving rehabilitation, hybrid closed-loop systems, and affective computing underdeveloped. Deep learning (RQ2) arrived sharply after 2017 and is now the fastest growing methodological strand, though classical SVM and LDA pipelines remain prevalent on small private datasets. Hardware selection (RQ3) is frequently misaligned with paradigm requirements, most visibly in the concentration of Emotiv consumer headsets in P300 speller studies. Open-science compliance (RQ4) is low: most studies share neither code nor data, and author-deposited datasets are rarer still. Finally, the

recurring self-reported limitations (RQ5), namely signal quality, accuracy ceiling, and real-time constraints, reflect structural features of the field rather than isolated study weaknesses, while reproducibility itself is almost never acknowledged as a limitation.

The most tractable interventions are not algorithmic. Researchers should: (i) select hardware based on paradigm requirements rather than convenience; (ii) deposit code to a persistent repository such as GitHub with a Zenodo DOI, and deposit EEG data to PhysioNet or OpenNeuro, as a default practice; (iii) report cross-subject or cross-session validation alongside within-subject offline results; and (iv) evaluate against established public benchmarks before collecting new private data. None of these steps is algorithmic; they are changes in editorial policy and community practice that would make open, requirement-driven research the default rather than the exception.

Study limitations. The search was limited to the Scopus database and did not include Web of Science, IEEE Xplore, or PubMed. Consequently, some relevant studies may not have been captured. Open science compliance was operationalised as the binary presence of a verifiable URL at search date, which is a conservative lower bound: repositories may have become unavailable after publication. Method classification relied on textual descriptions in papers rather than independent code verification. The 2025 publication count (n = 12) underestimates the true annual volume due to indexing lag.

References

Ang, K. K., Chin, Z. Y., Zhang, H., & Guan, C. (2012). Filter bank common spatial pattern algorithm on BCI competition IV datasets 2a and 2b. Frontiers in Neuroscience, 6, 39.

Arpaia, P., Esposito, A., Natalizio, A., & Parvis, M. (2022). How to successfully classify EEG in motor imagery BCI: A metrological analysis of the state of the art. Journal of Neural Engineering, 19(3), 031002. https://doi.org/10.1088/1741-2552/ac74e0

Chevallier, S., Carrara, I., Aristimunha, B., Guetschel, P., Sedlar, S., Lopes, B., Velut, S., Khazem, S., & Moreau, T. (2024). The largest EEG-based BCI reproducibility study for open science: the MOABB benchmark. arXiv preprint arXiv:2404.15319. https://doi.org/10.48550/arXiv.2404.15319

Farwell, L. A., & Donchin, E. (1988). Talking off the top of your head: Toward a mental prosthesis utilizing event-related brain potentials. Electroencephalography and Clinical Neurophysiology, 70(6), 510-523.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1), 159-174.

Lawhern, V. J., Solon, A. J., Waytowich, N. R., Gordon, S. M., Hung, C. P., & Lance, B. J. (2018). EEGNet: A compact convolutional network for EEG-based brain-computer interfaces. Journal of Neural Engineering, 15(5), 056013.

Lee, S., Kim, M., & Ahn, M. (2024). Evaluation of consumer-grade wireless EEG systems for brain-computer interface applications. Biomedical Engineering Letters, 14, 1433-1443. https://doi.org/10.1007/s13534-024-00416-w

Lotte, F., Bougrain, L., Cichocki, A., Clerc, M., Congedo, M., Rakotomamonjy, A., & Yger, F. (2018). A review of classification algorithms for EEG-based BCIs: A 10-year update. Journal of Neural Engineering, 15(3), 031005.

Nicolas-Alonso, L. F., & Gomez-Gil, J. (2012). Brain computer interfaces: A review. Sensors, 12(2), 1211-1279.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, n71.

Petersen, K., Vakkalanka, S., & Kuzniarz, L. (2015). Guidelines for conducting systematic mapping studies in software engineering: An update. Information and Software Technology, 64, 1-18. https://doi.org/10.1016/j.infsof.2015.03.007

Rashid, M., Sulaiman, N., Abdul Majeed, A. P. P., Musa, R. M., Ab Nasir, A. F., Bari, B. S., & Khatun, S. (2020). Current status, challenges, and possible solutions of EEG-based brain-computer interface: A comprehensive review. Frontiers in Neurorobotics, 14, 25.

Saibene, A., Ghaemi, H., & Dagdevir, E. (2024). Deep learning in motor imagery EEG signal decoding: A systematic review. Neurocomputing, 610, 128577. https://doi.org/10.1016/j.neucom.2024.128577

Wolpaw, J. R., Birbaumer, N., McFarland, D. J., Pfurtscheller, G., & Vaughan, T. M. (2002). Brain-computer interfaces for communication and control. Clinical Neurophysiology, 113(6), 767-791.

Zhang, J., Li, J., Huang, Z., Huang, D., Yu, H., & Li, Z. (2023). Recent progress in wearable brain-computer interface (BCI) devices based on electroencephalogram (EEG) for medical applications: A review. Health Data Science, 3, 0096. https://doi.org/10.34133/hds.0096

Appendix: Open-Science Resource Catalogue

This appendix catalogues the studies in the reviewed corpus that provide publicly accessible datasets (Table A1) or verified source-code repositories (Table A2). URLs were verified at time of corpus search (October 2025). Studies are listed alphabetically by first author. In Table A1, entries marked * cite a pre-existing benchmark repository and are not original author

deposits; entries without * deposited new data. Brophy et al. (2021) appears only in Table A2: the study used existing public datasets but did not deposit new data.

Table A1. Studies providing publicly accessible datasets (n = 37 total; 14 author-deposited, 23 benchmark reuse). Entries marked * are benchmark reuse only.

Citation	Dataset / Repository	URL / DOI	Paradigm
Cariello et al., 2023	Project data and resources page	dees.unict.it/mbucolo/index.php/resources	Rehabilitation
Clayton et al., 2020	FEIS: 14-ch imagined speech (21 subjects)	doi.org/10.5281/zenodo.3369179	Imagined speech
Fouad et al., 2021	P300 EEG dataset (Kaggle)	kaggle.com/islamfouad/p300dataset	P300 speller
Ganesan et al., 2025	*PhysioNet eegmmidb (BCI2000, 109 subjects)	physionet.org/content/eegmmidb/1.0.0/	Motor imagery
Kim et al., 2019	Nakanishi 2015 SSVEP (12-target, 10 subjects)	github.com/mnakanishi/12JFPM_SSVEP	SSVEP
Lawhern et al., 2018	*BCI Comp IV-2a; Kaggle ERN; RSVP P300	bbci.de/competition/iv/	P300, ERN, MI, SMR
Li et al., 2024	Semantic EEG dataset (26 days)	github.com/cfli20/EEG	Semantic decoding
Liu et al., 2024	MI EEG, 50 stroke patients, BIDS (Figshare)	doi.org/10.1038/s41597-023-02787-8	Motor imagery / rehab
Mattioli et al., 2021	*PhysioNet eegmmidb (BCI2000)	physionet.org/content/eegmmidb/1.0.0/	Motor imagery
Mezzina et al., 2021	*BCI Competition III dataset	bbci.de/competition/iii/	Motor imagery
Molina-Cantero et al., 2019	Blinking experiments dataset (ResearchGate)	researchgate.net/publication/331320020	EOG / blink
Moro et al., 2017	*BCI Competition II Dataset IIb (Wadsworth)	bbci.de/competition/ii/	ERP / P300
Novais et al., 2024	*BCI Competition II and III (referenced)	bbci.de/competition/ii/	Motor imagery (review)
Parui et al., 2022	*BCI Comp IV DS1 + BCI Comp III DSIVa	bbci.de/competition/iv/	Motor imagery
Peterson et al., 2020	MI-OpenBCI dataset (GitHub)	github.com/vpeterson/MI-OpenBCI	Motor imagery
Pires et al., 2011	*BCI Competition 2003 dataset	bbci.de/competition/ii/	P300 speller
Rajashekhar et al., 2021	*MIT-BIH Noise Stress Test DB	physionet.org/content/nstdb/	Artefact benchmarking
Raji et al., 2024	OpenNeuro ds004956 (Haupt et al.)	openneuro.org/datasets/ds004956/versions/1.0.1	ERP / cognitive
Ramele et al., 2018	*BNCI-Horizon 008-2014; BCI Comp II IIb	bnci-horizon-2020.eu/database/data-sets	P300 speller
Sahoo et al., 2024	Odia BCI dataset (IEEE Dataport)	dx.doi.org/10.21227/wg17-6068	P300 / language BCI
Saibene et al., 2023	*BCI Comp + PhysioNet eegmmidb (review)	physionet.org/content/eegmmidb/1.0.0/	Motor imagery (review)
Sampanna et al., 2018	*BCI Competition II and III	bbci.de/competition/iii/	Motor imagery
Schneider et al., 2020	*BCI Competition IV-2a	bnci-horizon-2020.eu/database/data-sets	Motor imagery
Sharma et al., 2024	Bonn Univ epilepsy; DEAP; UCI alcoholism (reviewed)	Multiple; see paper	Review of datasets
Stock et al., 2016	*BCI Competition II Dataset 3	bbci.de/competition/ii/	SSVEP
Sun et al., 2023	*BCI Competition IV-2a (survey reference)	bbci.de/competition/iv/	SSVEP / drone (survey)
Suri et al., 2023	*BCI Comp III + IV; PhysioNet eegmmidb (review)	bbci.de/competition/	Motor imagery (review)
Tahmasebzadeh et al., 2013	*BCI Competition 2003 IIb	bbci.de/competition/ii/	Motor imagery
Tiwari et al., 2022	*BCI Competition III Dataset V	bbci.de/competition/iii/	Motor imagery
Tiwari et al., 2023	Mind Big Dataset (MBD)	mindbigdata.com/opendb/	Mental imagery / digits
Yan et al., 2023	*BCI Comp IV-2a; High-Gamma Dataset	bbci.de/competition/iv/	Motor imagery

Citation	Dataset / Repository	URL / DOI	Paradigm
Zao et al., 2014	BCI Web Portal + RESTful API	bci.pet.cs.nctu.edu.tw/databank	Multi-modal / game BCI
Zhao et al., 2020	*IFMBE MEDICON 2019; BCI Comp III DS-II	medicon2019.org/scientific-challenge/	ERP / ASD BCI
Zhu et al., 2021	SSVEP 102-subject, 12-target dataset (Figshare)	doi.org/10.6084/m9.figshare.13560281	SSVEP
Pavan et al., 2025	MATLAB scripts + analysis data (OSF)	osf.io/d9eha/	Neurofeedback / ERP
Molina-Cantero et al., 2019	IR-OG and EOG hardware + software (GitHub)	github.com/aljemoca/BlinkingIROG	EOG / blink
Peterson et al., 2020 (b)	Acquisition + processing pipeline (GitHub)	github.com/vpeterson/MI-OpenBCI	Motor imagery

Table A2. Studies providing verified public source-code repositories (n = 15). All URLs confirmed live at time of corpus search (October 2025).

Citation	Description	Repository URL	Paradigm
Brophy et al., 2021	DenoiseEEG-GAN: GAN-based EEG denoising	github.com/Brophy-E/DenoiseEEG-GAN	Artefact removal
Cardona-Alvarez et al., 2023	BCI-Framework: open-source BCI development platform	github.com/UN-GCPDS/BCI-framework	General framework
Ingolfsson et al., 2024/2025	VowelNet: lightweight CNN for EEG vowel decoding	github.com/pulp-bio/VowelNet	Imagined speech
Larsen et al., 2024	Eye-tracking + Unity LSL integration	gitlab.stud.idi.ntnu.no/group-92/eye-tracking-unity-lsl	Game / eye-tracking
Lawhern et al., 2018	EEGNet: compact CNN for EEG-BCI (Python/Keras)	github.com/vlawhern/arl-eegmodels	Cross-paradigm
Legeay et al., 2022	NeuXus: modular BCI pipeline framework (Python/PyPI)	github.com/LaSEEB/NeuXus	General pipeline
Li et al., 2024	Semantic EEG analysis code (MATLAB)	github.com/cfli20/EEG	Semantic decoding
Mattioli et al., 2021	MI-EEG-ID-CNN: subject identification CNN	github.com/Kubasinska/MI-EEG-ID-CNN	Motor imagery
Molina-Cantero et al., 2019	IR-OG and EOG hardware + signal processing software	github.com/aljemoca/BlinkingIROG	EOG / blink
Parbez et al., 2020	Blink-Fruity: blink-controlled BCI game (Python)	github.com/ranadepto/blinkfruity	Binary / blink
Pavan et al., 2025	MATLAB analysis scripts + data (OSF)	osf.io/d9eha/	Neurofeedback / ERP
Peterson et al., 2020	MI-OpenBCI: acquisition + processing pipeline	github.com/vpeterson/MI-OpenBCI	Motor imagery
Reyes et al., 2021	BMI eyes blinking: blink detection BCI (Python)	github.com/Andresreyesf/brain_machine_interface_eyes_blinking	Binary / blink
Saini et al., 2022	tinymlgen: ML model export for microcontrollers	github.com/eloquentarduino/tinymlgen	Embedded / IoT BCI
Schneider et al., 2020	Q-EEGNet: quantised EEGNet (PULP platform)	github.com/pulp-platform/q-eegnet	Motor imagery

The Impact of Quantisation on Performance of Generative Adversarial Networks

Luka Lukač, Damjan Strnad
University of Maribor
Faculty of Electrical Engineering and Computer Science
Koroška cesta 46, 2000 Maribor, Slovenia
{luka.lukac, damjan.strnad}@um.si

Abstract. *This paper focuses on analyzing the influence of quantisation techniques on performance of generative adversarial networks (GANs) to reduce computational and memory requirements during their deployment. Two common techniques, post-training quantisation (PTQ) and quantisation-aware training (QAT), are incorporated into the adversarial training in multiple precision formats. Experimental results on two datasets from different modalities show that quantised models significantly reduced network and memory load with acceptable reduction of performance. The success of quantisation, however, strongly depends on the choice of quantisation strategy and targeted models.*

Keywords. machine learning, neural network, GAN, quantisation-aware training, post-training quantisation

1 Introduction

Modern generative adversarial networks (GANs) have exhibited remarkable results in sample generation across various domains (Aftabi et al., 2023; Majidpour et al., 2026; P. Wang and Sertel, 2023). Their performance, however, comes at the cost of requiring substantial computational resources and memory (Lin et al., 2021). Different compression schemes have been developed for conventional neural network (NN) architectures to address these challenges, including model pruning (Y. He et al., 2019), low-rank decomposition (Li et al., 2023), knowledge distillation (Qin et al., 2021) and quantisation (Rokh et al., 2023). Among these, quantisation stands out as a particularly effective and widely adopted approach for reducing model size and memory load. However, despite its frequent occurrence and success in other machine learning schemes, its application to GANs remains limited.

The main idea behind quantisation is to represent NN parameters with fewer bits than with standard 32-bit floating-point precision (Yang et al., 2019). This reduction leads to smaller model size and faster inference, enabling deployment on resource-constrained devices. However, naïve quantisation typically results in significant performance degradation (Y. Guo, 2018). Several more sophisticated approaches such as post-training quantisation (PTQ) and quantisation-aware training (QAT) (Hasan, 2024) were proposed to mitigate this issue. Despite the emergence of many effective quantisation methods, they have been rarely studied in GAN training (P. Wang et al., 2019), largely due to notorious instability and difficulty of adversarial training (Wiatrak et al., 2019). Therefore, this study investigates the impact of different quantisation strategies on performance of GANs.

The main contributions of this paper are:

- A study of quantisation techniques and their application on GANs with respect to their specifics for an efficient quantisation procedure.
- An experimental evaluation of PTQ and QAT approaches that demonstrate their effectiveness in quantising GAN models.

The remainder of the paper consists of 4 sections: Section 2 provides the background behind GANs and quantisation techniques, Section 3 describes the application of quantisation on GANs, Section 4 depicts and explains the results of the experimental work while Section 5 concludes the paper and provides cues for future work.

2 Background

This section consists of two parts. In the first part, the basic principles behind model quantisation are explained. After that, in the second part, GANs are shortly reviewed.

2.1 Quantisation

The fundamental idea behind quantisation is to represent a high-precision NN model (typically 32-bit floating point (FP32)) with lower-precision data types such as 16-bit floating point (FP16), 16-bit brain floating point (BF16), 8-bit integer (INT8) or 4-bit integer (INT4). Among these, BF16 is the most recent data type developed by Google Brain (S. Wang and Kanwar, 2019) and has the identical dynamic range to FP32 but

lower precision. Quantisation methods can be split into two main groups (Wei et al., 2024):

- Post-training quantisation (PTQ) where models are quantised after the training phase.
- Quantisation-aware training (QAT) that integrates the quantisation process into model training or fine-tuning.

2.1.1 Post-Training Quantisation

The less complex among the two is PTQ where a model is trained in high precision and is quantised only after the model has been trained to the full extent (Nagel et al., 2021). In the case of conversion to an integer precision data type, an original floating-point value r is rescaled to a quantised integer value q as:

$$q = \mathrm{round}(r/S + Z) \tag{1}$$

where S represents a positive downscaling factor while Z is a shifting zero-point value. After rounding, q is often clamped to the respective integer-type range to prevent overflow.

S and Z are usually determined during a calibration process to ensure maximum quantisation efficiency. To be able to effectively map floating-point values to a limited integer precision, statistics of the activations distributions are gathered. Consequently, the downside of PTQ is the need for a representative calibration dataset, with which the calibration is performed.

There exist several common PTQ algorithms for calculating S and Z (ApX Machine Learning, 2025). The most simple is MinMax quantisation where the minimum and maximum values of a quantised tensor are used directly to calculate S and Z. Despite the simple implementation, MinMax is highly sensitive to outliers. Percentile quantisation technique is often used as a more robust alternative, where values below the lower percentile threshold and above the upper percentile threshold (e.g., 0.1% and 99.9%) are clamped. After that, the threshold values are used in calculation of the quantisation parameters. Percentile quantisation scheme is therefore much more resistant to outliers but introduces saturation error for the clipped values. The final major algorithm is entropy quantisation, where clipping thresholds are iteratively determined via Kullback-Leibler (KL) divergence. This approach often results in best performance of the quantised model but is computationally much more expensive than MinMax and percentile quantisation.

2.1.2 Quantisation-Aware Training

PTQ can, in some cases, significantly degrade the performance of a quantised model (Hubara et al., 2021). Consequently, QAT, which integrates quantisation into the training or fine-tuning phase, was proposed. The full-precision model learns to adapt its parameters during the training in a way that the performance of the quantised model does not plunge. Consequently, the negative impact of QAT can be significantly reduced in comparison with PTQ.

Execution of the entire training process in low-precision arithmetic is unviable due to gradient calculation instability (Micikevicius et al., 2017). Therefore, QAT inserts special layers into the model computation graph during training, which quantise tensors and immediately dequantise them, simulating operations in lower-bit precision. The problem occurs in backward pass, as rounding and flooring operations, occurring in quantisation function, often have zero gradients. To solve this issue, the straight-through estimator (STE) technique is used (Bengio et al., 2013), which bypasses the problematic operations and calculates gradients only on truly differential steps in the training pipeline.

2.2 Generative Adversarial Networks

GANs are popular models that have been extensively used for sample generation purposes (Aggarwal et al., 2021; Creswell et al., 2018; Gui et al., 2023). A vanilla GAN consists of two NNs: the first is a generator, which produces samples from a randomly sampled latent space that resemble specimens from the original data distribution, and the second is a discriminator, which is trained to qualify samples as either real or artificially generated. The both models are trained alternately to the extent where the discriminator cannot distinguish between real and fake samples. A schematic overview of a GAN is shown in Fig.1.

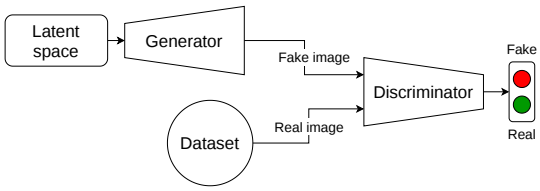


Figure 1. A basic GAN scheme.

Mode collapse is one of the most frequent issues appearing in GAN training (Kossale et al., 2022). Several techniques have approached this challenge in different ways in the past (Metz et al., 2016; Nguyen et al., 2017; Salimans et al., 2016). However, a real breakthrough in tackling mode collapse was the introduction of Wasserstein GANs (WGANs) (Arjovsky et al., 2017). By using the Wasserstein distance instead of divergence losses, the stability of the training was greatly improved while mode collapse was much rarely encountered. Several modern GAN architectures emerged after WGAN, including BigGAN (Brock et al., 2018) and StyleGAN (Karras et al., 2018) that were successfully applied to different fields (Z. Guo et al., 2026; Vanhala and Kurvinen, 2026; X.-Y. Zhang et al., 2025).

3 Method

This section contains three parts. Firstly, the GAN training is explained briefly. After that, in the following two parts, the methodology behind PTQ and QAT is described.

3.1 Training of a GAN

Let D be a discriminator and let G be a generator of a GAN, which is trained for observing quantisation influence on its performance. To enable control over generated samples, a conditional GAN (cGAN) (Mirza and Osindero, 2014) is employed. Discriminator loss function $\mathcal{L}_D$ is based on a modern hinge loss (Kavalerov et al., 2021) and is structured as:

$$\mathcal{L}_D = \mathbb{E}_{(x,y)\sim\mathbb{P}_r}\left[\max(0, 1 - D(x,y))\right] + \mathbb{E}_{z\sim\mathcal{N}(0,I),\, y\sim\mathbb{P}_r}\left[\max\left(0, 1 + D(G(z|y)|y)\right)\right] \tag{2}$$

where x represents a real data sample from the data distribution $\mathbb{P}_r$, z represents a latent vector that is sampled from the standard Gaussian distribution $\mathcal{N}(0, I)$, y represents a class label and $\sim$ represents a sampling operator. To impose 1-Lipschitz constraint inside the critic, spectral normalisation is used (Z. Zhang et al., 2020).

On the other hand, the generator loss function $\mathcal{L}_G$ is defined as:

$$\mathcal{L}_G = -\mathbb{E}_{z\sim\mathcal{N}(0,I),\, y\sim\mathbb{P}_r}\left[D(G(z|y)|y)\right] + \lambda_{div}\mathcal{L}_{div} \tag{3}$$

where λ_{div} is a scaling factor and $\mathcal{L}_{div}$ is a diversity term defined as:

$$\mathcal{L}_{div} = -\mathbb{E}\left[\frac{\|G(z_1, y) - G(z_2, y)\|_1}{\|z_1 - z_2\|_1 + \epsilon}\right] \tag{4}$$

where z_1 and z_2 are different latent vectors and ϵ is a term that prevents dividing by 0.

Algorithmically, the GAN training loop is presented in Algorithm 1. The training is executed in a classical way for N iterations (Line 7), where D and G are trained in an adversarial fashion. In Lines 10–12, D is trained with real and generated data while in Lines 14–16, G is trained with generated data only. The result of the training are the trained G and D in Line 18.

Algorithm 1 Training loop of the GAN.

1: **function** GAN-TRAIN(G, D, N, B, η_G, η_C)
2: ▷ N: number of iterations
3: ▷ B: batch size
4: ▷ η_G, η_C: generator and critic learning rate
5: ▷ λ_{div}: diversity loss factor
6:
7: **for** N iterations **do**
8: $x, y \leftarrow$ LoadRealSamples(B)
9:
10: $z \leftarrow$ SampleLatentVectors(B)
11: $\mathcal{L}_D \leftarrow$ DiscLoss(x, z, y) ▷ Eq. 2
12: $D \leftarrow$ Backpropagation($\mathcal{L}_D$, η_C)
13:
14: $z \leftarrow$ SampleLatentVectors(B)
15: $\mathcal{L}_G \leftarrow$ GeneratorLoss(z, y, λ_{div}) ▷ Eq. 3
16: $G \leftarrow$ Backpropagation($\mathcal{L}_G$, η_G)
17: **end for**
18: **return** G, D
19: **end function**

3.2 Post-Training Quantisation

In the PTQ approach, both G and D are trained in full floating-point precision. After the training is complete, only G is quantised as there are no use cases for D in the inference phase. In the case of quantisation to integer-type format, more sensitive activations are calibrated with entropy quantisation while weights are calibrated per channel with a MinMax approach. The advantage in GAN quantisation is that there is no need for a specific calibration dataset as additional samples can be easily generated with G.

3.3 Quantisation-Aware Training

During QAT, G and D are also trained in full precision. However, fake quantisation modules are inserted to one or both NNs. Tensors are then quantised and immediately dequantised during a forward pass, simulating a realistic quantisation error of lower-precision data types. After the training is done, G is converted to the target precision and is used for inference purposes.

4 Results

The main goal of the experiments was to analyse the impact of quantisation on performance of GANs. Two datasets from different modalities were used: a vector dataset HWCD (Lukač, 2026) and a raster image dataset CIFAR-10 (Krizhevsky et al., 2009). The experiments were conducted on a personal computer with Intel i9-12900K CPU, NVIDIA RTX 3080 GPU and 64 GB of RAM on Windows 11 Education OS.

Both G and D were designed as conditional residual networks (K. He et al., 2016) where G progressively increases sample resolution while on the other

hand D progressively decreases it. The HWCD generator consisted of 4 residual blocks and an output 1D convolutional layer with linear upsampling between them while its discriminator consisted of three fully-connected layers. On the other hand, the CIFAR-10 generator consisted of three residual blocks, self-attention block and an output 2D convolutional head. The CIFAR-10 discriminator was designed with three downsampling residual blocks and a linear output layer. Adam optimizer was used in all experimental settings with betas 0.0 and 0.9. The training hyperparameters are collected in Table 1.

Table 1. Training hyperparameters.

	HWCD	CIFAR-10
Number of iterations	100,000	
Batch size	512	128
Latent vector size	100	128
Generator learning rate	0.002	0.001
Discriminator learning rate	0.002	0.001
Label embedding size	32	
Diversity term factor	0.5	

Results of quantisation on the vector dataset HWCD are collected in Table 2. FP32 represents the baseline model while FP16, BF16 and INT8 represent quantised models. QAT was performed on G only, D only and on both G and D, while PTQ was applied exclusively to G. Although quantising D alone in QAT has no practical use in the inference phase, these experiments were performed for completeness purposes. A pre-trained classifier trained on a subset of HWCD was used for performance estimation on the generated samples with a trained G. It can be seen that the FP32 baseline yielded the best results, but quantisation to FP16 and BF16 introduced slight deterioration in classfication accuracy. When quantising to INT8, the classification accuracy was on-par if applying quantisation to D only but it plunged more when quantising G or both G and D. Furthermore, QAT proved to be superior to PTQ in most of the test scenarios, which can be attributed to the training adapted to quantified activations and weights.

Table 2. Classification accuracy of generated HWCD samples in regards to different quantisation and casting setups.

	QAT			PTQ
	G	D	G+D	G
FP32	97.93%			
FP16	97.88%	97.66%	96.56%	97.07%
BF16	97.74%	97.51%	97.81%	96.93%
INT8	93.63%	97.89%	88.76%	2.55%

Table 3 collects the quantisation results on the standard benchmark raster image dataset CIFAR-10. Unlike at HWCD, the quantisation performance was measured with the Fréchet inception distance (FID) (Heusel et al., 2017), which indicates better results in the case of lower values. As expected, FP32 yielded the best results. Larger differences between quantisation approaches emerged: the results were significantly better when only D is quantised than when G or both G and D underwent the QAT. Quantised G to INT8 resulted in unstable training, the result of which was modest performance. When quantising to FP16 or BF16, PTQ appeared to be superior to QAT, mainly due to the more stable training in FP32.

Table 3. FID between real and generated CIFAR-10 samples in regards to different quantisation and casting setups.

	QAT			PTQ
	G	D	G+D	G
FP32	19.55			
FP16	29.71	22.27	26.66	20.05
BF16	29.54	22.11	26.73	20.31
INT8	81.32	23.10	121.15	360.89

Table 4 summarises the G size under different numerical precisions for the HWCD and CIFAR-10 datasets. Reducing the numerical precision leads to decreased model sizes (50% for FP16 and BF16, and 75% for INT8). This indicates that quantisation can significantly shrink the G size during the inference phase without modifying the model architecture. Careful numerical precision selection, however, is crucial to preserve good model performance while reducing memory load.

Table 4. Memory usage of the quantized generator.

	HWCD	CIFAR-10
FP32	12.40 MB	16.53 MB
FP16, BF16	6.20 MB	8.26 MB
INT8	3.10 MB	4.13 MB

Due to lower memory usage and calculations in lower precision, the inference speed is higher and energy consumption is lower. However, it is difficult to directly measure those two effects as they significantly vary across different GPU architectures and underlying operating systems. Therefore, they are not evaluated directly in this work.

5 Conclusion

This paper presents the incorporation of quantisation to GAN models and analyses its impact on their performance. Two common quantisation techniques were

used: PTQ where one or both models were quantised in the post-training phase, and QAT where quantisation was applied already during the training. Both approaches were examined on a vector dataset HWCD and a raster image dataset CIFAR-10. The experiments indicate that quantisation can be highly beneficial for memory requirements and inference speed. However, the quantisation strategy must be selected carefully to obtain negligible deterioration in model performance.

In the future, the performance of quantised models should be tested in practical applications. Deriving from this, a thorough analysis of energy savings would be highly beneficial to observe the trade-off between computational efficiency and performance of GANs. Furthermore, it would be beneficial if the impact of quantisation were observed on other datasets from diferent data modalities.

Acknowledgments

The authors acknowledge the financial support from the Slovenian Research and Innovation Agency as a part of Young Researcher Funding under Grant 0796-59772, and Research Core Funding P2-0041.

References

Aftabi, S. Z., Ahmadi, A., & Farzi, S. (2023). Fraud detection in financial statements using data mining and GAN models. *Expert Syst. Appl.*, *227*, 120144. https://doi.org/10.1016/j.eswa.2023.120144

Aggarwal, A., Mittal, M., & Battineni, G. (2021). Generative adversarial network: An overview of theory and applications. *Int. J. Inf. Manag. Data Insights*, *1*(1), 100004. https://doi.org/10.1016/j.jjimei.2020.100004

ApX Machine Learning. (2025). Common PTQ algorithms. Retrieved May 7, 2026, from https://apxml.com/courses/practical-llm-quantization/chapter-2-post-training-quantization-ptq/common-ptq-algorithms

Arjovsky, M., Chintala, S., & Bottou, L. (2017, January). Wasserstein GAN. https://doi.org/10.48550/arXiv.1701.07875

Bengio, Y., Léonard, N., & Courville, A. (2013). Estimating or propagating gradients through stochastic neurons for conditional computation. https://doi.org/10.48550/arXiv.1308.3432

Brock, A., Donahue, J., & Simonyan, K. (2018). Large scale GAN training for high fidelity natural image synthesis [Version Number: 2]. https://doi.org/10.48550/arXiv.1809.11096

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., & Bharath, A. A. (2018). Generative adversarial networks: An overview. *IEEE Signal Process. Mag.*, *35*(1), 53–65. https://doi.org/10.1109/MSP.2017.2765202

Gui, J., Sun, Z., Wen, Y., Tao, D., & Ye, J. (2023). A review on generative adversarial networks: Algorithms, theory, and applications. *IEEE Trans. Knowl. Data Eng.*, *35*(4), 3313–3332. https://doi.org/10.1109/TKDE.2021.3130191

Guo, Y. (2018). A survey on methods and theories of quantized neural networks. https://doi.org/10.48550/arXiv.1808.04752

Guo, Z., Du, W., Li, C., & Liu, Z. (2026). Generative adversarial networks in health management of rotating machinery: Survey, recent advances and perspectives. *Neurocomputing*, *681*, 133390. https://doi.org/10.1016/j.neucom.2026.133390

Hasan, J. (2024). Optimizing large language models through quantization: A comparative analysis of PTQ and QAT techniques. https://doi.org/10.48550/ARXIV.2411.06084

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778. https://doi.org/10.1109/CVPR.2016.90

He, Y., Liu, P., Wang, Z., Hu, Z., & Yang, Y. (2019). Filter pruning via geometric median for deep convolutional neural networks acceleration. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). GANs trained by a two time-scale update rule converge to a local Nash equilibrium. *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6629–6640.

Hubara, I., Nahshan, Y., Hanani, Y., Banner, R., & Soudry, D. (2021). Accurate post training quantization with small calibration sets. In M. Meila & T. Zhang (Eds.), *Proceedings of the 38th International Conference on Machine Learning* (pp. 4466–4475, Vol. 139). PMLR. https://proceedings.mlr.press/v139/hubara21a.html

Karras, T., Laine, S., & Aila, T. (2018). A style-based generator architecture for generative adversarial networks [Version Number: 3]. https://doi.org/10.48550/arXiv.1812.04948

Kavalerov, I., Czaja, W., & Chellappa, R. (2021). A multi-class hinge loss for conditional GANs. *Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1289–1298. https://doi.org/10.1109/WACV48630.2021.00133

Kossale, Y., Airaj, M., & Darouichi, A. (2022). Mode collapse in generative adversarial networks: An overview. *2022 8th International Conference on Optimization and Applications (ICOA)*, 1–6. https://doi.org/10.1109/ICOA55659.2022.9934291

Krizhevsky, A., Nair, V., & Hinton, G. (2009, April). *Learning multiple layers of features from tiny images* (tech. rep.). Department of Computer Science,

University of Toronto. Toronto, ON, Canada. http://www.cs.toronto.edu/~kriz/cifar.html

Li, H., Wang, Z., Yue, X., Wang, W., Tomiyama, H., & Meng, L. (2023). An architecture-level analysis on deep learning models for low-impact computations. *Artif. Intell. Rev.*, *56*(3), 1971–2010. https://doi.org/10.1007/s10462-022-10221-5

Lin, J., Chen, W.-M., Cai, H., Gan, C., & Han, S. (2021). MCUNetV2: Memory-efficient patch-based inference for tiny deep learning. *NIPS'21: Proceedings of the 35th International Conference on Neural Information Processing Systems*.

Lukač, L. (2026, April). HWCD: HandWritten Character Dataset. https://doi.org/10.5281/ZENODO.19553096

Majidpour, J., Ahmed, H. A., Ahmed, M. H., Jalal, S. I., & Arabi, H. (2026). Applications of GAN models in breast cancer detection: A comprehensive review. *Arch. Computat. Methods Eng.*, *33*(1), 859–915. https://doi.org/10.1007/s11831-025-10323-7

Metz, L., Poole, B., Pfau, D., & Sohl-Dickstein, J. (2016, November). Unrolled generative adversarial networks. https://doi.org/10.48550/arXiv.1611.02163

Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., Ginsburg, B., Houston, M., Kuchaiev, O., Venkatesh, G., & Wu, H. (2017). Mixed precision training. https://doi.org/10.48550/arXiv.1710.03740

Mirza, M., & Osindero, S. (2014, November). Conditional generative adversarial nets. https://doi.org/10.48550/arXiv.1411.1784

Nagel, M., Fournarakis, M., Amjad, R. A., Bondarenko, Y., van Baalen, M., & Blankevoort, T. (2021). A white paper on neural network quantization. https://doi.org/10.48550/arXiv.2106.08295

Nguyen, T., Le, T., Vu, H., & Phung, D. (2017). Dual discriminator generative adversarial nets. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/e60e81c4cbe5171cd654662d9887aec2-Paper.pdf

Qin, D., Bu, J.-J., Liu, Z., Shen, X., Zhou, S., Gu, J.-J., Wang, Z.-H., Wu, L., & Dai, H.-F. (2021). Efficient medical image segmentation based on knowledge distillation. *IEEE Trans. Med. Imaging*, *40*(12), 3820–3831. https://doi.org/10.1109/TMI.2021.3098703

Rokh, B., Azarpeyvand, A., & Khanteymoori, A. (2023). A comprehensive survey on model quantization for deep neural networks in image classification. *ACM Trans. Intell. Syst. Technol.*, *14*(6), 1–50. https://doi.org/10.1145/3623402

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., & Chen, X. (2016). Improved techniques for training GANs. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, & R. Garnett (Eds.), *NIPS'16: Proceedings of the 30th International Conference on Neural Information Processing Systems* (Vol. 29). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2016/file/8a3363abe792db2d8761d6403605aeb7-Paper.pdf

Vanhala, M., & Kurvinen, E. (2026). A generative adversarial network-based methodology in systematic machine design at concept phase for solutions. *Adv. Mech. Eng.*, *18*(2), 16878132261420739. https://doi.org/10.1177/16878132261420739

Wang, P., & Sertel, E. (2023). Multi-frame super-resolution of remote sensing images using attention-based GAN models. *Knowl.-Based Syst.*, *266*, 110387. https://doi.org/10.1016/j.knosys.2023.110387

Wang, P., Wang, D., Ji, Y., Xie, X., Song, H., Liu, X., Lyu, Y., & Xie, Y. (2019). QGAN: Quantized generative adversarial networks. https://doi.org/10.48550/arXiv.1901.08263

Wang, S., & Kanwar, P. (2019, August). BFloat16: The secret to high performance on cloud TPUs. Retrieved May 14, 2026, from https://cloud.google.com/blog/products/ai-machine-learning/bfloat16-the-secret-to-high-performance-on-cloud-tpus

Wei, L., Ma, Z., Yang, C., & Yao, Q. (2024). Advances in the neural network quantization: A comprehensive review. *Appl. Sci.*, *14*(17), 7445. https://doi.org/10.3390/app14177445

Wiatrak, M., Albrecht, S. V., & Nystrom, A. (2019). Stabilizing generative adversarial networks: A survey. https://doi.org/10.48550/arXiv.1910.00927

Yang, J., Shen, X., Xing, J., Tian, X., Li, H., Deng, B., Huang, J., & Hua, X.-s. (2019). Quantization networks. *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7300–7308. https://doi.org/10.1109/CVPR.2019.00748

Zhang, X.-Y., Zhu, Q.-X., Ke, W., He, Y.-L., Zhang, M.-Q., & Xu, Y. (2025). Regression loss-assisted conditional style generative adversarial network for virtual sample generation with small data in soft sensing. *Eng. Appl. Artif. Intell.*, *147*, 110306. https://doi.org/10.1016/j.engappai.2025.110306

Zhang, Z., Zeng, Y., Bai, L., Hu, Y., Wu, M., Wang, S., & Hancock, E. R. (2020). Spectral bounding: Strictly satisfying the 1-Lipschitz property for generative adversarial networks. *Pattern Recognit.*, *105*, 107179. https://doi.org/10.1016/j.patcog.2019.107179

Assessing the importance of knowledge graph formatting in improving LLM answer accuracy using data augmentation

Sven Leo Kupres, Marko Horvat, Marko Delimar
University of Zagreb Faculty of Electrical Engineering and Computing
Department of Applied Computing, Department of Applied Computing, Department of Energy and Power Systems
Unska 3, HR-10000 Zagreb, Croatia
{Sven-Leo.Kupres, Marko.Horvat, Marko.Delimar}@fer.unizg.hr

Abstract. *While Large Language Models (LLMs) offer intuitive natural language interfaces, producing falsehoods limits their reliability. Therefore, this study examines which abstraction level and data format optimize question accuracy and scalability, as well as asserts how Knowledge Graphs (KGs) can improve LLM responses through a neuro-symbolic approach. Using synthetic datasets based on the FOAF ontology, six query attributes per entity were evaluated through in-context prompting and supervised fine-tuning. Results reveal an accuracy-complexity trade-off: KGs improve factual precision but increase token consumption and preprocessing demands. The findings provide general guidelines for data augmentation and balancing symbolic representation with knowledge scalability.*

Keywords. Large language models, Knowledge graphs, Ontologies, Knowledge representation, Data augmentation, Natural language processing

1 Introduction

Despite the remarkable capabilities of Large Language Models (LLMs) in natural language understanding and generation, their inherent limitations, most notably factual inconsistencies and interpretability challenges, impede their reliable deployment in specialized domains (Yang et al., 2024). While LLMs offer natural language interaction, extensive knowledge coverage and generalized reasoning capabilities, their architecture intrinsically lacks mechanisms to entirely prevent the fabrication of false information, leading to persistent hallucination errors (Chen, 2023).

This reliability gap necessitates the exploration of neuro-symbolic AI architectures that combine the generative power of LLMs with the structured precision of knowledge graphs (KGs) (Hogan et al., 2021). Knowledge graphs are particularly effective at representing factual information as well as complex, relational properties between entities. By grounding generative outputs in verifiable sources, KGs offer deterministic reasoning paths and significantly reduce the likelihood of hallucinations by establishing a logic model that can be efficiently traced, providing AI reliability and explainability (Tilwani et al., 2024; Dehal et al., 2025).

Frameworks such as Retrieval-Augmented Generation (RAG) (Zhao et al., 2026) have been instrumental in this integration, helping to keep generation focused by providing external context. However, the efficiency of these augmentation techniques relies heavily on the quality, structure, and formatting of the provided data. Combining knowledge graphs within LLM architectures introduces critical complexities regarding computational overhead, token limit constraints, and the dynamic updating of knowledge bases (Lin et al., 2025). Therefore, it remains important to assess how the specific knowledge representation format of the augmented data affects the model's reasoning capabilities.

This research examines the potential benefits of integrating large language models with knowledge graphs in the context of neuro-symbolic AI to improve data augmentation methods, while addressing the associated increases in token consumption and data preprocessing demands (Pan et al., 2024; Škrlj et al., 2025). Specifically, different levels of abstraction in knowledge representation, coupled with various data formats are explored as well as how they influence the accuracy and scalability of neuro-symbolic LLM architectures. Knowledge representations at various abstraction levels are described, especially those comparing numeric integers versus human-readable full names as entity identifiers. Different knowledge representation formats, including standardized JSON, XML, and RDF schemes based on the FOAF ontology (Kalemi & Martiri, 2011), are tested, as well as unstructured natural language (Choudhary & Reddy, 2023; Dai et al., 2025; Narvekar et al., 2025; Wu & Tsioutsiouliklis, 2024).

By comparing both in-context system prompting and supervised fine-tuning approaches, this systematic exploration aims to bridge the gap between the generative capacity of LLMs and the structured precision of KGs (Lavrinovics et al., 2024; Yang et al., 2025).

To achieve this, the present paper examines the characteristics of augmentation data used in LLM pipelines,

identifying the structural issues that lead to suboptimal representation, and analyzes how such inputs can be optimized to reduce incorrect answers and hallucinations. Our experimental dataset uses a manageable smaller-scale knowledge graph built on the standardized FOAF ontology and includes entities with heterogeneous properties. Importantly, the LLM's ability to deduce multi-hop relational connections that a human would intuitively make is assessed, determining the extent to which the specific data format influences the model's reasoning capability. Finally, augmentation of the model's context is directly compared to supervised fine-tuning over identical data, providing a comprehensive discussion of the operational benefits, drawbacks, and computational trade-offs of each method.

2 Related Work

Recently published research extensively documents the contradictory nature of LLMs: while they have remarkable natural language capabilities, they suffer from persistent factual inconsistencies. Interpretability challenges as the primary barriers to deploying LLMs in specialized domains were identified, emphasizing that fact-aware language modeling is required (Yang et al., 2024). Furthermore, it was emphasized that current generative architectures do not have the mechanisms required to prevent hallucination errors (Chen, 2023). This limitation is most visible in the LLMs inability to reliably recall long-term or long-tail knowledge, an area where generative models consistently fail without explicit external grounding (Kandpal et al., 2023). As a result, improving factuality through external verification has become an important priority in developing trustworthy large knowledge models.

To bridge this reliability gap, researchers are increasingly turning to neuro-symbolic AI, which combines neural generation and structured symbolic logic. The mutually beneficial relationship between LLMs and knowledge graphs was investigated, demonstrating the effectiveness of structured conceptualization in supporting AI results (Jaldi et al. 2025, Dehal et al. 2025). This claim was confirmed by demonstrating how neuro-symbolic approaches provide deterministic, traceable reasoning paths that improve explainability (Tilwani et al., 2024). A roadmap for the implementation of this synergy was proposed, emphasizing the trade-offs between neural flexibility and symbolic rigidity. While KGs reduce hallucinations, their integration requires significant computational complexity and data preprocessing (Pan et al., 2024, Škrlj et al., 2025).

The Retrieval-Augmented Generation paradigm was unequivocally established as the most important framework for addressing knowledge-intensive natural language processing tasks (Lewis et al., 2020). The RAG has rapidly become the accepted operational standard for LLM-KG integration (Arslan et al., 2024; Zhao et al., 2026). While recent advances, such as Active Retrieval Augmented Generation (Jiang et al., 2023), attempt to iteratively query external sources to verify mid-generation tokens, the efficacy of augmenting LLMs is significantly limited by context window limits and the structural formatting of the provided data (Lin et al., 2025).

As a result, the impact of data formats and prompt structures on LLM comprehension has emerged as a distinct and critical research topic. Ayala and Bechard explicitly demonstrate that maintaining structured outputs in RAG pipelines depends extensively on the retrieval format, which has a direct impact on hallucination rates. This dependency is thoroughly supported by some researchers (He et al., 2024), who systematically investigate prompt formatting across plain text, Markdown, JSON, and YAML, revealing that model performance can vary by up to 40 % simply by changing the structural template (Ayala and Bechard, 2024). This research was expanded on how to leverage these formats by providing a comprehensive evaluation of prompt engineering techniques such as RAG and In-Context Learning (ICL), demonstrating that the optimal strategy is highly dependent on the specific LLM architecture and the structure of the target dataset (Son, Won, and Lee, 2025). Also, complex logical reasoning over KGs have been have investigated (Dai et al., 2025, Choudhary & Reddy, 2023) and it was concluded that structured data schemas influence reasoning accuracy (Wu & Tsioutsiouliklis, 2024). Importantly, the effects of input formats on model performance were compared, and a preference for JSON format for both large and small models was demonstrated. Importantly, smaller LLMs, which have more limited capacity and context windows than their larger counterparts, benefit more significantly from structured input (Narvekar et al., 2025). Also published research shows that the FOAF ontology (Kalemi & Martiri, 2011) and other foundational semantic web vocabularies continue to be part of established standards for representing and evaluating these structured relational graphs.

Finally, the field of neuro-symbolic AI is still actively working to determine the best operational method for integrating structured knowledge. Comprehensive surveys to assess the differences between supervised fine-tuning and in-context system prompting methodologies have been conducted (Lavrinovics et al., 2025, Yang et al., 2025). It was shown that both approaches seek to combine the generative capacity of LLMs with the structured precision of KGs, but they have different operational trade-offs with regard to scalability and overhead.

The overview of recent literature shows that optimizing data augmentation requires a careful balance between symbolic representation format and computational scalability, which this study attempts to systematically evaluate

3 Methodology

The following subsections explain in detail the methodology used throughout the research process in order for them to be reproducible. The first subsection explores the synthetic dataset created for the purposes of this research. The second subsection deals with the model and exactly how the data was used to augment it. Finally, the third subsection outlines in detail the evaluation framework employed to evaluate the generated responses, incorporating the necessary qualitative analysis to accurately measure hallucination rates and reasoning efficacy.

3.1 Synthetic Dataset

The utilized dataset models a person entity and their relationships. A person entity has a first name (string), last name (string), age (integer), gender (enumerator string, either discrete value "male" or "female") and a list of friends (list of other people), where "list" refers to the relationship of a person entity to multiple other entities.

For such a conceptual model, several ways were constructed to present the information to the LLM. The simplest format to model information is through natural language. Naturally, English was selected, as it is a logical choice considering most models were trained on predominantly English data and therefore work in conjunction with it. For instance: "Nada Sudar is 19 years old." or "Zoran Goreta is friends with Nada Sudar.", where "Nada Sudar" and "Zoran Goreta" represent person entities, "Nada" and "Zoran" are the respective first names of persons, while "Sudar" and "Goreta" are their respective last names.

This format displays data the same way two people would display it to one another during interpersonal communication, however dry or simple it may be. A more complex approach is to encapsulate these same sentences within a JSON structure. Each natural language sentence would be a "statement", and all od them together are "statements". An example is depicted in Fig. 1.

A different approach is reducing language complexity and treating data as simple properties. A person has properties and those properties have their own values, e.g. the value of a property "age" is an integer larger than zero (excluding newborn children). However, as will be explained later in more detail, during a person's age generation the number generated could not be smaller than six.

To return to a previous example, the person "Nada Sudar", like all people, has the property "age", and the value of the property is e.g. 25. XML and JSON markup languages are utilized with the exact purpose to save data with such formatting (World Wide Web Consortium, 2006; Bassett, 2015). These two languages are popular and easy to use, with many free tools for parsing and building expressions, and therefore both were used throughout this research.

```
"statements": [
  {
    "statement": "Saša Škrtić is 44 years old."
  },
  {
    "statement": "Saša Škrtić is a male."
  },
  {
    "statement": "Saša Škrtić and Zoran Goreta are friends."
  },
  {
    "statement": "Saša Škrtić and Nikola Prelošnjak are
      friends."
  },
  {
    "statement": "Saša Škrtić and Monika Rimac are friends."
  },
  {
    "statement": "Saša Škrtić and Mirjam Devčić are friends."
  },
  {
    "statement": "Saša Škrtić and Antonija Salinović are
      friends."
  },
  {
    "statement": "Saša Škrtić and Cvijeta Svetličić are
      friends."
  },
  {
    "statement": "Duje Ravlić is 18 years old."
  },
  {
    "statement": "Duje Ravlić is a male."
```

Figure 1. Example of the JSON format which encapsulates natural language. One can see that that each separate "statement" is grouped into the list "statements".

Finally, inspired by the Friend-of-a-friend ontology (FOAF) (Kalemi & Martiri, 2011), the standardized RDF/XML format was also used. FOAF is a vocabulary which describes people and their relationships to other concepts, like their friends, their interests and their creations. It also describes other people's properties like age, gender, home location and appearance.

Each data point, or rather a person entity, must have a unique identifier in order to be separated from other people, as well as making it possible to be referenced when talking about a person's friends. All three of the above-listed structured formats (i.e., XML, JSON, and RDF) had two variations each. The first variation used people's full names as IDs. However, this, of course, is not a good practice in real applications since different people can share a full name, but it was made impossible for the purposes of this research. The second variation used integers as identifiers. This was most relevant for listing friends, since the person's identifier would be used as the reference for another friend entity. To go back to the previous example, if "Zoran Goreta" and "Nada Sudar" were friends, "Zoran Goreta"'s "friend" attribute could either reference "Nada Sudar" or simply "1", which ever was the person's unique identifier for the variation. An example of these differences between variations is depicted in Fig. 2.

During data generation, the first step is to decide on the number of data points, or rather, the number of people entities it consists of. Afterwards, for each person entity the data was generated in the following steps, the order not necessarily important:

```
<Humans>
    <Human id="0">
        <name>Saša</name>
        <surname>Škrtić</surname>
        <age>44</age>
        <gender>MALE</gender>
        <friends>
            <friend ref="3"/>
            <friend ref="9"/>
            <friend ref="11"/>
            <friend ref="14"/>
            <friend ref="16"/>
            <friend ref="17"/>
        </friends>
    </Human>
    <Human id="1">
        <name>Duje</name>
        <surname>Ravlić</surname>
        <age>18</age>
        <gender>MALE</gender>
        <friends>
            <friend ref="2"/>
            <friend ref="5"/>
            <friend ref="7"/>
            <friend ref="8"/>
            <friend ref="9"/>
            <friend ref="14"/>
            <friend ref="15"/>
            <friend ref="17"/>
        </friends>
    </Human>
```

```
<Humans>
    <Human id="Saša Škrtić">
        <name>Saša</name>
        <surname>Škrtić</surname>
        <age>44</age>
        <gender>MALE</gender>
        <friends>
            <friend ref="Nikola Prelošnjak"/>
            <friend ref="Cvijeta Svetličić"/>
            <friend ref="Monika Rimac"/>
            <friend ref="Zoran Goreta"/>
            <friend ref="Antonija Salinović"/>
            <friend ref="Mirjam Devčić"/>
        </friends>
    </Human>
    <Human id="Duje Ravlić">
        <name>Duje</name>
        <surname>Ravlić</surname>
        <age>18</age>
        <gender>MALE</gender>
        <friends>
            <friend ref="Lea Grubić"/>
            <friend ref="Nikola Prelošnjak"/>
            <friend ref="Gordan Lucić"/>
            <friend ref="Vladimir Svalina"/>
            <friend ref="Mirjam Devčić"/>
            <friend ref="Eduard Tutić"/>
            <friend ref="Tomislav Šoštarić"/>
            <friend ref="Cvijeta Svetličić"/>
        </friends>
    </Human>
```

Figure 2. Comparison of the XML format variations. Top image: numeric IDs; bottom image: full names as IDs.

1. People's first and last names were created using an online generator.
2. People's age was sampled from a discrete uniform distribution *Unif(6, 70)*. It was selected arbitrarily because the expectation is that a distribution with a smaller or larger interval would have little impact on the results.
3. People's gender was decided before name generation, so as to have an equal amount of male and female people.
4. People's friends were generated by sampling from a continuous uniform distribution $U_{[\frac{1}{4},\frac{1}{3}]}$ of all friend combinations. The distribution was selected in order for the number of friendships in a dataset to not be predictable and so the number of friendships couldn't be too large. "Friendship" is symmetrical - if one person is friends with another, then the opposite is also always true. "Friendship" also isn't transitive - a person entity who has a friend is not also necessarily a friend of all their friends. A person cannot be a friend to themselves.

In conclusion, eight different formats were utilized in the experiment: 1) natural English language , 2) natural English language encapsulated in JSON, 3) XML with numbers as IDs, 4) XML with full names as IDs, 5) JSON with numbers as IDs, 6) JSON with full names as IDs, 7) RDF with numbers as IDs, and 8) RDF with full names as IDs. For each of these eight described formats, four datasets were constructed, each with a different number of people. In order, the datasets described six, ten, fourteen and twenty people to evaluate how scalable each of the formats is.

3.2 Used model and data integration

The model used was Meta's Llama-3.2-3B-Instruct (meta-llama/Llama-3.2-3B-Instruct, 2024). A simple model, it was just small enough to run on the hardware, but has three billion parameters, which is enough to acquire meaningful results. It is a great baseline model to prove a concept which can then be expanded with a more advanced model with more parameters. The model therefore allows the results to be reproducible.

Two methods were attempted in how a model could access information. The first method was inputting the data into the model's system prompt. Henceforth such models will be referred to in the text as "augmented models" because they are augmented by the data in the system prompt.

The second method was fine-tuning the model on the data. Each data format was inputted or fine-tuned and then evaluated independent of each other. The fine-tuning process was done over fifteen epochs, with a learning rate of 5×10^{-4}. In both cases the model's parameter precision was reduced in order to accommodate running the tests. Other model hyperparameters were never manipulated.

The augmented models received an additional system prompt after the data was given: "Answer questions about this data in as few words as possible.". A similar system prompt was also given to the fine-tuned models: "Answer questions in as few words as possible."

The main point of using both methods is to compare the complexity of the two methods and their respective strengths and weaknesses. The methods differ significantly, with augmentation being relatively simple to accomplish and automate, and fine-tuning requiring more time to perform successfully. Not as much time was dedicated to perform optimal fine-tuning, as the hyperparameters were selected by only applying general good practice rules. This gives an insight into the trade-offs of each method.

3.3 Evaluation process

Each format was evaluated on the same set of questions. For each person in the dataset, six questions were asked:

- What is [full_name]'s first name?
- What is [full_name]'s last name?
- What is [full_name]'s age?
- What is [full_name]'s gender?
- List all [full_name]'s friends.
- How many friends does [full_name] have?

The model's full answer was recorded and stored for subsequent analysis. The answers were human-evaluated since in some instances the model runs out of tokens and cuts the answer short or adds irrelevant information after the question was answered, thus making automatic evaluation unreliable.

Model's response would be classified in one of three categories: "true", "false" or "half true". A "true" answer could contain additional irrelevant information as long as it was not false or contradictory. A "half true" answer could mean multiple things, but it usually meant an incomplete answer. For instance, the model could run out of tokens while generating the list of friends, but all the names it generated were correct. Otherwise, "half true" or "partially true" answers could be ones that provided technically correct information which likely resulted from a lack of specificity in the system prompt, e.g. the model was not explicitly instructed to present the list of friends as their full names, so it returned a list of numbers representing the people's corresponding IDs. However, if the model proceeds to simply output the same data which it was given (JSON or XML structure for example), that is impractical, illegible, and undesired for a human and therefore considered false. Any answer that could be interpreted as misleading was marked as "false", for instance if a person has no friends and a model responds that it does not possess any information about that person's friends. An important part of the evaluation is that a "half true" answer is not quantifiable as 50% of a correct or incorrect answer, but more so an answer that can not be considered either "true" or "false" and should be viewed as its own independent category.

In conclusion, the first set of experiments consists of four separate datasets: 6-, 10-, 14- and 20-person entity dataset. Each dataset evaluation was performed with all eight described formats. The LLM was asked six questions about each person in the dataset. Equation 1 calculates the total number of data points after evaluation for the first set of experiments.

$$8 * 6 * (6 + 10 + 14 + 20) = 2400 \tag{1}$$

Equation 2 calculates the total number of data points for the second set of experiments. Only three of the four experimental datasets were utilized, with the 20-person entity dataset left out. Additionally, only 7 formats were evaluated, with the format "natural.json" left out.

$$7 * 6 * (6 + 10 + 14) = 1260 \tag{2}$$

4 Results

The following subsections showcase the results. Subsection 4.1 examines into how many tokens each data format has been vectorized and establishes whether a direct link exists between how many tokens a format contains and its success rate. Subsection 4.2 compares fine-tuning and augmenting the model on the same data, the strengths and weaknesses of each approach. The following formats deal only with the augmented format; subsection 4.3 explores the differences between formats, mainly establishing a difference between the formats using natural language and structured text. The final subsection 4.4 looks into the variations of structured formats and how slight changes can affect results.

4.1 Token number per format

A potential difference between models that must first be ruled out is if the model's success simply depends on the number of tokens the format is encoded into. In Table 1 one can see how many tokens each of the formats was encoded into. Logically, the more people are present in the format, the more tokens are required. This growth is not purely linear due to how the friends list was constructed, as described in Subsection 3.1.

While the formats comprised of natural language sentences "natural.txt" and "natural.json" start with the least tokens when there are only six people, that amount grows at a quicker rate than the other "structured" formats.

Fig. 3 shows the corelation between the accuracy of the augmented model and the number of tokens the format is encoded into. While it is obvious that each individual format trends downwards, comparing formats leads to a conclusion that data structure does in fact matter and the accuracy is not simply completely dependent on the number of tokens, although it likely plays a role considering more tokens means the LLM has more context to process.

Table 1. Number of tokens for each combination of formats and data points (number of people).

	6	10	14	20
natural.txt	261	599	1055	2272
natural.json	486	1136	2000	4201
foaf.xml	464	845	1297	2224
foaf_names.xml	519	993	1578	2885
foaf.json	496	896	1365	2319
foaf_names.json	551	1044	1646	2980
foaf.rdf	754	1369	2111	3670
foaf_names.rdf	807	1512	2375	4308

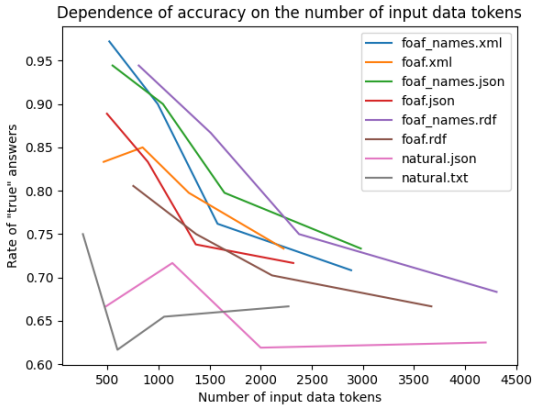


Figure 3. Correlation between the accuracy of the augmented model and the number of tokens used for data encoding.

Table 2. Augmented model accuracy for each combination of formats and number of people entities.

	6	10	14	20
natural.txt	75.00	61.67	65.48	66.67
natural.json	66.66	71.67	61.90	62.50
foaf.xml	83.33	85.00	79.76	73.33
foaf_names.xml	97.22	90.00	76.19	70.83
foaf.json	88.89	83.33	73.81	71.67
foaf_names.json	94.44	90.00	79.76	73.33
foaf.rdf	80.56	75.00	70.24	66.67
foaf_names.rdf	94.44	86.67	75.00	68.33

4.2 Difference between fine-tuned and augmented models

Fine-tuned models had an overall worse accuracy than augmented models, as presented in Table 2 and Table 3. In theory such a result is expected since fine-tuning the model requires it to fit in the new data with its already existing knowledge. The training hyper-parameters were likely not properly selected as well. However, that is a case against fine-tuning models for such tasks since it requires more work to find optimal hyper-parameters and achieve comparable results (Hoffmann et al., 2022). The model in this case was likely under-fitted. If the model were over-fitted on the data, it would "forget" the data it was pretrained on, making it only useful for a specific dataset and losing its ability to generalize. Additionally, the process would have to repeat every time the data was updated, making its scalability poor. However it is worth noting that the execution time for fine-tuned models was significantly faster since the model did not have to process the large system prompt with each query. When measured, the time it took the augmented model to execute all the questions for the "foaf.xml" format, dataset of 10 people, was 302.5 seconds. In comparison, the measured time for the same format when executed by the fine-tuned model was 19.9 seconds, and the other execution times are comparable. This execution time will of course stay largely the same as the dataset grows, while being able to execute augmented model queries with larger datasets requires chunking the data. The experiment execution times for datasets with 6, 10, 14 and 20 people on the "foaf.xml" format are respectively: 121.8, 302.5, 392.9 and 750.0 seconds. Of course, the number of queries per experiment also increases, but even accounting for that, the times per question are: 3.38, 5.04, 4.68, 6.25 seconds.

Table 3. Fine-tuned model accuracy for each combination of formats and number of people entities.

	6	10	14
natural.txt	50.00	43.33	45.24
natural.json	50.00	46.67	46.43
foaf_names.xml	50.00	48.33	48.81
foaf_names.json	50.00	48.33	46.43
foaf_names.rdf	50.00	43.33	46.43

4.3 Difference between formats for augmented models

Per Table 2, there is a noticeable difference in accuracy between "natural.txt" and "natural.json", formats which use natural language, and the more structured formats. The problem arises from the fact that the natural language format accuracy is not as heavily impacted by the increase in dataset size, which can also be seen in Fig. 3. However, the structured format accuracy is still overall superior, as well as the data token amount (which means the computing speed is also faster).

When considering the formats "foaf.json", "foaf.xml" and "foaf.rdf", the first two formats are quite similar when it comes to accuracy, but "foaf.xml" scales better in terms of token amount, as per the Table 1. "foaf.rdf" format is by far the worst of these three - the smallest dataset has the highest amount of tokens of all the formats and scales faster then any other format (besides "foaf_names.rdf", but that is a topic for later). Additionally, it displays poor

accuracy compared to "foaf.json" and "foaf.xml". Even considering the format sees standard use for describing graph data and the model's pretraining data could have included such a format, the results show that building a proprietary lighter-weight framework leads to improved performance.

4.4 Difference between the two format variations

As described in Subsection 3.1, each person is assigned a unique identifier. Changing it from a unique number to the person's full name increases the size of the text and therefore the token amount: 11.85% for the "xml" format, 11.09% for the "json" format and 7.03% for the "rdf" format. This also means using numbers is more scalable, as with a dataset size increase identifiers would have to be used more. However, using names as identifiers is more legible for humans and as it turns out - somewhat for LLM-s as well. Using names as identifiers leads to an improvement in accuracy for all three aforementioned formats (with a caveat being a dataset of 14 and 20 in case of the "foaf_names" format, which defies explanation as of yet). The improvement in accuracy can actually be explained very well by the answers the model gave to a specific question: "List all [full_name]'s friends". The model would usually return the answers using the same format that it found them in - it would normally return full names when using full names as identifiers and return numbers when using numbers as identifiers. The validity of such answers is normally up to the user, but the grading process for this paper was explained in Subsection 3.3.

According to Table 4, taking "half true" answers into account makes the difference between format variations less noticeable and in some cases numbers even outperform full names as identifiers. Additionally, the output format problem could be solved or at least mitigated by improving the system prompt.

Table 4. Augmented model accuracy when regarding "half true" answers as "true".

	6	10	14	20
natural.txt	80.56	66.67	73.81	73.34
natural.json	69.45	78.34	67.85	72.50
foaf.xml	83.33	93.33	83.33	84.16
foaf_names.xml	97.22	96.67	85.71	82.50
foaf.json	88.89	95.00	80.95	80.00
foaf_names.json	94.44	96.67	88.09	84.16
foaf.rdf	80.56	80.00	77.38	71.67
foaf_names.rdf	94.44	90.00	83.33	80.83

Fig. 4 shows the difference between the format variations in a more granular view, on the level of question types. The major difference between variations occurs for the question "List all [full_name]'s friends", but as per Table 4, the difference is lessened when also taking into account "half true" answers, colored green in the figures. The largest source of mistakes are the questions relating to a person entity's friends, which makes sense considering it is the most complex property. When retrieving a list, the model does not need to simply retrieve a single word or number, but multiple elements, which requires generation of a larger number of tokens and therefore has more chances to make a mistake. Giving the number of friends is also a highly complex task since it required the model to count how many specific elements are present.

5 Discussion and conclusion

Table 3 shows the results of the fine-tuned models. The results are overall significantly worse than the augmented model, as per Table 2. However, the fine-tuned model does not display the same drastic accuracy drop-off with an increase in dataset size. This could be explained by the model having more new data to interpret, and so being able to better focus on it, instead of it being lost in the noise of its pretraining data (Hoffmann et al., 2022). However with a substantial increase in dataset size, the same problem as stated in Subsection 4.2 is likely - the model loses its ability to generalize (Luo et al., 2025). The results point to the fact that fine-tuning, while more complex to perform correctly, could lead to better results. However, it is significantly less flexible than simply augmenting the model, which can lead to positive accuracy results with less time dedicated to achieving it, as well as it being simple to change the data the model has access to. It is important to note that the fine-tuning accuracy results by themselves are not indicative of the quality of the data integration method, as they require proper context.

However, the main goal of this paper was to discover how much of an impact the data format presented to an LLM's ability to recall information. When comparing natural language and more structured data, there is a clear difference: structured data outperforms natural language in terms of knowledge accuracy, data scalability and execution speed.

Since there are obvious nuances between the XML, JSON and RDF formats, however slight, further research needs to be done as to the cause of the discrepancy between their results before concluding an objectively superior format among them and if any of the formats considered in this study even are the best. Certain data might prefer to be presented in a specific format based on its qualities, like its size, complexity, domain of knowledge and similar. Another possibility is that the utilized LLM and its base training data might be making a difference - if the training data includes more JSON than XML structure, it would make sense that the LLM would be better at interpreting it (Pan et al., 2026), and vice versa. On a dataset that is as easily chunkable as the one used throughout this research, the

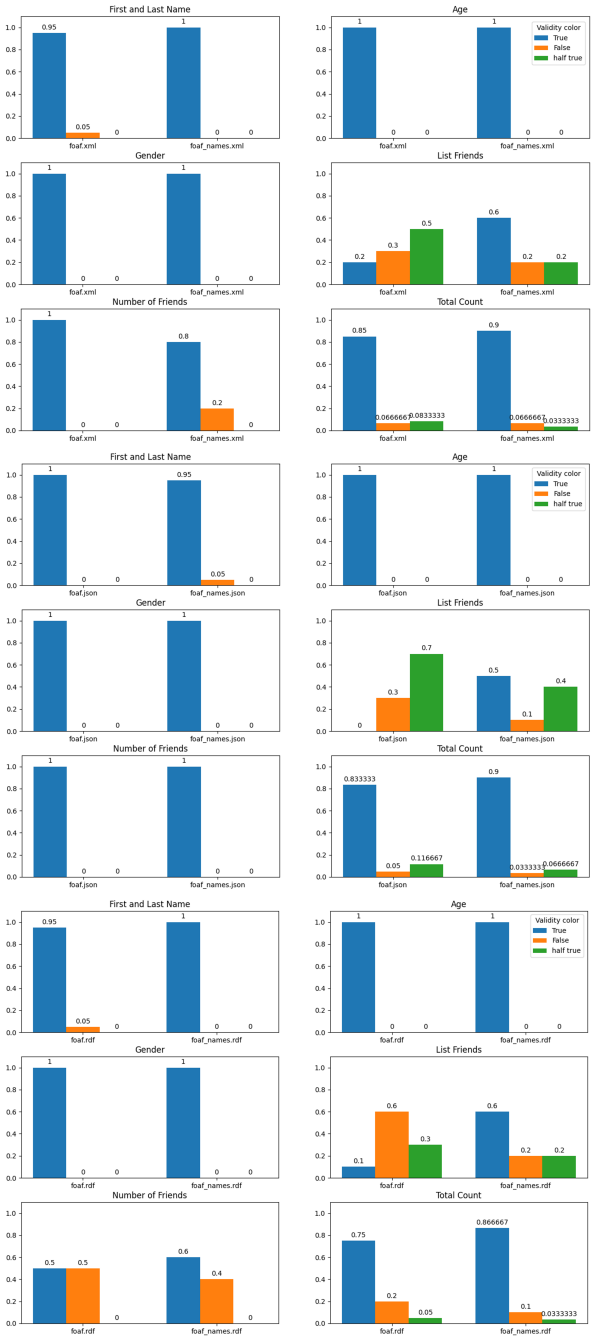


Figure 4. Accuracy of the formats on the 10 person dataset divided by questions. First and second name questions are grouped in the top left bar chart. The bottom right bar chart depicts the average accuracy across all questions. Top image: "foaf.xml" and "foaf_names.xml"; middle image: "foaf.json" and "foaf_names.json"; bottom image: "foaf.rdf" and "foaf_names.rdf".

best approach in order to produce the highest model accuracy seems to be a structured JSON or XML format with minimally complex annotations and label names, as well as using human-readable metadata instead of abstract terms like numbers, which in this case were people entities' IDs.

Furthermore, it is highly important to construct a format which can be easily chunked to avoid large execution times without denying the model proper context. That, however, would need to include a data retrieval mechanism to aid the augmented model in retrieving the required information, which means implementing a system similar to RAG. More research needs to done to conclude if such a retrieval mechanism can also benefit from improved data formatting - would the embedding vectors of the data be easier to look up if the embedding were, for example, formatted as JSON instead of natural language? Does the format used during retrieval have to match the format used during augmentation?

Finally, future research will expand this evaluation to include larger language models and real-world knowledge graph datasets. Evaluating these neuro-symbolic integrations with different LLM architectures and knowledge formats should clarify how model size and context capacity influence LLMs' ability to interpret complex structured data.

Acknowledgments

This paper is based on the research developed through the project "LIVE Quantum – Development of an Integrated AI Platform for Multi-channel Personalized Management of User Requests" supported by the European Union within the Operational Programme Competitiveness and Cohesion in the financial period 2021–2027 (PK.1.1.12.0099).

References

Arslan, M., Ghanem, H., Munawar, S., & Cruz, C. A Survey on RAG with LLMs In *Procedia Computer Science Volume 246, 2024* (pp. 3781-3790)

Ayala, O., Bechard, P. Reducing hallucination in structured outputs via Retrieval-Augmented Generation in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)* (pp. 228-238)

Bassett, L. (2015). Introduction to JavaScript object notation: a to-the-point guide to JSON. " O'Reilly Media, Inc.".

Chen, H. (2023). Large knowledge model: Perspectives and challenges. arXiv preprint arXiv:2312.02706.

Choudhary, N., & Reddy, C. K. (2023). Complex logical reasoning over knowledge graphs using large language models. arXiv preprint arXiv:2305.01157.

Dai, X., Hua, Y., Wu, T., Sheng, Y., Ji, Q., & Qi, G. (2025). Large language models can better understand knowledge graphs than we thought. Knowledge-Based Systems, 312, 113060.

Dehal, R. S., Sharma, M., & Rajabi, E. (2025). Knowledge graphs and their reciprocal relationship with large language models. Machine Learning and Knowledge Extraction, 7(2), 38.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., ... & Sifre, L. (2022). An empirical analysis of compute-optimal large language model training. Advances in neural information processing systems, 35, 30016-30030.

Hogan, A., Blomqvist, E., Cochez, M., dâAmato, C., Melo, G. D., Gutierrez, C., ... & Zimmermann, A. (2021). Knowledge graphs. ACM Computing Surveys (Csur), 54(4), 1-37.

Jaldi, C. D., Saini, A., Ghiasi, E., Divine Eziolise, O., & Shimizu, C. (2025, November). Knowledge conceptualization impacts rag efficacy. In International Knowledge Graph and Semantic Web Conference (pp. 208-223). Cham: Springer Nature Switzerland.

Jiang, Z., Xu, F., Gao L., Sun, Z., Liu Q., Dwivedi-Yu J., Yang Y., Callan J., & Neubig, G. Active Retrieval Augmented Generation In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 7969-7992)

Kalemi, E., & Martiri, E. (2011, November). FOAF-academic ontology: a vocabulary for the academic community. In 2011 Third International Conference on Intelligent Networking and Collaborative Systems (pp. 440-445). IEEE.

Kandpal, N., Deng, H., Roberts, A., Wallace, E., Raffel, & C. (2023). Large Language Models Struggle to Learn Long-Tail Knowledge. Retrieved from https://arxiv.org/abs/2211.08411

Lavrinovics, E., Biswas, R., Bjerva, J., & Hose, K. (2025). Knowledge graphs, large language models, and hallucinations: An nlp perspective. Journal of Web Semantics, 85, 100844.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Article 793, pp. 9459-9474). Curran Associates Inc.

Lin, Y. H., Chen, Q. H., Cheng, Y. J., Zhang, J. R., Liu, Y. H., Hsia, L. Y., & Chen, Y. N. (2025). Llm inference enhanced by external knowledge: A survey. arXiv preprint arXiv:2505.24377.

Luo, Y., Yang, Z., Meng, F., Li, Y., Zhou, J., & Zhang, Y. (2025). An empirical study of catastrophic forgetting in large language models during continual fine-tuning. IEEE Transactions on Audio, Speech and Language Processing.

Model meta-llama/Llama-3.2-3B-Instruct (2024). Retrieved from https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct, retrieved 14 May 2026.

Narvekar, A., Mukherjee, T., & Tanuwijaya, J. (2025, November). Evaluating the Impact of Input Format on LLM Performance: JSON vs. YAML vs. Text. In 2025 IEEE International Conference on Future Machine Learning and Data Science (FMLDS) (pp. 572-577). IEEE.

Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., & Wu, X. (2024). Unifying large language models and knowledge graphs: A roadmap. IEEE Transactions on Knowledge and Data Engineering, 36(7), 3580-3599.

Pan, Z., Wang, S., Pengfei, L., & Li, J. (2026). Understanding llm behaviors via compression: Data generation, knowledge acquisition and scaling laws. Advances in Neural Information Processing Systems, 38, 168404-168454.

Škrlj, B., Koloski, B., Pollak, S., & Lavrač, N. (2025). From Symbolic to Neural and Back: Exploring Knowledge GraphâLarge Language Model Synergies. Challenges and Algorithms for Knowledge Discovery from Data: Essays Dedicated to Arno Siebes on the Occasion of His 67th Birthday, 181-197.

Tilwani, D., Venkataramanan, R., & Sheth, A. P. (2024). Neurosymbolic ai approach to attribution in large language models. IEEE Intelligent Systems, 39(6), 10-17.

Yang, L., Chen, H., Li, Z., Ding, X., & Wu, X. (2024). Give us the facts: Enhancing large language models with knowledge graphs for fact-aware language modeling. IEEE Transactions on Knowledge and Data Engineering, 36(7), 3091-3110.

Yang, W., Some, L., Bain, M., & Kang, B. (2025). A comprehensive survey on integrating large language models with knowledge-based methods. Knowledge-Based Systems, 318, 113503.

World Wide Web Consortium. (2006). Extensible markup language (XML) 1.1.

Wu, X., & Tsioutsiouliklis, K. (2024). Thinking with knowledge graphs: Enhancing LLM reasoning through structured data. arXiv preprint arXiv:2412.10654.

Zhao, P., Zhang, H., Yu, Q., Wang, Z., Geng, Y., Fu, F., ... & Cui, B. (2026). Retrieval-augmented generation for ai-generated content: A survey. Data Science and Engineering, 1-29.

Language-Aware Evaluation of AI-Driven Search: A Benchmark Study of Albanian–English Asymmetry in a Low-Resource Setting

Lazerta Rista
Assistant Lecturer, Faculty of Information Technology,
"Aleksandër Moisiu" University Durrës, Albania
Doctoral Student, University of New York Tirana (UNYT), Tirana, Albania
lazertarista@uamd.edu.al

Abstract. *AI-driven search increasingly mediates information access, yet query language remains underexamined in low-resource settings. AlbSearchBench is a pilot benchmark of 30 matched Albanian–English query pairs tested on Google Search and Gemini across market, institutional, and cultural modules. In Google's top-five results, Albanian queries surfaced Albanian-owned domains more often than English queries (Local Visibility Rate: 0.73 vs 0.49; Holm-adjusted $p < 0.001$). Across both systems, Albanian-language outputs received higher cultural-grounding ratings (4.77 vs 4.13; $p < 0.001$). Intermediary Saturation (IS) and Institutional-module Official Access Rate (OAR) showed directionally consistent patterns, although neither reached statistical significance. These findings support larger language-aware evaluations.*

Keywords. AI-driven search, language asymmetry, low-resource languages, generative AI, visibility allocation, Albanian language

1 Introduction

Information retrieval is shifting from conventional document retrieval toward AI-driven synthesis (Metzler et al., 2021). By producing direct responses rather than only ranked links, generative systems may reduce users' search burden while increasing the mediation between users and underlying information sources. This study examines how that mediation is associated with access to local direct providers, public institutions, and culturally grounded information across languages.

An underexplored question is whether these systems provide equivalent access when the same intent is expressed in different languages. English benefits from substantially greater representation in NLP datasets and technologies, whereas languages such as Albanian remain comparatively under-resourced (Joshi et al., 2020; Kastrati & Biba, 2022). These resource imbalances may affect both the visibility and cultural specificity of information returned across languages.

This study investigates whether query language is associated with differences in procedural access, defined here as the pathway through which a search system directs users to information. Specifically, it examines whether equivalent English and Albanian queries differ in the ownership of surfaced access points, local or international, and in their actor type, direct provider, official source, or intermediary. It also examines whether the cultural fidelity of returned information differs by language. The term procedural asymmetry describes systematic cross-language differences in these routing patterns.

Research Question: To what extent is query language associated with differences in visibility allocation, access pathways, and cultural fidelity in AI-driven search within a low-resource language setting?

The core contributions of this paper are threefold. First, it introduces AlbSearchBench, a matched-query benchmark for evaluating language-conditioned access in AI-driven search across three modules: Market Visibility, Institutional Access, and Cultural/Contextual Grounding. The benchmark combines intent-matched Albanian–English query pairs, a standardized collection protocol, and evaluation across Google Search and Gemini. It further distinguishes between structural visibility analysis based on Google's top-five results and human assessment of outputs from both systems. Second, the study operationalizes procedural access through a reproducible classification framework that codes each surfaced access point along two dimensions: domain ownership, local or international, and actor type, direct provider, official source, or intermediary. These classifications support three visibility metrics: Local Visibility Rate (LVR), Intermediary Saturation (IS), and Official Access Rate (OAR). LVR is treated as the primary visibility measure because it directly captures whether query language is associated with differences in the visibility allocated to Albanian-owned domains. IS describes intermediary-controlled results, while OAR describes access to official sources and is inferentially evaluated within the Institutional module. Third, the study provides pilot-scale empirical evidence that query language is associated with differences in both procedural access and cultural fidelity. Significant differences are observed for local visibility and cultural grounding, whereas intermediary saturation and Institutional-module official access show directionally consistent but statistically non-significant patterns. Together, these contributions establish a reproducible foundation for larger-scale, language-aware audits of visibility allocation, access pathways, and cultural representation in low-resource language settings.

2 Related Work

Research on fairness in natural language processing has increasingly examined how language technologies may reproduce societal biases and generate representational and downstream harms (Bender et al., 2021; Dev et al., 2022). In multilingual settings, these concerns extend beyond model accuracy. Users expressing equivalent information needs in different languages may encounter differences in information quality, source visibility, access to authoritative resources, and the representation of culturally embedded concepts. However, existing evaluation research has largely concentrated on semantic performance, cross-lingual generalization, and retrieval relevance rather than on whether query language changes the ownership and actor type of the sources surfaced to users.

Multilingual NLP benchmarks have substantially broadened the evaluation of cross-lingual model performance. XTREME evaluates multilingual representations across a range of language-understanding tasks, while XTREME-R extends this framework through more challenging tasks and broader cross-lingual evaluation (Hu et al., 2020; Ruder et al., 2021). These benchmarks provide important evidence about whether models can transfer linguistic knowledge across languages. Nevertheless, their principal unit of analysis is model performance on predefined datasets and tasks. They do not directly examine how operational search systems distribute visibility when the same real-world information need is expressed in different languages.

Information-retrieval benchmarks provide a related but distinct evaluation perspective. BEIR supports heterogeneous zero-shot retrieval evaluation across multiple domains, while MTEB evaluates text embeddings across retrieval, reranking, classification, clustering, and other tasks (Thakur et al., 2021; Muennighoff et al., 2023). Multilingual resources such as mMARCO and MIRACL further extend passage-ranking and ad hoc retrieval evaluation across languages (Bonifacio et al., 2021; Zhang et al., 2023). These resources measure retrieval effectiveness, ranking performance, text representation, and cross-lingual transfer. Their settings, however, differ from the present study because they generally evaluate systems against predetermined document collections and relevance judgments. They do not treat domain ownership, direct-provider access, intermediary dependence, official-source visibility, or cultural grounding as primary evaluation outcomes. Albanian is also not consistently represented across these widely used resources, leaving its behaviour in multilingual retrieval and AI-mediated search comparatively underexamined.

A parallel literature on exposure and ranking fairness provides a conceptual basis for studying visibility allocation. Singh and Joachims (2018) formalize fairness in terms of how exposure is distributed through ranked positions. Biega, Gummadi, and Weikum (2018) examine whether attention is allocated equitably among ranked subjects over time, while Diaz et al. (2020) develop exposure-based approaches for evaluating stochastic rankings. This literature demonstrates that ranking systems do more than organize information: they allocate attention among items, individuals, and providers. However, exposure-fairness research generally examines allocation within a single ranking or language condition. It does not directly address whether changing the language of an otherwise equivalent query redistributes visibility between locally and internationally controlled access points. The present study extends this perspective to a matched cross-language setting by examining both domain ownership and actor type.

The growing use of generative systems introduces an additional dimension to information-access evaluation. Conventional search presents identifiable links and domains, whereas generative systems may synthesize information into direct responses and increase the distance between users and the underlying sources. As a result, relevance alone may not capture the quality of the resulting access pathway. Usefulness, source transparency, confidence, and cultural grounding also become important, particularly when information is produced in or about a comparatively under-resourced language. AlbSearchBench therefore distinguishes between structural visibility analysis, based on Google's ranked results, and human evaluation of outputs from both Google Search and Gemini.

Research specifically addressing Albanian has documented the language's structural characteristics and persistent scarcity of digital resources, datasets, and NLP tools (Kastrati & Biba, 2022). More recent work has examined how pre-trained language models perform across multilingual tasks and how explainable-AI methods can be used to interpret their behaviour (Kastrati et al., 2025). Within the Albanian digital ecosystem, preliminary studies have also identified growing use of AI in marketing alongside continuing linguistic challenges for local enterprises, particularly in audience-facing activities such as copywriting and local search optimization (Thanasi Biti & Rista, 2025; Rista & Thanasi Biti, 2026a; Rista & Thanasi Biti, 2026b). These studies establish the linguistic and applied context of the present research, but they do not investigate how live search and generative systems distribute access, visibility, and cultural representation between Albanian and English query conditions.

Taken together, previous research provides established methods for evaluating multilingual model performance, retrieval effectiveness, text representation, and ranking exposure. Limited empirical work, however, examines whether intent-matched queries expressed in different languages systematically alter the ownership and actor type of surfaced access points or the cultural fidelity of returned information. This study addresses that gap by shifting the evaluation focus from semantic performance alone to procedural access, connecting multilingual evaluation with visibility allocation, access pathways, and cultural representation in operational AI-driven search.

3 Methodology

AlbSearchBench evaluates whether expressing the same information need in Albanian or English leads AI-driven search systems to surface different access points and provide differently evaluated outputs. The benchmark examines two dimensions. First, it measures structural access in Google's top-five results by classifying each surfaced domain according to actor type, direct provider, official source, or intermediary, and domain ownership, local or international. Second, it evaluates outputs from Google Search and Gemini through multi-rater assessment of relevance, usefulness, source transparency, cultural grounding, and

confidence. AlbSearchBench is organized into four layers: a Query Bank of intent-matched Albanian–English pairs distributed across the Market Visibility, Institutional Access, and Cultural/Contextual Grounding modules; a Result Classification Layer assigning each Google result an actor type and ownership category; a Metric Layer computing Local Visibility Rate, Intermediary Saturation, and Official Access Rate; and an Evaluation Layer applying human assessment, with inter-rater reliability assessed using the intraclass correlation coefficient.

3.1 Dataset Construction

The pilot dataset comprises 30 matched Albanian–English query pairs, with 10 pairs in each of three modules: Market Visibility, covering searches for local commercial services; Institutional Access, covering civic, governmental, and public-service information; and Cultural/Contextual Grounding, covering culturally embedded concepts and meanings. These modules are referred to hereafter as Market, Institutional, and Cultural, respectively. Each query pair was executed on Google Search and Gemini, producing 120 system–language runs. The 60 Google runs were used to compute the structural visibility metrics, while outputs from both systems were included in the human evaluation.

Queries were formulated as short, natural search expressions representing realistic user information needs rather than artificial keyword strings. Market queries reflected local transactional intentions, such as “Hotel në Berat me parkim” / “Hotel in Berat with parking.” Institutional queries addressed practical procedures and access to authoritative information, such as “Si të regjistroj një biznes në Shqipëri” / “How to register a business in Albania.” Cultural queries examined culturally embedded concepts, such as “Çfarë përfaqëson Kanuni në kulturën shqiptare” / “What does the Kanun represent in Albanian culture.” Within each pair, the Albanian and English formulations were matched as closely as practicable in information need, specificity, phrasing style, and length. The paired design was intended to limit differences arising from query construction, allowing observed variation to be interpreted as associated with the query-language condition.

Data collection was conducted in April 2026 from Durrës, Albania, using controlled incognito browser sessions to reduce the influence of search history, stored cookies, and account-based personalization. The same matched query formulations were used across systems, and collection conditions were kept consistent within each pair. The complete dataset and supporting materials are available in the OSF repository.

3.2 Classification Rule and Metrics

Each of the top-five Google results was assigned exactly one actor type according to a strict access-point ownership rule. Classification was based on the organization controlling the surfaced domain rather than on the relevance or origin of the content displayed on it. A verified website controlled by a public institution was classified as Official, including government portals and verified domains of ministries, municipalities, agencies, and other public authorities. A non-governmental entity’s own website was classified as a Direct Provider. Domains controlled by third-party platforms, directories, aggregators, or hosting services, including booking.com, tripadvisor.com, facebook.com, instagram.com, and youtube.com, were classified as Intermediaries, regardless of whose content appeared on them. When ownership or institutional status was not evident from the domain name, it was verified through the website’s About, contact, institutional, or legal-information pages.

Each result was also assigned one ownership category: local or international. Local denotes an access point controlled by an Albanian organization or institution at the national level; it does not necessarily indicate geographic proximity to the city named in a query. International denotes an access point controlled by an organization based outside Albania. Within each classification axis, the categories were mutually exclusive and collectively exhaustive: every result received one actor-type label and one ownership label.

The rule is deliberately conservative. An Albanian business’s Facebook page or booking.com listing is classified as an international intermediary even when the content concerns a local provider, because the surfaced access point is controlled by a foreign platform. This treatment affects Albanian queries more frequently than English queries: 21 of the 150 Albanian-query results appeared on foreign-controlled platforms, compared with 13 of the 150 English-query results. Accordingly, the ownership-based measure provides a conservative estimate of the Albanian advantage in locally relevant exposure.

Three visibility metrics were calculated for each Google run using the five classified results. Each metric is expressed as a proportion ranging from 0 to 1. Local Visibility Rate (LVR), defined in Eq. (1), is the proportion of the top-five results controlled by Albanian organizations or institutions. It is the primary visibility metric because it directly measures whether Albanian- and English-language queries differ in the visibility allocated to Albanian-controlled access points. Intermediary Saturation (IS), defined in Eq. (2), is the proportion of results controlled by third-party platforms, directories, or aggregators rather than by a provider or official source. Official Access Rate (OAR), defined in Eq. (3), is the proportion of results controlled by a government body, public institution, or official service portal.

$$LVR = (\text{Number of local results of a query}) / 5 \quad (1)$$
$$IS = (\text{Number of intermediary results of a query}) / 5 \quad (2)$$
$$OAR = (\text{Number of official results of a query}) / 5 \quad (3)$$

An LVR of 0.80 indicates that four of the five results are controlled by Albanian actors, whereas an LVR of 0.40 indicates that two are locally controlled. An IS of 0.60 indicates that three of the five results are intermediaries, while an OAR of 0.20 indicates that one result is an official source. A value of 0 means that none of the five results belongs to the relevant category, and a value of 1 means that all five do. Because LVR is derived from the ownership axis, whereas IS and OAR are derived from the actor-type axis, the three metrics are not expected to sum to 1. OAR was calculated descriptively for all Google runs, but its inferential comparison was restricted to the 10 Institutional pairs because official sources are an expected access outcome for civic and public-service queries, not for most Market or Cultural queries.

The visibility metrics were computed only for Google Search because Gemini did not consistently return a fixed ranked list of five source links during the April 2026 collection period. Gemini runs without a complete ranked

source list were excluded from the structural metrics rather than coded as zero. Gemini was instead examined through the human-evaluation results and descriptive source analysis reported in Sections 4.5 and 4.6.

This restriction reflects the observed behaviour of the Gemini 3.1 Flash consumer interface during the collection window. Source links were inconsistent and particularly sparse for general and explanatory queries. A later rerun displayed materially greater citation coverage, indicating temporal variation in source presentation. The findings are therefore interpreted as a snapshot of the system during the recorded period.

3.3 Human Evaluation and Reliability

Three independent raters, a researcher, an SEO specialist, and a business owner, represented analytical, professional, and end-user perspectives. Each independently assessed all 120 system-language runs on five-point scales for relevance, usefulness, source transparency, cultural grounding, and rating confidence. None was an author, and raters could not view one another's scores. The procedure produced 360 ratings per dimension and 1,800 ratings overall.

Predefined criteria guided scoring. Relevance measured whether the output addressed the query; usefulness, whether a user could act on it; source transparency, whether supporting sources were visible and credible; and rating confidence, the evaluator's certainty in the assigned score. Cultural grounding was a composite measure of whether an output reflected the Albanian context through appropriate language, relevant local providers or institutions, and, for cultural queries, accurate cultural meaning. Endpoint anchors ranged from 1 (very poor, absent, or low) to 5 (excellent, clear, or high).

For analysis, the three ratings for each run and dimension were averaged into one run-level score. Inter-rater reliability was assessed with a two-way random-effects, absolute-agreement intraclass correlation model.

Although the three raters were purposively selected to represent distinct evaluative perspectives, analytical, professional, and end-user, rather than randomly sampled from an undifferentiated rater pool, the random-effects specification was retained because the intended target of inference was the reliability of an equivalently constituted evaluation panel rather than agreement among these three specific individuals. ICC(2,3) therefore estimates the reliability of the mean judgment produced by the three-rater panel under this generalization assumption. Because only one rater represented each perspective, however, individual-rater effects cannot be separated from perspective-specific effects, and generalization beyond similarly constituted panels should be made cautiously.

Both single-measure reliability, ICC(2,1), and average-measure reliability, ICC(2,3), were calculated. ICC(2,1) estimates the reliability of one rater, whereas ICC(2,3) estimates the reliability of the mean across all three raters. The coefficients are reported in Section 4.1.

No formal calibration round preceded independent scoring. The strong reliability estimates reported in Section 4.1, particularly for relevance, cultural grounding, and usefulness, indicate substantial convergence under the predefined criteria. Nevertheless, a calibration session is planned for the full-scale study so that shared interpretation can be established prospectively rather than inferred solely from post hoc reliability estimates.

3.4 Statistical Analysis and Interpretation

Matched conditions were compared using the Wilcoxon signed-rank test (Woolson, 2007), which does not require normally distributed measurements. Language comparisons paired the Albanian and English versions of the same query within a system; system comparisons paired Google Search and Gemini for the same query and language. The test ranks non-zero paired differences and evaluates whether they are systematically centered above or below zero. Statistical significance was assessed at $p < 0.05$.

For descriptive language comparisons, differences were calculated as SQ minus EN. Positive differences therefore favour Albanian, while zero differences are ties and are omitted from the signed-rank calculation. Rank-biserial correlation, r, was reported as the effect size. Because the statistical output used the English-minus-Albanian contrast, negative r values favour Albanian. Absolute values below 0.10 were treated as negligible, 0.10-0.29 as small, 0.30-0.49 as medium, and 0.50 or above as large.

The Hodges-Lehmann estimate expresses the typical paired shift in the original units of a metric. For LVR, a shift of 0.20 corresponds to one additional Albanian-controlled result among five. Holm correction was applied to three prespecified structural comparisons: overall LVR, overall IS, and Institutional-module OAR, controlling familywise error (Holm, 1979). Descriptive patterns with $p >= 0.05$ are reported as directionally consistent but statistically non-significant, rather than as evidence of no difference.

Human-evaluation means were calculated from the run-level averages. ICC values were interpreted following Koo and Li (2016): below 0.50, poor; 0.50-0.75, moderate; 0.75-0.90, good; and above 0.90, excellent reliability.

4 Results

The results are presented in six subsections. Inter-rater reliability is reported first, followed by the structural visibility and human-evaluation findings. Two cross-language effects reached statistical significance at pilot scale: Local Visibility Rate and cultural grounding. Intermediary Saturation and Institutional-module Official Access Rate showed directionally consistent but statistically non-significant differences. Cross-system comparisons revealed task-dependent usefulness and lower source transparency for Gemini.

4.1 Inter-rater Reliability

Inter-rater reliability ranged from moderate to excellent for individual raters and from good to excellent for the mean of three raters (Table 1). Average-measure reliability was excellent for relevance, cultural grounding, and usefulness, and good for rating confidence and source transparency.

Table 1. Inter-rater reliability across 120 runs and three raters.

Dimension	ICC(2,1)	ICC(2,3)
Relevance	0.92	0.97
Cultural grounding	0.996	0.999
Usefulness	0.78	0.91
Rating confidence	0.71	0.88
Source transparency	0.58	0.80

The near-perfect cultural-grounding coefficient indicates highly consistent ratings under the applied criteria. It should nevertheless be interpreted cautiously because Albanian-language outputs were concentrated at the upper boundary of the scale, creating a ceiling effect and limiting discrimination among those outputs.

4.2 Local Visibility Rate

Across all 30 matched pairs, Albanian queries produced a substantially higher mean LVR than English queries (0.73 vs 0.49; W=5.0, p<0.001; rank-biserial r=−0.95; Hodges–Lehmann shift=0.20, 95% CI [0.10,0.30]). The difference remained statistically significant after Holm correction across the three prespecified structural comparisons (adjusted p<0.001). Eighteen pairs favoured Albanian, one favoured English, and 11 were tied.

Mean LVR was descriptively higher for Albanian in every module (Table 2). Module-level values are reported to show the consistency of direction across the 10 pairs in each module; inferential testing is reserved for the prespecified overall LVR comparison.

Table 2. Local Visibility Rate by module for Google Search (10 pairs per module; module values descriptive).

Module	SQ	EN
Market	0.78	0.54
Institutional	0.78	0.52
Cultural	0.64	0.40
All 30 pairs	0.73	0.49

A worked example illustrates how query language affected both ownership and actor type. For "Si të aplikoj për pasaportë shqiptare" / "How to apply for an Albanian passport," all five Albanian results were locally controlled, including four official sources. The English results contained two local official sources and three international intermediaries (Table 3).

Table 3. Worked example: top-five Google results for the passport query pair.

Rank	English SERP	Albanian SERP
1	e-albania.al (Official/L)	e-albania.al (Official/L)
2	futurepassports.com (Intermediary/Int)	identitek.al (Direct/L)
3	global-relocate.com (Intermediary/Int)	e-albania.al (Official/L)
4	ambasadat.gov.al (Official/L)	ambasadat.gov.al (Official/L)
5	wikipedia.org (Intermediary/Int)	punetejashtme.gov.al (Official/L)

For this pair, LVR was 1.00 for Albanian and 0.40 for English; IS was 0.00 and 0.60; and OAR was 0.80 and 0.40, respectively. In Table 3, L denotes local ownership and Int denotes international ownership.

4.3 Intermediary Saturation and Official Access Rate

Across all 30 pairs, mean IS was higher for English than Albanian queries (0.57 vs 0.51; W = 36.0, p = 0.10, r = 0.47; Holm-adjusted p = 0.21). The difference was directionally consistent but statistically non-significant.

OAR was evaluated inferentially only within the Institutional module, where official sources are a substantively expected access outcome. Albanian queries produced a higher mean OAR than English queries (0.44 vs 0.30; r = -0.54, p = 0.30; Holm-adjusted p = 0.30), but the difference was not statistically significant. Market and Cultural OAR values were near zero in both languages and are retained descriptively in Table 4 because those query types rarely call for official sources.

Table 4. IS and OAR by module for Google Search; only Institutional OAR is tested inferentially.

Module	IS SQ	IS EN	OAR SQ	OAR EN
Market	0.28	0.40	0.00	0.00
Institutional	0.28	0.44	0.44	0.30
Cultural	0.98	0.86	0.00	0.04
All 30 pairs	0.51	0.57	0.15	0.11

Metric behaviour also reflected module structure. In the Cultural module, almost all results in both language conditions were intermediaries, producing near-ceiling IS values of 0.98 for Albanian and 0.86 for English. LVR was less affected because it derives from the separate ownership axis.

4.4 Cultural Grounding and Conceptual Flattening

The second significant cross-language effect concerned cultural fidelity. Across all 60 matched query-system units, Albanian-language outputs received higher mean cultural-grounding ratings than English-language outputs (4.77 vs 4.13; p < 0.001; r = -0.81).

Within the Cultural module, Albanian outputs averaged 5.00 compared with 3.45 for English, and every matched unit favoured Albanian (r = -1.00). The gap was larger for Google Search than for Gemini (Table 5).

Table 5. Cultural-grounding ratings within the Cultural module (10 pairs per system).

System	SQ	EN
Google Search	5.00	2.90
Gemini	5.00	4.00

Qualitative inspection of the structured issue log illustrates a pattern termed conceptual flattening, understood here as the reduction of culturally specific ambiguity, variation, or context in the English-language condition. For the Zana query, the principal difference concerned characterization rather than translation: the English output more strongly emphasized benevolent and protective fairy-like roles, whereas the Albanian output more explicitly retained the figures' ambivalence, presenting them as beautiful and protective but also potentially dangerous or punitive. A comparable difference appeared in the black-cat query. The English output stated more generally that black cats are commonly considered bad luck in Albania, while the Albanian output qualified the belief as mixed, partly borrowed or urban, and culturally variable. In the traditional-food query, the English output provided a general list of Albanian dishes, desserts, and drinks, whereas the Albanian output connected particular foods with specific holidays and celebrations. These observations show how an

output may remain broadly relevant while losing local variation, symbolic ambiguity, or occasion-specific meaning. They are used to interpret, rather than independently establish, the statistical cultural-grounding result.

4.5 Task-Dependent Usefulness

The comparison between Google Search and Gemini revealed a task-dependent usefulness pattern. Usefulness scores were pooled across the Albanian and English conditions within each module, producing 20 matched query-language units per module. Relevance was consistently high for both systems (Gemini 5.00; Google 4.86), while usefulness differed by module (Table 6). For Market queries, Google was rated more useful (5.00 vs 3.52; $p < 0.001$; $|r| = 1.00$, favouring Google in all 20 matched units), frequently providing actionable elements such as map listings, telephone details, opening hours, and directions.

For Cultural queries, Gemini was rated more useful (4.70 vs 3.60; $p < 0.001$; $|r| = 1.00$, favouring Gemini in all 20 matched units), reflecting the value of synthesized responses for cultural and explanatory information needs. Institutional queries also favoured Google (4.87 vs 4.00; $p < 0.001$; $|r| = 1.00$, favouring Google in all 20 matched units), as these tasks often required direct access to an official portal or service.

Table 6. Mean usefulness ratings by module and system (20 query-language units per module; 60 overall).

Module	Google	Gemini
Market	5.00	3.52
Institutional	4.87	4.00
Cultural	3.60	4.70
All runs	4.49	4.07

Although the overall mean favoured Google, the module-level results reveal a crossover pattern. Google was significantly more useful for actionable Market and Institutional queries, whereas Gemini was significantly more useful for Cultural queries requiring synthesized explanation.

4.6 Source Transparency and Verification Risk

Unlike usefulness, source transparency consistently favoured Google. Gemini received a substantially lower mean rating than Google (1.96 vs 4.07; $p < 0.001$; $|r| = 1.00$, favouring Google in all 60 matched comparisons). Raters also reported lower rating confidence for Gemini outputs than for Google outputs (3.06 vs 4.28; $p < 0.001$; $|r| = 0.89$). Because source transparency showed moderate single-rater reliability (ICC(2,1) = 0.58) but good reliability for the averaged three-rater score (ICC(2,3) = 0.80), this comparison is interpreted using the run-level mean ratings.

The structured issue log records reduced or distorted treatments of culturally specific concepts and English result sets relying on low-specificity sources. These cases do not estimate a general hallucination rate; they show that sparse or weak sourcing increases the verification burden and makes unsupported reduction or distortion harder to detect.

Within Gemini, no visible sources were provided in 23 of 30 Albanian runs, compared with eight of 30 English runs. Google source-transparency ratings were nearly identical across languages (4.08 vs 4.06). A later rerun displayed materially greater citation coverage, so these findings describe Gemini 3.1 Flash during the April 2026 collection window rather than a stable property of later versions.

5 Discussion

The findings show that query language is associated with both the structure of information access and the cultural fidelity of returned information. Albanian queries allocated more visibility to Albanian-controlled domains, whereas English queries more often routed users through internationally controlled access points. Albanian-language outputs also preserved culturally embedded meaning more consistently. These patterns do not imply intentional bias or establish a causal mechanism; they describe systematic associations within a matched pilot dataset.

The practical implications differ by module. In local service search, reduced visibility for Albanian-controlled providers may shift attention and economic value toward international platforms. In institutional search, the Institutional-module OAR comparison suggests that query language may affect how directly users reach authoritative sources, although the difference was not statistically established. For cultural information, conceptual flattening can obscure distinctions that are central to local knowledge. Generative synthesis offered greater usefulness for explanation but imposed a higher verification burden, with the source-omission pattern appearing more pronounced in Albanian during the collection window.

Methodologically, AlbSearchBench separates two analytical layers. The structural layer identifies who controls the surfaced access point through reproducible ownership and actor-type classification. Its strength is consistency; its limitation is that a local provider hosted on a foreign platform is coded as international intermediation. The human-evaluation layer provides an initial assessment of what outputs deliver in terms of relevance, usefulness, transparency, and cultural grounding. Together, the layers extend multilingual evaluation beyond semantic accuracy toward visibility allocation and procedural access.

6 Limitations

Several constraints limit generalization. First, the pilot contains 30 matched pairs, 10 per module, and is designed to establish the framework and detect large effects rather than provide a census of search behaviour. The secondary overall IS and Institutional-module OAR comparisons are limited by the pilot sample and are therefore reported cautiously as directional. Second, the data form a cross-sectional snapshot from April 2026. Gemini 3.1 Flash returned sources inconsistently, and a later rerun showed greater citation coverage; system-level transparency findings are therefore bounded to the collection window. Third, all queries were executed from Durrës, Albania, so language is not fully separated from a localized IP and search environment. Fourth, all queries targeted Albanian providers, institutions, or cultural content. The design therefore cannot fully distinguish an effect of query language itself from an effect of alignment between the query language and the target

content. Future work should include reverse-case queries, such as Albanian-language queries for non-Albanian targets, to separate these mechanisms.

The ownership rule classifies access points rather than page content. This improves reproducibility but treats an Albanian provider's page on a foreign platform as international intermediation. The conceptual-flattening examples are illustrative and were not subjected to an independent qualitative coding scheme. Cultural-grounding scores also showed a ceiling effect for Albanian outputs, while relevance was maximal for one system, limiting discrimination. Because each evaluator perspective was represented by one rater, generalization of the random-effects ICC to comparable evaluators remains provisional. Finally, the five-point criteria used endpoint anchors rather than fully specified descriptors for every scale point; a more granular anchored protocol is planned for subsequent work.

7 Conclusion

AlbSearchBench identifies two significant cross-language asymmetries in this low-resource setting. Albanian queries produced greater visibility for Albanian-controlled domains in Google's top-five results, and Albanian-language outputs received higher cultural-grounding ratings across Google Search and Gemini. Overall Intermediary Saturation and Institutional-module Official Access Rate followed directionally consistent patterns but did not reach statistical significance at pilot scale.

The cross-system findings further show that usefulness depends on the task: Google was significantly more useful for actionable Market and Institutional queries, whereas Gemini was significantly more useful for Cultural explanation. This explanatory value was accompanied by lower source transparency and greater verification difficulty during the recorded Gemini 3.1 Flash collection window. Evaluating AI-driven search through semantic relevance alone would therefore miss how language affects the ownership of visible access points, the route to authoritative information, and the preservation of culturally specific meaning.

The benchmark provides a reproducible basis for larger language-aware audits. Future work will expand the query bank; add reverse-case tests, including Albanian-language queries for non-Albanian targets, to separate query-language effects from language-content alignment; introduce a coded content-level analysis; test additional systems and collection locations; and monitor temporal changes in ranking and source presentation.

References

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21), 610-623.

Biega, A. J., Gummadi, K. P., & Weikum, G. (2018). Equity of attention: Amortizing individual fairness in rankings. Proceedings of SIGIR 2018, 405-414.

Bonifacio, L., Jeronymo, V., Abonizio, H. Q., Campiotti, I., Fadaee, M., Lotufo, R., & Nogueira, R. (2021). mMARCO: A multilingual version of the MS MARCO passage ranking dataset. arXiv preprint arXiv:2108.13897.

Dev, S., Sheng, E., Zhao, J., Amstutz, A., Sun, J., Hou, Y., Sanseverino, M., Kim, J., Nishi, A., Peng, N., & Chang, K.-W. (2022). On measures of biases and harms in NLP. Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022.

Diaz, F., Mitra, B., Ekstrand, M. D., Biega, A. J., & Carterette, B. (2020). Evaluating stochastic rankings with expected exposure. Proceedings of CIKM 2020, 275-284.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics, 6(2), 65-70.

Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., & Johnson, M. (2020). XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. Proceedings of ICML 2020.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. Proceedings of the 58th Annual Meeting of the ACL, 6282-6293.

Kastrati, M., & Biba, M. (2022). Natural language processing for Albanian: A state-of-the-art survey. International Journal of Electrical and Computer Engineering, 12(6), 6432-6439.

Kastrati, M., Imran, A. S., Hashmi, E., Kastrati, Z., Daudpota, S. M., & Biba, M. (2025). Unlocking language barriers: Assessing pre-trained large language models across multilingual tasks and unveiling the black box with explainable artificial intelligence. Engineering Applications of Artificial Intelligence, 149, 110136.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of Chiropractic Medicine, 15(2), 155-163.

Metzler, D., Tay, Y., Bahri, D., & Najork, M. (2021). Rethinking search: Making domain experts out of dilettantes. ACM SIGIR Forum, 55(1), 1-27.

Muennighoff, N., Tazi, N., Magne, L., & Reimers, N. (2023). MTEB: Massive text embedding benchmark. Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, 2014-2037.

Rista, L., & Thanasi Biti, T. (2026a). The strategic integration of artificial intelligence in digital marketing: Global case studies and implications for the Albanian emerging market. Interdisciplinary Journal of Research and Development, 13(1), 177. https://doi.org/10.56345/ijrdv13n120

Rista, L., & Thanasi Biti, T. (2026b). A real-world mixed-methods investigation across service and e-commerce sectors in Albania. Proceedings of the 16th International Scientific Conference of Business Faculty (ISCBF 2026), 419-429.

Ruder, S., Constant, N., Botha, J., Siddhant, A., Firat, O., Fu, J., Liu, P., Hu, J., Garrette, D., Neubig, G., & Johnson, M. (2021). XTREME-R: Towards more challenging and nuanced multilingual evaluation. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 10215-10245.

Singh, A., & Joachims, T. (2018). Fairness of exposure in rankings. Proceedings of KDD 2018, 2219-2228.

Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., & Gurevych, I. (2021). BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1.

Thanasi Biti, T., & Rista, L. (2025). Leveraging AI for personalized digital marketing campaigns in Albania. Interdisciplinary Journal of Research and Development, 12(3 S1), 73. https://doi.org/10.56345/ijrdv12n3s110

Woolson, R. F. (2007). Wilcoxon signed-rank test. Wiley Encyclopedia of Clinical Trials.

Zhang, X., Thakur, N., Ogundepo, O., Kamalloo, E., Alfonso-Hermelo, D., Li, X., Liu, Q., Rezagholizadeh, M., & Lin, J. (2023). MIRACL: A multilingual retrieval dataset covering 18 diverse languages. Transactions of the Association for Computational Linguistics, 11, 1114-1131.

Machine Learning for Subject Classification of Books Using Open Bibliographic Metadata

Livija Ružić
Osnovna škola Prelog
Trg bana Jelačića 2, 40323, Prelog
livija.ruzic@gmail.com

Dino Vlahek
Faculty of Organization and Informatics, University of Zagreb
Department of Theoretical and Applied Foundations of Information Sciences
Pavlinska 2, 42000 Varaždin
dvlahek@foi.hr

Abstract. *This study evaluates automatic subject classification of books using open bibliographic metadata from Open Library. A balanced dataset of 15,000 records was constructed across ten broad subject categories, using only titles and descriptions as predictive input, while original subject labels were used only to derive target classes. Classical TF-IDF-based classifiers were compared with MiniLM sentence embeddings followed by Logistic Regression. The transformer-based representation achieved the best performance, with a macro-F1 score of 0.6173, outperforming the strongest classical baseline. The results indicate that semantic embeddings can support broad subject-category prediction, although expert review remains necessary for ambiguous records.*

Keywords. automatic subject classification, bibliographic metadata, Open Library, machine learning, TF-IDF, transformer embeddings

1 Introduction

Libraries and digital repositories nowadays provide access to large and heterogeneous bibliographic collections. In such collections, metadata play a central role in discovery, retrieval, cataloguing, and user access. Among different types of bibliographic metadata, subject information is particularly important, as it allows users to identify relevant books even when they do not know the exact title, author, or terminology used in the catalogue. Subject classification can therefore be viewed as a key component of knowledge organization, since it connects bibliographic records with controlled vocabularies, classification systems, and user-oriented access points (Golub, 2019; Golub, Hagelbäck, & Ardö, 2021). However, subject assignment is a demanding process. It often requires expert interpretation of bibliographic descriptions, knowledge of cataloguing practice, and the use of controlled subject vocabularies. This becomes increasingly difficult in large-scale digital collections, where records may be incomplete, inconsistently described, or available only through short textual fields. For this reason, automatic subject indexing and automatic subject classification have been studied as computational approaches for supporting metadata creation and enrichment (Golub et al., 2021; Suominen, 2019). These approaches are not intended to replace professional cataloguing, but to provide candidate subject categories that can be reviewed, corrected, or refined by experts (Mödden, 2022). Automatic classification of books is especially challenging because bibliographic records are usually short and semantically ambiguous. A title may contain only a small amount of information, while descriptions differ considerably in length, style, and quality. In addition, subject categories are not always mutually exclusive. For example, a book about a historical figure may be related to both biography and history, while a book about computing may overlap with science, technology, education, or business. Consequently, the task cannot be reduced only to keyword matching, since the same subject area may be expressed using different vocabularies, while different subject areas may share similar terminology.

Earlier approaches to automatic text classification commonly relied on lexical document representations, such as bag-of-words and TF-IDF (Term Frequency–Inverse Document Frequency), combined with classical classifiers (Salton & Buckley, 1988; Sebastiani, 2002). Such methods remain important baselines because they are computationally efficient, relatively transparent, and often effective when discriminative terms are present in the input text (Kowsari et al., 2019). In bibliographic classification, however, their performance depends strongly on the availability and quality of textual metadata. When records are short or

when subject information is expressed indirectly, lexical representations may fail to capture semantic similarity between texts that do not share the same vocabulary (Nordström, 2018). Transformer-based language models provide an alternative representation strategy by learning dense semantic embeddings of text. The transformer architecture introduced attention-based language modelling (Vaswani et al., 2017), while BERT demonstrated the effectiveness of deep bidirectional pre-training for language understanding tasks (Devlin, Chang, Lee, & Toutanova, 2019). Sentence-BERT further adapted transformer models for efficient sentence- and paragraph-level embeddings suitable for similarity comparison and downstream classification (Reimers & Gurevych, 2019). Such representations are relevant for bibliographic metadata because titles and descriptions are often short textual units whose subject meaning may be expressed indirectly. Recent work has also explored BERT-based models for assisted subject indexing, indicating their potential relevance for library metadata workflows (Chou, 2022). Despite these developments, the evaluation of automatic subject classification must be designed carefully. Bibliographic records often contain original subject labels or subject headings, which may be strongly related to the target category. If such fields are used as predictive input, the resulting evaluation may suffer from label leakage and produce overly optimistic results (Kaufman, Rosset, & Perlich, 2012). Therefore, a realistic experimental setting should distinguish between metadata used to define the target variable and metadata used as input to the classifier.

This study investigates whether broad subject categories of books can be predicted from open bibliographic metadata. The study is based on records obtained from the Open Library works data dump (Open Library, 2025). From these records, a balanced dataset of 15,000 books across ten broad subject categories is constructed: Art, Biography, Business, Education, Fiction, History, Medicine, Religion, Science, and Technology. The predictive input is limited to the book title and description, while the original subject labels are used only to construct the target category. The objective of this study is to compare classical lexical representations with transformer-based semantic embeddings for broad book subject classification. For this purpose, TF-IDF features are evaluated using Decision Tree, Random Forest, Linear Support Vector Machine, k-Nearest Neighbors, and a multilayer perceptron classifier. In addition, MiniLM sentence embeddings are combined with Logistic Regression in order to evaluate a compact transformer-based representation under the same classification setting. The contributions of this paper are as follows: (i) A reproducible dataset construction procedure is presented for deriving broad subject categories from open bibliographic metadata while avoiding the direct use of subject labels as predictive features. (ii) Classical TF-IDF-based machine learning models are compared with transformer-based semantic embeddings for book subject classification using the same input fields and target categories. (iii) The obtained results are discussed from the perspective of semi-automatic metadata enrichment, where automatic classification is used as support for expert cataloguing rather than as a fully autonomous replacement for it.

2 Related Work

As the problem addressed in this paper combines bibliographic metadata processing, automatic subject indexing, and supervised text classification, the related work must be considered from several perspectives. First, automatic subject indexing and metadata enrichment in library and information science are reviewed. Classical machine learning approaches for text classification are then discussed, since they represent important and interpretable baselines. Finally, transformer-based text representations are considered as a more recent alternative for capturing semantic relationships in short bibliographic descriptions.

2.1 Automatic Subject Indexing and Bibliographic Metadata

Automatic subject classification has been studied as part of broader efforts to support metadata creation, cataloguing, and knowledge organization in libraries. Traditional cataloguing relies on expert interpretation of bibliographic records and controlled vocabularies. However, the increasing scale of digital and digitized collections has made fully manual subject assignment difficult to maintain, especially when metadata are incomplete, inconsistent, or available at different levels of granularity. In order to address these challenges, automatic subject indexing approaches have been introduced to assist the assignment of subject labels, classification codes, and other descriptive metadata (Golub, 2019; Golub et al., 2021). Practical systems such as Annif further demonstrate that machine-learning-based subject indexing can be implemented as a tool for suggesting subject headings and supporting metadata creation workflows (Suominen, 2019). In this context, automatic classification is not usually considered a replacement for professional cataloguing, but rather as a decision-support mechanism. Machine learning models can suggest candidate subject categories that are then verified, corrected, or refined by domain experts. This is particularly important in bibliographic classification, where subject categories may be semantically overlapping, hierarchically organized, and dependent on cataloguing practice. Similar machine-learning approaches have also been investigated for assigning controlled subject headings to scholarly documents, further indicating the relevance of automatic classification for large-scale bibliographic organization (Kazi, Thompson, Choudhury, & Oard, 2021). Recent work

on artificial intelligence and machine learning in bibliographic control has therefore emphasized the role of such methods in supporting, rather than replacing, expert-based knowledge organization workflows (Mödden, 2022).

2.2 Classical Machine Learning for Text Classification

Earlier approaches to automatic text and bibliographic classification commonly relied on lexical representations of documents. In such approaches, textual fields are transformed into high-dimensional feature vectors using bag-of-words or TF-IDF representations, where terms are weighted according to their importance within a document and across a collection (Salton & Buckley, 1988). These representations are then used together with classical classifiers, such as Naive Bayes, Support Vector Machines, logistic regression, decision trees, and ensemble-based models. Such machine-learning-based text categorization approaches have been widely studied for assigning documents to predefined categories (Sebastiani, 2002). Although these methods are relatively simple, they remain important baselines because they are computationally efficient, robust on small and medium-sized datasets, and more interpretable than many neural approaches (Kowsari et al., 2019). In bibliographic classification, the usefulness of these methods depends strongly on the quality and richness of the available metadata. Titles, subtitles, descriptions, and abstracts may contain discriminative terms that indicate the subject of a work. However, bibliographic records are often short, noisy, or incomplete, while different books belonging to the same broad subject area may use substantially different vocabularies. As a result, lexical methods can perform well when subject categories are associated with distinctive terminology, but their performance is limited when the same subject is expressed indirectly or when different subjects share similar terms (Nordström, 2018). Another limitation of classical lexical approaches is that they usually represent documents according to surface-level word occurrence. Consequently, semantic similarity is captured only indirectly through shared vocabulary. This can be problematic in book classification, where short descriptions may refer to the same topic using different words, or where the subject of a book is implied by context rather than stated explicitly. Nevertheless, such methods are still useful in empirical studies because they provide transparent and computationally inexpensive baselines against which more complex representation-learning approaches can be compared.

2.3 Transformer-Based Text Representations

More recent approaches to text classification rely on neural language models that learn dense semantic representations of text. The transformer architecture introduced the attention-based modelling principle on which many modern language models are based (Vaswani et al., 2017). Building on this paradigm, BERT introduced deep bidirectional pre-training as a general framework for language understanding tasks (Devlin et al., 2019). Unlike TF-IDF representations, which encode documents primarily through weighted word occurrence, transformer-based models can capture contextual meaning and semantic similarity between texts even when they do not share the same vocabulary. In library and information science, BERT-based models have also been explored for assisted subject indexing, indicating their relevance for automatic metadata enrichment and subject classification tasks (Chou, 2022). Sentence-BERT further adapted transformer models for efficient sentence-level and paragraph-level representation by producing embeddings suitable for similarity comparison and downstream classification tasks (Reimers & Gurevych, 2019). Such representations are especially relevant for bibliographic metadata because book titles and descriptions are often short textual units. In these cases, the ability to encode semantic similarity beyond exact word overlap may improve classification performance, particularly for broad subject categories whose textual expressions vary across records. However, transformer-based approaches also introduce several practical limitations. They are generally less transparent than classical lexical models, require more computational resources, and their performance may depend on the choice of the pre-trained model and downstream classifier (Chou, 2022; Devlin et al., 2019; Reimers & Gurevych, 2019). In addition, when bibliographic subject labels are available in the metadata, care must be taken to avoid using them as input features, because this may lead to label leakage and produce overly optimistic classification results (Kaufman et al., 2012).

Although previous studies report encouraging results for automatic subject indexing and text classification, direct numerical comparison is difficult because datasets, metadata fields, target vocabularies, and evaluation protocols differ substantially. Some studies predict controlled subject headings, some use classification codes, while others classify documents into broad topical groups. Reported F1-scores are therefore not directly comparable to the present ten-category Open Library setting. Nevertheless, prior work generally indicates that performance decreases when the input is restricted to short metadata fields and when categories are broad or semantically overlapping. The present study contributes to this line of work by reporting a reproducible comparison on title–description metadata only, where the best model achieved a macro-F1 score

of 0.6173.

3 Methodology

This section describes the experimental procedure used for automatic subject classification of books from open bibliographic metadata. The methodology consists of five main steps: acquisition and parsing of Open Library records, extraction of relevant bibliographic fields, construction of broad subject categories, transformation of textual metadata into numerical representations, and evaluation of classification models.

3.1 Data Source and Record Extraction

The dataset used in this study was constructed from the Open Library works data dump. Open Library provides bibliographic records in a tab-separated format, where each row contains the record type, record key, revision, last modification date, and a JSON representation of the corresponding work (Open Library, 2025). In this study, the works dump was parsed line by line in order to extract records suitable for supervised subject classification.

For each valid JSON record, the following fields were extracted: the Open Library work key; the book title; the textual description; the original subject labels; the last modification date. The description field in Open Library records may appear either as a plain string or as a structured object containing a textual value. Both cases were handled during preprocessing. The title and description were cleaned by removing line breaks, tab characters, and repeated whitespace. They were then concatenated into a single textual input field. Records with insufficient textual information were excluded by requiring the resulting title–description text to contain at least five words.

Importantly, original subject labels were not used as input features. They were used only for constructing the target variable, as described in Section 3.2. This distinction was necessary in the present experimental design because the ten broad target categories were derived from the same subject-label field. Therefore, using subject labels as predictive input would introduce label leakage and lead to an unrealistic estimate of classification performance (Kaufman et al., 2012). This restriction is specific to the present setup. If the broad categories were obtained independently, for example through manual annotation by human experts, subject labels could represent a legitimate and valuable source of predictive metadata.

3.2 Construction of Target Categories

Open Library subject labels are highly diverse, fine-grained, and not always standardized. Therefore, direct classification using the original labels would result in a large and sparse label space. In order to obtain a controlled experimental setting, the original subject labels were mapped into ten broad subject categories: Art, Biography, Business, Education, Fiction, History, Medicine, Religion, Science, and Technology.

Table 1: Dataset construction audit during Open Library parsing.

Filtering outcome	Records
Parsed JSON records	357,133
Missing/invalid subject list	174,688
No broad category matched	85,078
More than one category matched	23,291
Exactly one category matched	74,076
Fewer than five text words	36,648
Excluded after class cap	22,428
Final usable records	15,000

For each category, a manually defined set of representative keywords was used. For example, the Fiction category included terms such as *fiction*, *novel*, *stories*, *fantasy*, and *mystery*; the Technology category included terms such as *technology*, *engineering*, *computer*, *programming*, *software*, and *machine learning*; while the Science category included terms such as *science*, *physics*, *chemistry*, *biology*, *mathematics*, and *astronomy*.

All subject labels were normalized before matching by converting text to lowercase, removing non-alphanumeric characters, and reducing repeated whitespace. Keyword matching was performed using word-aware regular expressions in order to reduce accidental substring matches. A record was assigned to a broad category only if its subject labels matched exactly one of the predefined categories. Records that matched more than one category were excluded, as were records that could not be mapped reliably to any selected category. This procedure was used to construct a balanced dataset of 15,000 records, with 1,500 books per category. During parsing, at most 500,000 lines of the Open Library works dump were scanned, and the procedure stopped earlier if the required number of records was obtained for all categories. The balanced design was used to prevent frequent categories from dominating the training process and to make the comparison of classification performance across categories more reliable.

Because no independent manually annotated gold standard was available for the broad categories, precision and recall of the keyword mapping itself could not be estimated in the strict supervised sense. The mapping was therefore used as a deterministic dataset construction rule rather than as an evaluated classifier. To reduce obvious mapping noise, only records that matched exactly one broad category were retained. A manual validation of a stratified sample of mapped records would be required to estimate the mapping precision and recall directly.

Table 1 summarizes the main filtering outcomes observed during dataset construction. Out of 357,133 parsed JSON records, 85,078 records did not match

any of the selected broad categories, while 23,291 records matched more than one broad category and were therefore excluded from the single-label dataset. In addition, 36,648 records were excluded because the title–description text contained fewer than five words. Among the records that matched exactly one broad category, 36,648 out of 74,076 records were excluded because the title–description text contained fewer than five words, corresponding to 49.47%. The final dataset contained 15,000 records, corresponding to 1,500 records for each of the ten broad subject categories.

This sampling strategy creates a controlled and balanced benchmark, but it may also introduce selection bias. Records with clear and unambiguous subject labels are more likely to be retained, while sparse, noisy, or genuinely multi-subject records are more likely to be excluded. Therefore, the final dataset should not be interpreted as a representative sample of all Open Library records, but as a filtered benchmark for evaluating classification from title–description metadata under clean single-label conditions.

The task was formulated as single-label classification in order to obtain a controlled benchmark for comparing textual representations and classifiers under the same evaluation protocol. This design choice simplifies the bibliographic reality, since books may naturally belong to more than one subject area. However, retaining only records that matched exactly one broad category made it possible to construct a balanced dataset and to evaluate all models using standard multiclass metrics. The implications of this simplification are considered in the error analysis and in the discussion of future multi-label extensions.

3.3 Text Representation

Let x_i denote the textual input of the i-th book record, obtained by concatenating its title and description, and let $y_i \in \mathcal{C}$ denote the corresponding broad subject category. The dataset is therefore represented as

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N},$$

where $N = 15000$.

Two types of textual representation were evaluated. First, a classical lexical representation was obtained using term frequency–inverse document frequency (TF-IDF). The TF-IDF vectorizer was configured to use lowercasing, English stop-word removal, unigrams and bigrams, a maximum vocabulary size of 50,000 terms, a minimum document frequency threshold of three documents (`min_df=3`), a maximum document frequency threshold of 90% of the training documents (`max_df=0.90`), and sublinear term-frequency scaling. This representation was used as input for the classical machine learning models.

Second, a transformer-based semantic representation was evaluated using the `sentence-transformers/all-MiniLM-L6-v2` model. Each title–description text was encoded into a dense sentence embedding. The embeddings were normalized and then used as input to a Logistic Regression classifier. This setting does not perform full transformer fine-tuning; instead, the transformer is used as a fixed sentence encoder, followed by a conventional supervised classifier. Because the `sentence-transformers/all-MiniLM-L6-v2` encoder truncates inputs longer than 256 word pieces by default, the tokenized length of each title–description input was additionally inspected. This check was used to determine how many records may have been affected by truncation before embedding generation.

For k-nearest neighbors and the multilayer perceptron classifier, dimensionality reduction was applied before classification. This was not intended to imply that nearest-neighbor methods are generally unsuitable for TF-IDF representations. In fact, nearest-neighbor classification can be applied to sparse TF-IDF vectors when an appropriate similarity measure, such as cosine similarity, is used. In the present experiment, truncated singular value decomposition was used as a practical design choice to obtain a compact dense representation, reduce computational cost, and use the same reduced representation for both the KNN and MLP baselines. The reduced vectors were then standardized before classification.

This configuration was selected deliberately as a lightweight and reproducible setting. The objective was to evaluate whether frozen semantic embeddings can improve over classical lexical representations when computational resources and training data are limited. Full transformer fine-tuning was not included in the present comparison because it introduces additional computational cost, hyperparameter sensitivity, and a different experimental setting.

3.4 Classification Models

Six classification approaches were compared. Five approaches used TF-IDF-based lexical features, while one approach used transformer-based sentence embeddings:

1. TF-IDF with Decision Tree;
2. TF-IDF with Random Forest;
3. TF-IDF with Linear Support Vector Machine;
4. TF-IDF with truncated SVD and k-Nearest Neighbors;
5. TF-IDF with truncated SVD and multilayer perceptron neural network;
6. `sentence-transformers/all-MiniLM-L6-v2` sentence embeddings with Logistic Regression.

The Decision Tree classifier was trained with a maximum depth of 60 and a minimum of three samples per leaf. The Random Forest classifier used 300 trees, a minimum of two samples per leaf, and balanced subsampling of class weights. The Linear Support Vector Machine used a linear decision function with $C = 1.0$ and balanced class weights. The k-Nearest Neighbors classifier used $k = 7$, distance-based weighting, and cosine distance. The multilayer perceptron consisted of two hidden layers with 256 and 128 neurons, ReLU activation, L_2 regularization parameter $\alpha = 10^{-4}$, initial learning rate 10^{-3}, batch size 64, and a maximum of 50 training iterations. Finally, the transformer-based approach used Logistic Regression with balanced class weights and a maximum of 2000 iterations.

The selected models represent different classification principles. Decision Tree and Random Forest models represent tree-based learning, Linear Support Vector Machine is a strong linear baseline for sparse textual data, k-Nearest Neighbors represents similarity-based classification, and the multilayer perceptron provides a neural baseline on top of reduced TF-IDF features. The MiniLM-based approach was included to evaluate whether dense transformer embeddings provide an advantage over lexical representations when only title and description are available as input.

In addition to the main fair comparison, a separate diagnostic upper-bound experiment was conducted using the original Open Library subject labels as predictive text. This experiment was not treated as part of the main comparison, because the broad target categories were derived from the same metadata field. Its purpose was only to estimate how much performance may increase when the classifier receives metadata that is directly related to the target construction rule.

3.5 Training and Testing Procedure

The final dataset was randomly shuffled using a fixed random seed and divided into training and test subsets. A stratified split was used, with 75% of the records assigned to the training set and 25% assigned to the test set. Stratification preserved the same class distribution in both subsets. Since the dataset was balanced, each category was represented equally in the training and test data.

All models were trained only on the training subset. The test subset was used exclusively for final evaluation. This ensured that the reported results measured the ability of the models to generalize to previously unseen bibliographic records.

3.6 Evaluation Metrics

Model performance was evaluated using standard multiclass classification metrics. The following measures were reported: accuracy; balanced accuracy; macro-averaged precision; macro-averaged recall; macro-averaged F1-score; weighted F1-score; and total runtime. Since the dataset was balanced, accuracy and balanced accuracy are directly comparable, and accuracy is not dominated by more frequent classes. Macro-F1 was nevertheless used as the primary comparison metric because it combines class-level precision and recall and is sensitive to uneven error patterns across categories. This is relevant in the present task because some broad categories are semantically more overlapping than others, even though all categories contain the same number of test examples. Weighted-F1 was reported additionally as an aggregate measure that accounts for the number of samples per class. In addition to aggregate metrics, a full classification report was generated for each model. This allowed the comparison of precision, recall, and F1-score at the category level.

3.7 Analysis of Misclassification Patterns

After model evaluation, the models were ranked according to macro-F1. The confusion matrix of the best-performing model was then generated and analyzed. This analysis was used to identify common misclassification patterns and subject categories that were frequently confused. Such an analysis is particularly important in bibliographic classification because not all errors have the same interpretation. Misclassifications between categories such as Biography and History, Science and Technology, or Business and Education may reflect genuine semantic overlap in the bibliographic descriptions. Therefore, the confusion matrix was interpreted not only as an error analysis tool, but also as an indicator of the conceptual difficulty of broad subject classification. Finally, example predictions of the best-performing model were saved for qualitative inspection. These examples included the original record information, the true category, the predicted category, and an indicator of whether the prediction was correct. This allowed a more detailed interpretation of individual correct and incorrect classifications.

4 Results

In this section, the results of the automatic subject classification experiment described in Section 3 are presented. The evaluation was conducted on the stratified test set containing 3,750 bibliographic records, with 375 records from each of the ten subject categories. Since the dataset was balanced, accuracy and balanced accuracy are directly comparable. Macro-F1 is used as the primary performance indicator because it summarizes class-level precision and recall and is sensitive to uneven error patterns across semantically overlapping categories.

The MiniLM truncation audit showed that 324 out of 15,000 records exceeded the 256 word-piece limit, corresponding to 2.16% of the dataset. The median input

length was 15 word pieces, the mean input length was 35.0 word pieces, and the 95th percentile was 166 word pieces. Therefore, truncation affected only a small subset of records, although some descriptions were substantially longer, with the maximum observed length reaching 1,732 word pieces. This should be considered when interpreting the transformer-based results for records with longer descriptions.

4.1 Model Comparison

Table 2 reports the performance of the evaluated classification models. The compared approaches include five TF-IDF-based classical machine learning models and one transformer-based semantic representation using MiniLM embeddings followed by Logistic Regression.

The reported runtime denotes the total measured execution time of the evaluated pipeline. For TF-IDF-based models, this includes TF-IDF feature construction, model training, and prediction on the test set. For the MiniLM-based model, it includes embedding generation for the training and test texts, Logistic Regression training, and prediction. Therefore, the runtime values should be interpreted as end-to-end pipeline times rather than isolated classifier training times.

The diagnostic subject-label experiment achieved an accuracy of 0.9995 and a macro-F1 score of 0.9995. This result should not be interpreted as a fair comparison with the title–description models, because the subject labels were also used to construct the broad target categories. However, the near-perfect performance confirms that subject labels contain direct category information. This supports their exclusion from the main predictive input and shows why the title–description setting provides a more realistic estimate of classification performance.

The best result was obtained by the transformer-based model, which achieved an accuracy of 0.6157 and a macro-F1 score of 0.6173. The strongest classical baseline was TF-IDF + Linear SVM, with a macro-F1 score of 0.5323, followed by TF-IDF + Random Forest with a macro-F1 score of 0.5238. Thus, the MiniLM-based representation improved macro-F1 by approximately 8.5 percentage points compared with the best TF-IDF-based baseline. This result indicates that dense semantic embeddings are more suitable for this task than purely lexical representations. The advantage is expected, since the input consists of relatively short bibliographic texts in which the subject category may be expressed indirectly. Nevertheless, the TF-IDF + Linear SVM and TF-IDF + Random Forest models achieved competitive results, especially considering their substantially lower execution time. Classical lexical models therefore remain relevant baselines for automatic bibliographic classification, particularly in settings where interpretability and computational efficiency are important. Figure 1 provides a visual comparison of the macro-F1 scores across all evaluated models.

4.2 Per-Class Performance

Since the MiniLM embeddings combined with Logistic Regression achieved the highest macro-F1 score, this model was analyzed further at the level of individual subject categories. Table 3 reports precision, recall, and F1-score for each class.

The highest F1-score was achieved for Technology, followed by Business, Medicine, Religion, and Education. These categories are characterized by relatively distinctive vocabularies, such as terms related to computing, engineering, management, health, theology, or teaching. In contrast, the lowest F1-scores were obtained for Biography and History. This is consistent with the structure of the task, since these two categories are often semantically close. Historical books may focus on individuals, while biographical books frequently contain historical context. The obtained class-level results therefore show that the classification difficulty is not uniform across categories. Categories with more specialized terminology are classified more reliably, while categories with broader or overlapping semantic scope produce more errors.

4.3 Confusion Matrix Analysis

In order to examine the structure of classification errors, the confusion matrix of the best-performing model was analyzed. Table 4 reports the confusion matrix, where rows correspond to true labels and columns correspond to predicted labels, while Figure 2 provides its graphical representation.

The confusion matrix confirms that the model correctly classified the majority of records in all categories. The highest number of correct classifications was obtained for Technology, with 274 correctly classified records, followed by Business with 259, Religion with 252, and Medicine with 245. These results are consistent with the per-class F1-scores reported in Table 3. The most evident confusion occurs between semantically related categories. Biography is often confused with History, with 58 Biography records classified as History. Conversely, 42 History records were classified as Biography. This confirms that the boundary between these categories is not always clear when only title and description are used. A similar pattern can be observed between Science and Technology. Technology achieved the highest overall class performance, but 36 Technology records were classified as Science, while 25 Science records were classified as Technology. This indicates that applied scientific and technical descriptions may share similar vocabulary. Additional confusions are also meaningful from a bibliographic perspective. Art is sometimes classified as History, which may correspond to art history records.

Table 2: Model comparison on the Open Library subject classification task.

Model	Accuracy	Balanced Acc.	Macro Prec.	Macro Recall	Macro-F1	Weighted-F1	Total runtime (s)
Transformer MiniLM embeddings + Logistic Regression	**0.6157**	**0.6157**	**0.6205**	**0.6157**	**0.6173**	**0.6173**	92.88
TF-IDF + Linear SVM	0.5301	0.5301	0.5359	0.5301	0.5323	0.5323	1.68
TF-IDF + Random Forest	0.5173	0.5173	0.5397	0.5173	0.5238	0.5238	2.99
TF-IDF + SVD + MLP Neural Network	0.4739	0.4739	0.4875	0.4739	0.4766	0.4766	36.92
TF-IDF + SVD + KNN	0.4200	0.4200	0.4294	0.4200	0.4222	0.4222	7.16
TF-IDF + Decision Tree	0.2877	0.2877	0.6171	0.2877	0.3172	0.3172	1.35

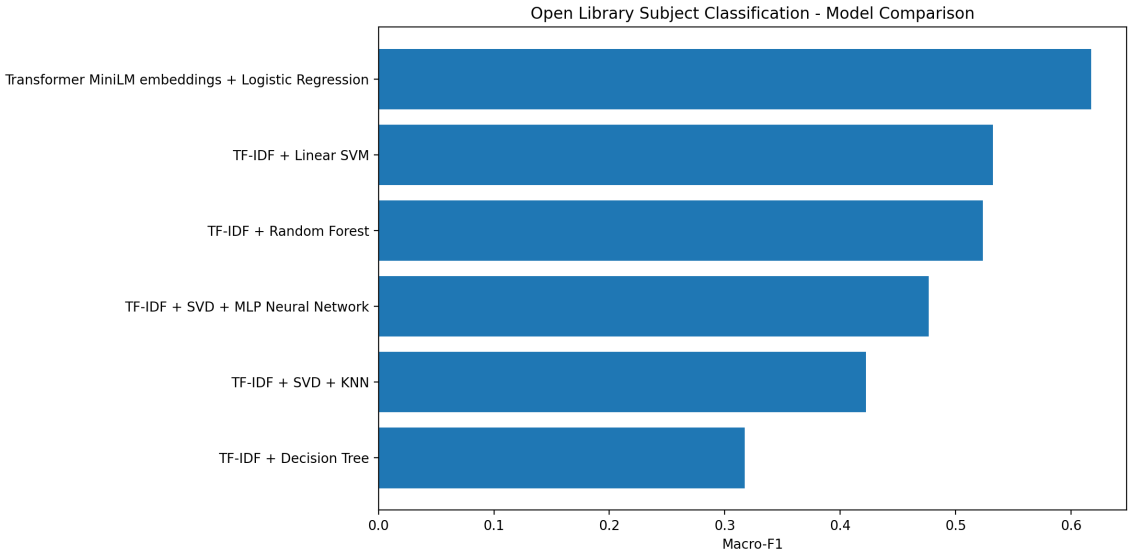


Figure 1: Comparison of the evaluated classification models according to macro-F1. The transformer-based model achieved the highest score, followed by TF-IDF + Linear SVM and TF-IDF + Random Forest.

Table 3: Per-class performance of the best-performing model: Transformer MiniLM embeddings + Logistic Regression.

Class	Precision	Recall	F1-score	Support
Art	0.587	0.576	0.581	375
Biography	0.454	0.483	0.468	375
Business	0.687	0.691	0.689	375
Education	0.712	0.613	0.659	375
Fiction	0.608	0.608	0.608	375
History	0.462	0.525	0.492	375
Medicine	0.712	0.653	0.682	375
Religion	0.648	0.672	0.660	375
Science	0.624	0.605	0.614	375
Technology	**0.712**	**0.731**	**0.721**	375

Fiction is sometimes confused with Biography, which may reflect memoir-like, biographical, or narrative descriptions. Religion also shows confusions with Biography and History, which may occur in works about religious figures, traditions, or historical religious movements. These results suggest that a part of the observed error is not caused only by model limitations, but also by the inherent ambiguity of broad subject classification. In such cases, a single-label prediction may be too restrictive, since some books can reasonably belong to more than one subject area. Consequently, the reported single-label accuracy may underestimate the practical usefulness of the model in a decision-support setting. A prediction counted as incorrect under the single-label protocol may still be useful if it represents a plausible secondary category for librarian review. For this reason, the present results should be interpreted as conservative evidence for broad category suggestion rather than as a complete evaluation of multi-subject cataloguing.

4.4 Summary of Results

The experimental results show that transformer-based sentence embeddings achieved the best overall performance on the Open Library subject classification task. The MiniLM embeddings combined with Logistic Regression achieved a macro-F1 score of 0.6173, outperforming the strongest classical baseline, TF-IDF + Linear SVM, by approximately 8.5 percentage points.

At the same time, the results show that the task remains challenging. Even the best model achieved moderate, rather than near-perfect, classification performance. This is expected because the input was intentionally restricted to title and description, while original subject labels were excluded from the predictive representation. Therefore, the task reflects a realistic setting in which the model must infer the subject category from incomplete and sometimes ambiguous bibliographic text.

Overall, the results support the use of semantic text representations for automatic subject-category predic-

Table 4: Confusion matrix for the best-performing model. Rows represent true labels and columns represent predicted labels.

True / Pred.	Art	Biography	Business	Education	Fiction	History	Medicine	Religion	Science	Technology
Art	**216**	34	6	11	17	43	7	12	9	20
Biography	36	**181**	11	5	38	58	8	28	10	0
Business	6	13	**259**	12	3	23	16	4	15	24
Education	15	10	12	**230**	16	22	18	19	15	18
Fiction	14	49	4	4	**228**	19	9	25	17	6
History	35	42	19	9	23	**197**	10	20	13	7
Medicine	8	24	21	14	19	8	**245**	14	13	9
Religion	18	34	6	7	18	23	6	**252**	9	2
Science	9	12	17	21	13	22	17	12	**227**	25
Technology	11	0	22	10	0	11	8	3	36	**274**

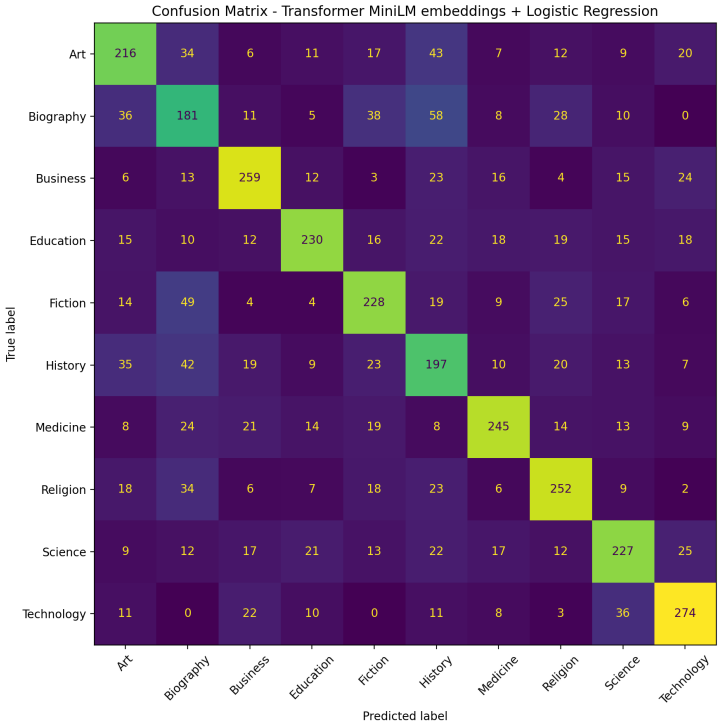


Figure 2: Confusion matrix for the best-performing model, Transformer MiniLM embeddings + Logistic Regression. Diagonal values represent correct classifications, while off-diagonal values represent misclassifications.

tion from bibliographic metadata. However, the observed misclassification patterns also show that automatic classification should be interpreted as a decision-support mechanism. The model can provide useful candidate categories, but expert review remains important, especially for records located near the semantic boundaries between broad subject areas. In a practical cataloguing workflow, the model should be used as a decision-support tool rather than as an autonomous cataloguing system. For each new record, the model could provide the most likely broad category together with alternative high-ranking categories for librarian review. Since the present study did not perform probability calibration or threshold optimization, no fixed confidence threshold for automatic acceptance can be recommended. In deployment, such a threshold should be selected on a validation set according to institutional requirements. For example, high-confidence predictions could be accepted automatically only when precision exceeds a predefined operational target, while lower-confidence and semantically borderline cases should be routed to expert review.

5 Conclusion

This study presented a reproducible experiment for automatic subject classification of books based on open bibliographic metadata from Open Library. The predictive input was intentionally limited to title and description, while original subject labels were used only to construct the target categories. In this way, the evaluation avoided label leakage and reflected a more realistic setting for metadata enrichment. The results showed that MiniLM sentence embeddings combined with Logistic Regression achieved the best overall per-

formance, outperforming classical TF-IDF-based models. However, the obtained macro-F1 score of 0.6173 also shows that the task remains challenging. Many errors occurred between semantically related categories, such as Biography and History or Science and Technology, indicating that some misclassifications reflect genuine ambiguity in bibliographic subject assignment.

Overall, the findings suggest that semantic text representations can support broad subject-category prediction from limited bibliographic descriptions. Nevertheless, automatic classification should be understood as a decision-support mechanism rather than a replacement for expert cataloguing. The dataset construction procedure may favor records with clearer and less ambiguous subject metadata, since ambiguous, sparse, and unmapped records were excluded. In addition, the transformer model was used only as a fixed sentence encoder. Fine-tuning MiniLM or another transformer model on the same training data may improve performance, but it would require a separate experimental design and additional hyperparameter control. Future work should reformulate the task as multi-label category recommendation, where several plausible subject categories are ranked and presented to librarians for review.

References

Chou, C.-Y. (2022). An analysis of bert (nlp) for assisted subject indexing. *Cataloging & Classification Quarterly*, *60*(8), 683–702. doi: 10.1080/01639374.2022.2138666

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: Human language technologies* (Vol. 1, pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics. doi: 10.18653/v1/N19-1423

Golub, K. (2019). Automatic subject indexing of text. In B. Hjørland & C. Gnoli (Eds.), *Isko encyclopedia of knowledge organization.* International Society for Knowledge Organization. Retrieved from `https://www.isko.org/cyclo/automatic`

Golub, K., Hagelbäck, J., & Ardö, A. (2021). Automated subject indexing: An overview. *Cataloging & Classification Quarterly*, *59*(8), 702–719. doi: 10.1080/01639374.2021.2012311

Kaufman, S., Rosset, S., & Perlich, C. (2012). Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data*, *6*(4), 15. doi: 10.1145/2382577.2382579

Kazi, N., Thompson, B., Choudhury, S. R., & Oard, D. W. (2021). Automatically cataloging scholarly articles using library of congress subject headings. In *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: Student research workshop* (pp. 69–75). Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-srw.7

Kowsari, K., Meimandi, K. J., Heidarysafa, M., Mendu, S., Barnes, L. E., & Brown, D. E. (2019). Text classification algorithms: A survey. *Information*, *10*(4), 150. doi: 10.3390/info10040150

Mödden, E. (2022). Artificial intelligence, machine learning and bibliographic control: Ddc short numbers: Towards machine-based classifying. *JLIS: Italian Journal of Library, Archives and Information Science= Rivista italiana di biblioteconomia, archivistica e scienza dell'informazione: 13, 1, 2022*, 256–264.

Nordström, J. (2018). *Automated classification of bibliographic data using svm and naive bayes* (Unpublished master's thesis). Uppsala University.

Open Library. (2025). *Open library data dumps.* `https://openlibrary.org/developers/dumps`. (Accessed: 2026-05-17)

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing* (pp. 3982–3992). doi: 10.18653/v1/D19-1410

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, *24*(5), 513–523. doi: 10.1016/0306-4573(88)90021-0

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, *34*(1), 1–47. doi: 10.1145/505282.505283

Suominen, O. (2019). Annif: Diy automated subject indexing using multiple algorithms. *LIBER quarterly*, *29*(1), 1–25.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (Vol. 30).

Cross-Method Analysis of Explainability in EEG-Based Major Depressive Disorder Classification

Antonia Šarčević, Alan Jović, Nikolina Frid
University of Zagreb Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Croatia
{antonia.sarcevic, alan.jovic, nikolina.frid}@fer.unizg.hr

Jakša Vukojević
University Psychiatric Hospital Vrapče
Bolnička 32, 10000 Zagreb, Croatia
jaksa.vukojevic@bolnica-vrapce.hr

Abstract. *Deep learning models have shown promising performance for EEG-based Major Depressive Disorder (MDD) detection, but their clinical use requires methods that can provide insight into model predictions. This study compares multiple post-hoc explainability techniques, including gradient-, perturbation-, and Shapley-based approaches, applied to an InceptionTime classifier trained on resting-state clinical EEG recordings. The results reveal partial agreement between attribution methods, alongside notable differences arising from their underlying methodological assumptions. These findings indicate that post-hoc explanations provide complementary but method-dependent perspectives on model behavior, highlighting the importance of assessing cross-method consistency when interpreting EEG-based depression classifiers.*

Keywords. Major Depressive Disorder, EEG, Explainable Artificial Intelligence, Deep Learning, InceptionTime

1 Introduction

Major depressive disorder (MDD) is among the most prevalent psychiatric disorders worldwide (GBD 2017 Disease and Injury Incidence and Prevalence Collaborators, 2018). Despite extensive research, diagnosis remains primarily based on clinical assessment, motivating the search for objective measures that could support MDD evaluation. In this context, electroencephalography (EEG) has attracted considerable interest as a non-invasive and accessible modality for assessing MDD. Numerous studies have reported alterations in spectral, connectivity, and nonlinear EEG characteristics (Čukić et al., 2020; Kaiser et al., 2018; Simmatis et al., 2023); however, no objective biological marker has yet been established for routine clinical use (Kendler, 2016).

The increasing availability of EEG data has motivated the application of machine learning methods capable of identifying complex patterns that may not be apparent through conventional analysis alone. Both classical feature-based approaches and more recent deep learning models have reported promising results for MDD detection (Aleem et al., 2022; Avots et al., 2022; Dutua and Sultana, 2025; Wang et al., 2024). However, modern deep learning and time-series classification models learn high-dimensional representations whose relationship to underlying neurophysiological processes is often unclear, limiting direct human interpretation of model behavior (Elton, 2020; Jovic et al., 2025). Such limitations are particularly important in high-stakes domains such as healthcare (Rudin, 2019).

To improve the interpretability of the model, post-hoc explainability methods are commonly applied to black-box models (Linardatos et al., 2021; Marcinkevičs and Vogt, 2023). Widely used approaches include Shapley-value-based methods, gradient-based attribution methods, and perturbation-based techniques (Lundberg and Lee, 2017; Sundararajan et al., 2017; Zeiler and Fergus, 2014). Beyond improving transparency, such methods may help relate model predictions to physiologically meaningful EEG characteristics and support scientific interpretation of learned representations (Hicks et al., 2021; Mulc et al., 2025). However, these approaches rely on different assumptions and approximation strategies, which may lead to substantially different explanations for the same prediction (Jovic et al., 2025; Krishna et al., 2022). Consequently, explanation stability, robustness, and physiological plausibility remain important concerns. Despite these challenges, EEG studies commonly apply a single explainability technique, while systematic investigation of agreement and disagreement between multiple post-hoc explainability methods remains relatively limited.

In this work, we apply multiple post-hoc explainability methods to an InceptionTime model (Ismail Fawaz

et al., 2020) trained for EEG-based MDD detection. We investigate whether different explainability methods produce convergent attribution patterns and examine the extent to which their explanations remain consistent across the same classification task. Through this analysis, we assess both the potential usefulness and the limitations of post-hoc explainability for interpreting black-box EEG models in depression detection.

2 Materials and Methods

2.1 Dataset Description

EEG recordings were obtained at the University Psychiatric Hospital Vrapče using a Nihon Kohden EEG-1200K system from 140 adults: 70 patients with current moderate-to-severe MDD diagnosed by senior psychiatrists according to ICD-10 and 70 controls screened to exclude psychiatric or neurological disorders and psychopharmacotherapy. Groups were matched for age and sex as closely as possible (MDD: 37 women, 36.86±10.22 years; 33 men, 45.24±12.10; controls: 32 women, 35.88±9.96; 38 men, 36.16±11.35). Further cohort criteria are reported in (Mulc et al., 2025). Signals were sampled at 200 Hz from 19 electrodes positioned according to the international 10–20 system. Each 30-minute session began with resting-state recording and continued with activation protocols. Preprocessing included band-pass filtering (0.1–40 Hz), common-average re-referencing, and ICA-based artifact removal using ICLabel. The first 300 seconds of resting-state EEG were segmented into 2-second windows with a 0.5-second (25%) overlap, corresponding to a 1.5-second stride.

2.2 Model Training and Evaluation

The InceptionTime architecture (Ismail Fawaz et al., 2020) was selected due to its strong performance in time-series classification tasks (Middlehurst et al., 2024), including EEG analysis (Kendjel et al., 2026). Experiments were conducted using the InceptionTime implementation provided by the tsai library (Oguiza, 2023), which employs parallel convolutions with multiple kernel sizes to capture temporal patterns at different temporal scales and residual connections to facilitate optimization.

Channel-wise z-score normalization was performed using statistics computed exclusively from the training data within each fold. Model evaluation was conducted using subject-level stratified 5-fold cross-validation, with all epochs from each participant retained in the same fold to prevent information leakage. Within each training fold, 10% of subjects were reserved for validation. Epochs inherited the corresponding participant-level diagnosis as weak supervision, while inference and evaluation remained subject-level. For each subject, epoch-wise logits were aggregated using a trimmed-mean operator and converted into a single probability estimate before assigning the final class label. All reported metrics were calculated from these subject-level predictions. Training was performed using the Adam optimizer, cross-entropy loss, and One-Cycle learning-rate scheduling, and the checkpoint with the lowest validation loss was retained for subsequent explainability analysis.

2.3 Explainability Framework

To analyze model behavior, we evaluated five post-hoc explainability methods representing three major attribution paradigms: DeepSHAP (Lundberg and Lee, 2017; Shrikumar et al., 2017), Integrated Gradients (IG) (Sundararajan et al., 2017), Grad-CAM (Selvaraju et al., 2017), Occlusion (Zeiler and Fergus, 2014), and Permutation Feature Importance (PFI) (Breiman, 2001; Fisher et al., 2019). For each EEG segment, channel–time attributions were aggregated into channel-level importance values and then averaged across segments, subjects, and cross-validation folds. Signed attributions were converted to absolute values before channel-level aggregation; the resulting profiles therefore represent attribution magnitude rather than direction.

DeepSHAP was implemented using Captum's `DeepLiftShap` module with a background reference set of 40 randomly selected training samples per fold.

Integrated Gradients was computed using 100 interpolation steps. To assess baseline sensitivity, both zero-valued and Gaussian-noise baselines were evaluated.

Grad-CAM was applied to the final Inception block prior to global average pooling. To obtain channel-level relevance estimates, Grad-CAM was supplemented with InputXGradient attribution.

Occlusion analysis was performed using sliding temporal masks of size $(1, 50)$ and stride $(1, 25)$.

Permutation Feature Importance was computed by evaluating each EEG channel independently. For a selected channel, its temporal sequence was randomly permuted across the evaluation batch while all remaining channels were left unchanged. Channel importance was then quantified by the resulting decrease in target-class prediction confidence.

Frequency-band attribution was derived from the DeepSHAP and IG channel–time maps. A real FFT (rFFT) was applied along the temporal axis, and the resulting magnitudes were pooled into delta (0.5–4 Hz), theta (4–8 Hz), alpha (8–13 Hz), and beta (13–30 Hz) bands. Each score array was min–max normalized to the range 0–1; arrays with a range below 10^{-8} were set to zero. Class-specific scores were averaged across segments, subjects, and folds and then summed across channels within each region. The heatmaps therefore represent cumulative regional spectral attribution magnitude, which may depend partly on the number of channels assigned to each region, rather than EEG band

power.

Explanation robustness was evaluated using the coefficient of variation (CoV) for fold-wise variability, the Gini coefficient for attribution concentration, and Kendall's τ for cross-method ranking agreement. Higher Gini values indicate greater concentration. Fold-wise rank stability was additionally assessed using an overlap-based SRA score adapted from the sequential ranking principle of Ekstrøm et al. (2019). For each fold pair, the top-k Jaccard overlap at depth k was weighted by $1/k$, normalized by $\sum_{k=1}^{d} 1/k$, and averaged across fold pairs. SRA ranges from 0 to 1, with higher values indicating greater agreement. Regional SRA applies the same procedure after aggregating channels into cortical regions. In addition, a spatial-temporal variation metric (STV) (Mortaheb et al., 2019) was used to quantify the smoothness of attribution maps across neighboring EEG channels.

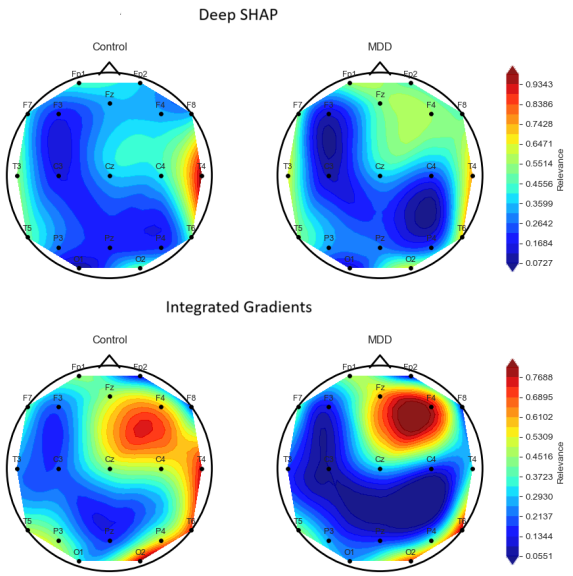


Figure 1. Topographic attribution maps

3 Results

3.1 Classification Performance

Classification performance was evaluated using subject-level predictions aggregated across EEG segments within a stratified 5-fold cross-validation framework. Mean performance metrics are summarized in Table 1. The model achieved strong and consistent performance across folds, with an average ROC AUC of 0.976 and PR AUC of 0.980. High precision (0.926) and recall (0.914) further indicate reliable identification of both MDD and control subjects. These internally validated subject-level results provide a suitable basis for analyzing model behavior but do not establish clinical validity or external generalizability.

Table 1. Subject-level classification performance (mean $\pm$ standard deviation).

Metric	Mean $\pm$ STD
Accuracy	0.914 ± 0.048
Precision	0.926 ± 0.032
Recall	0.914 ± 0.048
F1-score	0.913 ± 0.051
ROC AUC	0.976 ± 0.019
PR AUC	0.980 ± 0.014

3.2 Explainability Results

For each explainability method, attribution maps were computed for individual EEG segments and aggregated across subjects and cross-validation folds to obtain global class-level relevance profiles for the Control and MDD groups. Preliminary temporal attribution analysis did not reveal consistent patterns across folds or subjects. Consequently, the following analysis focuses on aggregated spatial and spectral attribution trends. Detailed visualizations are presented for DeepSHAP and Integrated Gradients, while channel-level agreement across all evaluated explainability methods is examined through the subsequent cross-method comparison and consensus analyses.

3.2.1 Global attribution patterns

DeepSHAP and Integrated Gradients (IG) produced broadly similar spatial attribution patterns, indicating partial agreement between the two explainability approaches (Fig. 1). Both methods highlighted the importance of right-hemisphere regions and revealed a broader spatial attribution distribution in the MDD group compared to Controls. For DeepSHAP, attributions in the Control group remained comparatively localized around the right temporal region, with T4 emerging as the most influential channel. In contrast, the MDD group exhibited a wider attribution pattern involving additional right-hemisphere channels, particularly T6, O2, F4, and the midline frontal channel Fz. Integrated Gradients showed stronger emphasis on frontal regions, especially F4 and Fz, in both groups and more prominently in MDD. While right temporal relevance remained evident for Controls, IG also attributed greater importance to parietal and occipital regions, including T5, T6, and O2.

The regional-frequency attribution maps generated using DeepSHAP (Fig. 2) indicate substantial contributions of beta-band activity in both groups. In the MDD group, attribution was more broadly distributed across delta-, theta-, and alpha-band components, with the strongest emphasis observed in frontal regions.

The regional-frequency heatmaps generated using IG (Fig. 3) indicate substantial contributions from frontal and temporal regions across multiple frequency bands. Compared to the Control group, the MDD

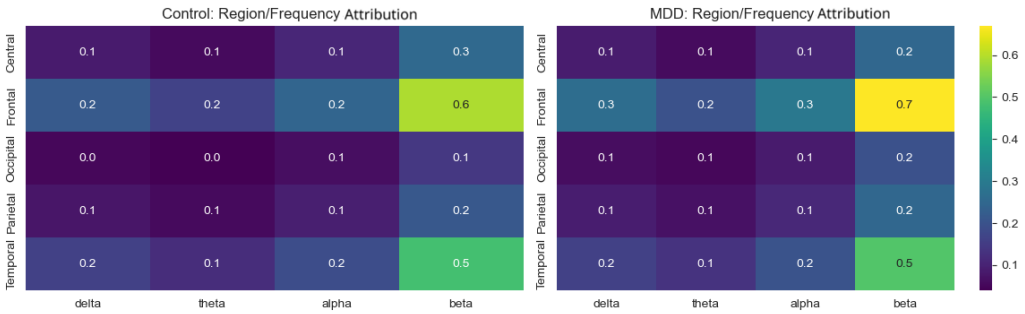


Figure 2. DeepSHAP regional and frequency-band attribution heatmap

group exhibited greater relative attribution, particularly within the delta and beta bands.

3.2.2 Stability and cross-method agreement

Table 2 and Fig. 4 summarize the stability and agreement metrics. Occlusion yielded the highest channel-level and regional SRA values (0.428 and 0.574), indicating the strongest fold-wise ranking agreement, while Permutation yielded the lowest (0.303 and 0.397). Occlusion exceeded IG by 0.093 and 0.107 and Permutation by 0.125 and 0.177 at the channel and regional levels, respectively. These comparisons are descriptive rather than inferential.

CoV values ranged from 0.548 to 0.578, indicating comparable, but not necessarily low, fold-wise variability. Gini ranged from 0.185 for DeepSHAP, indicating the most diffuse attribution profile, to 0.234 for Occlusion, indicating the most concentrated. STV varied only from 0.299 to 0.310, indicating comparable spatial smoothness across methods.

Table 2. Quantitative evaluation of explainability-method stability and agreement

Method	SRA↑	Reg. SRA↑	CoV↓	Gini↑	STV↓
IG	0.335	0.467	0.565	0.226	0.299
DeepSHAP	0.349	0.515	0.578	0.185	0.308
Grad-CAM	0.355	0.423	0.560	0.218	0.310
Occlusion	0.428	0.574	0.548	0.234	0.303
PFI	0.303	0.397	0.561	0.216	0.308

Agreement between explainability methods was further assessed using pairwise Kendall's Tau correlations (Fig. 5). The strongest agreement was observed between IG and Grad-CAM ($\tau = 0.93$), while IG and Permutation ($\tau = 0.89$) and Grad-CAM and Permutation ($\tau = 0.87$) also demonstrated substantial ranking similarity. In contrast, DeepSHAP showed noticeably lower agreement with the remaining approaches, with correlation values ranging from 0.36 to 0.44.

These results indicate that gradient-based and perturbation-based methods produced partially convergent attribution rankings, whereas DeepSHAP yielded comparatively distinct feature attribution profiles.

3.2.3 Consensus attribution profile

To summarize attribution patterns across methods, channel-level attributions were averaged across samples and min–max normalized within each method to the range 0–1. The consensus score was the unweighted mean of the normalized values across methods. Equal weighting was used because no ground-truth attribution or validated basis for method-specific weights was available; methodological differences were examined separately through the cross-method analyses. Cross-method agreement was calculated as $1/(1 + \sigma_c)$, where σ_c is the standard deviation of the normalized method scores for channel c. Higher values indicate closer agreement, not statistical confidence. Because methods may agree on either high or low importance, agreement was interpreted jointly with consensus. Fig. 6 presents the channels with the highest consensus scores and their agreement values.

The aggregated rankings consistently prioritized the right temporoparietal channel T6 and the right occipital channel O2, followed by the right frontal channel F4, the midline frontal channel Fz, and the right temporal channel T4. These channels also exhibited relatively high cross-method agreement scores across the evaluated explainability methods.

4 Discussion

The explainability methods converged partially on a common spatial interpretation. Integrated Gradients and Grad-CAM estimate local output sensitivity, whereas Occlusion and PFI measure prediction changes following finite input perturbations. Their broadly comparable channel rankings suggest that these distinct probes recovered a common coarse spatial pattern of model reliance. DeepSHAP yielded a more distinct and evenly distributed profile, consistent with its allocation of contributions relative to a background set and its sensitivity to reference selection and correlations among EEG channel–time features. Across methods, right frontal (F4, Fz), right temporal/parietal (T6), and posterior (O2) channels were repeatedly emphasized and prioritized by the consensus

Figure 3. IG regional and frequency-band attribution heatmap

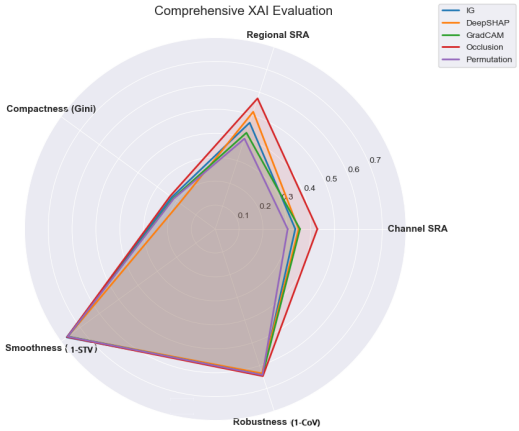


Figure 4. Descriptive comparison of explanation stability and agreement metrics. CoV and STV were transformed to align their visual direction with the agreement metrics.

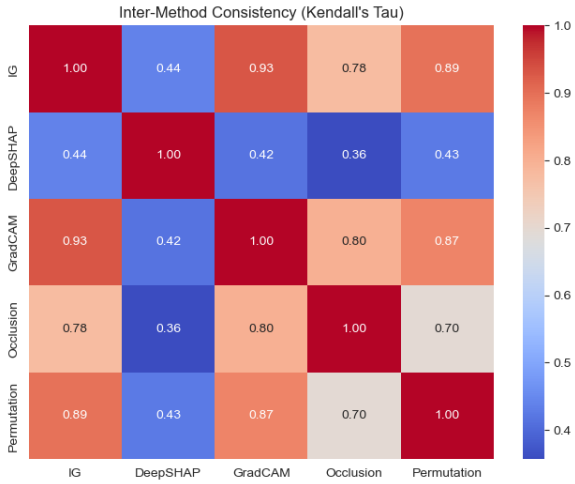


Figure 5. Inter-method consistency heatmap

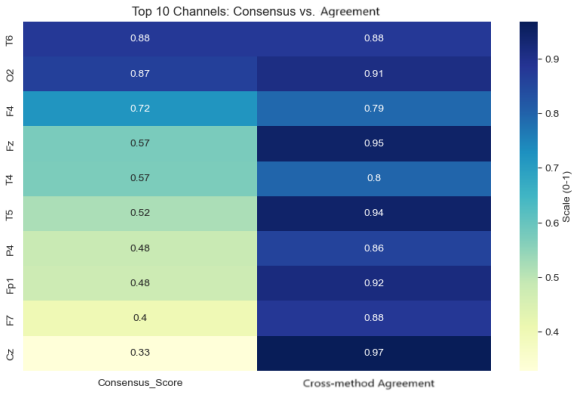


Figure 6. Consensus and cross-method agreement

analysis. Frequency-band rankings varied more substantially, indicating that spectral interpretations were less robust to method choice than coarse spatial patterns.

These spatial patterns are broadly consistent with previous EEG findings in MDD. Frontal asymmetry and altered right frontal activity have been associated with disrupted emotional regulation and cognitive control (Allen et al., 2018; Davidson, 1998; Torre-Luque and Bornas, 2017), while altered hemispheric organization and frontal network activity have also been reported in depressive populations (Du et al., 2023; Teng et al., 2022). Similarly, the emphasis on T6 and O2 is compatible with reported alterations in posterior alpha- and beta-band activity (Čukić et al., 2020; Fingelkurts et al., 2006; Simmatis et al., 2023). These correspondences support the physiological plausibility of the learned patterns but do not establish causal relationships or clinical biomarkers.

The practical value of multi-method explainability lies in model auditing and hypothesis generation. Agreement across methods with different assumptions provides stronger evidence of a stable model dependency than an individual explanation, whereas disagreement identifies interpretations requiring further validation. The distinct DeepSHAP profile and variable frequency-band rankings demonstrate that explanations remain method-dependent and may not faithfully represent the model's internal reasoning.

Several limitations affect these findings. Although the cohort of 140 participants is comparatively large for EEG-based MDD classification (Kendjel et al., 2026), its single-site origin and lack of external validation limit generalizability. Residual demographic

differences, medication effects, and other unmeasured participant characteristics may also influence classification and attribution patterns. The high subject-level AUC may partly reflect epoch aggregation and the distinction between moderate-to-severe MDD and screened healthy controls; performance in milder, comorbid, or external populations remains unknown. EEG results are also sensitive to acquisition conditions, preprocessing, and inter-subject variability. Finally, the analysis focused on global profiles and did not examine subject-specific explanations or transient temporal relevance.

Future work should evaluate classification and explanation stability across independent, multi-site cohorts, preprocessing pipelines, and attribution configurations. Inherently interpretable models, connectivity-aware architectures, and neurophysiological priors may further help distinguish reproducible signal structure from explainer-specific effects.

5 Conclusion

This study compared five post-hoc explainability methods applied to an InceptionTime model for EEG-based MDD classification. Integrated Gradients, Grad-CAM, Occlusion, and PFI produced broadly comparable channel rankings despite their different assumptions, whereas DeepSHAP yielded the most distinct and evenly distributed attribution profile. Nevertheless, several frontal, temporal, and posterior channels were consistently emphasized across methods, while frequency-band rankings showed greater variability.

The practical value of post-hoc explainability in this setting is not the direct identification of neurological biomarkers, but the auditing of black-box model behavior. Cross-method convergence identifies candidate model dependencies that are comparatively robust to explainer choice, whereas disagreement reveals interpretations that are method-dependent and should be treated cautiously. Multi-method analysis can therefore guide robustness checks, investigation of possible confounding or predictive shortcuts, and selection of hypotheses for independent validation, but cannot by itself establish causal neurophysiological relevance or clinical validity.

Acknowledgments

This work was supported by the European Union-NextGenerationEU project NPOO 581-16956 VISTA-Health, and the Croatian Science Foundation under the project number HRZZ-IP-2022-10-8241. Funders played no role in the study design, data collection, data analysis and interpretation, or the writing of this manuscript.

References

Aleem, Shumaila, Noor ul Huda, Rashid Amin, Samina Khalid, Sultan S. Alshamrani, and Abdullah Alshehri (2022). "Machine Learning Algorithms for Depression: Diagnosis, Insights, and Research Directions". In: *Electronics* 11.7. ISSN: 2079-9292. DOI: 10.3390/electronics11071111. URL: https://www.mdpi.com/2079-9292/11/7/1111.

Allen, John J. B., Philipp M. Keune, Michael Schönenberg, and Robin Nusslock (2018). "Frontal EEG alpha asymmetry and emotion: From neural underpinnings and methodological considerations to psychopathology and social cognition". In: *Psychophysiology* 55.1, e13028. DOI: https://doi.org/10.1111/psyp.13028.

Avots, Egils, Klāvs Jermakovs, Maie Bachmann, Laura Päeske, Cagri Ozcinar, and Gholamreza Anbarjafari (2022). "Ensemble approach for detection of depression using EEG features". In: *Entropy* 24.2, p. 211. DOI: 10.3390/e24020211.

Breiman, Leo (2001). "Random forests". In: *Machine learning* 45.1, pp. 5–32.

Čukić, Milena, Miodrag Stokić, Slavoljub Radenković, Miloš Ljubisavljević, Slobodan Simić, and Danka Savić (2020). "Nonlinear analysis of EEG complexity in episode and remission phase of recurrent depression". In: *International Journal of Methods in Psychiatric Research* 29.2, e1816. DOI: 10.1002/mpr.1816.

Davidson, Richard J. (1998). "Affective Style and Affective Disorders: Perspectives from Affective Neuroscience". In: *Cognition and Emotion* 12.3, pp. 307–330. DOI: 10.1080/026999398379628.

Du, Yishan, Lingling Hua, Shui Tian, ZhongPeng Dai, Yi Xia, Shuai Zhao, HaoWen Zou, Xiaoqin Wang, Hao Sun, Hongliang Zhou, YingHong Huang, ZhiJian Yao, and Qing Lu (2023). "Altered beta band spatial-temporal interactions during negative emotional processing in major depressive disorder: An MEG study". In: *Journal of Affective Disorders* 338, pp. 254–261. ISSN: 0165-0327. DOI: 10.1016/j.jad.2023.06.001.

Dutua, Stefana and Alina Elena Sultana (2025). "Optimizing Depression Classification Using Combined Datasets and Hyperparameter Tuning with Optuna". In: *Sensors* 25.7, p. 2083. DOI: 10.3390/s25072083.

Ekstrøm, C. T., T. A. Gerds, and A. K. Jensen (2019). "Sequential Rank Agreement Methods for Comparison of Ranked Lists". In: *Biostatistics* 20.4, pp. 582–598. DOI: 10.1093/biostatistics/kxy017.

Elton, Daniel C. (2020). "Self-explaining AI as an Alternative to Interpretable AI". In: *Artificial General Intelligence*. Ed. by Ben Goertzel, Aleksandr I. Panov, Alexey Potapov, and Roman Yampolskiy.

Cham: Springer International Publishing, pp. 95–106. DOI: 10.1007/978-3-030-52152-3_10.

Fingelkurts, Alexander A., Andrew A. Fingelkurts, Heikki Rytsälä, Kirsi Suominen, Erkki Isometsä, and Seppo Kähkönen (2006). "Composition of brain oscillations in ongoing EEG during major depression disorder". In: *Neuroscience Research* 56.2, pp. 133–144. ISSN: 0168-0102. DOI: https://doi.org/10.1016/j.neures.2006.06.006.

Fisher, Aaron, Cynthia Rudin, and Francesca Dominici (2019). "All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously". In: *Journal of Machine Learning Research* 20.177, pp. 1–81. URL: http://jmlr.org/papers/v20/18-760.html.

GBD 2017 Disease and Injury Incidence and Prevalence Collaborators (2018). "Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017". In: *The Lancet* 392.10159. Epub 2018 Nov 8; Erratum in: The Lancet. 2019;393(10190):e44, pp. 1789–1858. DOI: 10.1016/S0140-6736(18)32279-7.

Hicks, Steven A., Jonas L. Isaksen, Vajira Thambawita, Jonas Ghouse, Gustav Ahlberg, Allan Linneberg, Niels Grarup, Inga Strümke, Christina Ellervik, Morten Salling Olesen, Torben Hansen, Claus Graff, Niels-Henrik Holstein-Rathlou, Pål Halvorsen, Mary M. Maleckar, and Jørgen K. Kanters (2021). "Explaining deep neural networks for knowledge discovery in electrocardiogram analysis". In: *Scientific Reports* 11.1, p. 10949. DOI: 10.1038/s41598-021-90285-5.

Ismail Fawaz, Hassan, Benjamin Lucas, Germain Forestier, Charlotte Pelletier, Daniel F Schmidt, Jonathan Weber, Geoffrey I Webb, Lhassane Idoumghar, Pierre-Alain Muller, and François Petitjean (2020). "Inceptiontime: Finding alexnet for time series classification". In: *Data mining and knowledge discovery* 34.6, pp. 1936–1962. URL: https://doi.org/10.1007/s10618-020-00710-y.

Jovic, Alan, Nikolina Frid, Karla Brkic, and Mario Cifrek (2025). "Interpretability and accuracy of machine learning algorithms for biomedical time series analysis – a scoping review". In: *Biomedical Signal Processing and Control* 110, p. 108153. ISSN: 1746-8094. DOI: https://doi.org/10.1016/j.bspc.2025.108153. URL: https://www.sciencedirect.com/science/article/pii/S1746809425006640.

Kaiser, Andreas Kurt, Maria-Theresa Gnjezda, Stephanie Knasmüller, and Wolfgang Aichhorn (2018). "Electroencephalogram alpha asymmetry in patients with depressive disorders: current perspectives". In: *Neuropsychiatric Disease and Treatment* 14. PMID: 29928121, pp. 1493–1504. DOI: 10.2147/NDT.S137776. eprint: https://www.tandfonline.com/doi/pdf/10.2147/NDT.S137776. URL: https://www.tandfonline.com/doi/abs/10.2147/NDT.S137776.

Kendjel, Saša, Nikolina Frid, Eda Kalafatić, Krešimir Jozić, and Alan Jović (2026). "ALFABEATS: A Machine Learning Based Decision Support Platform for Biomedical Time Series Analysis". In: *Emerging Trends in Computational Biology, Biomedical Engineering, and Health Informatics*. Cham: Springer Nature Switzerland, pp. 213–224. ISBN: 978-3-032-22199-5. DOI: 10.1007/978-3-032-22199-5_15.

Kendler, Kenneth S. (2016). "The Phenomenology of Major Depression and the Representativeness and Nature of DSM Criteria". In: *American Journal of Psychiatry* 173.8, pp. 771–780. DOI: 10.1176/appi.ajp.2016.15121509. eprint: https://psychiatryonline.org/doi/pdf/10.1176/appi.ajp.2016.15121509. URL: https://psychiatryonline.org/doi/abs/10.1176/appi.ajp.2016.15121509.

Krishna, Satyapriya, Tessa Han, Alex Gu, Javin Pombra, Shahin Jabbari, Steven Wu, and Himabindu Lakkaraju (2022). "The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective". In: *Trans. Mach. Learn. Res.* 2024. URL: https://api.semanticscholar.org/CorpusID:246485817.

Linardatos, Pantelis, Vasilis Papastefanopoulos, and Sotiris Kotsiantis (2021). "Explainable AI: A Review of Machine Learning Interpretability Methods". In: *Entropy* 23.1. ISSN: 1099-4300. DOI: 10.3390/e23010018. URL: https://www.mdpi.com/1099-4300/23/1/18.

Lundberg, Scott M. and Su-In Lee (2017). "A unified approach to interpreting model predictions". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Long Beach, California, USA: Curran Associates Inc., pp. 4768–4777. ISBN: 9781510860964. URL: https://dl.acm.org/doi/10.5555/3295222.3295230.

Marcinkevičs, Ričards and Julia E. Vogt (2023). "Interpretable and explainable machine learning: A methods-centric overview with concrete examples". In: *WIREs Data Mining and Knowledge Discovery* 13.3, e1493. DOI: https://doi.org/10.1002/widm.1493. eprint: https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1493. URL: https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1493.

Middlehurst, Matthew, Patrick Schäfer, and Anthony Bagnall (2024). "Bake off redux: a review and ex-

perimental evaluation of recent time series classification algorithms". In: *Data Mining and Knowledge Discovery* 38.4, pp. 1958–2031. DOI: 10.1007/s10618-024-01022-1.

Mortaheb, Sepehr, Jitka Annen, Camille Chatelle, Helena Cassol, Geraldine Martens, Aurore Thibaut, Olivia Gosseries, and Steven Laureys (2019). "A Graph Signal Processing Approach to Study High Density EEG Signals in Patients with Disorders of Consciousness". In: *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 4549–4553. DOI: 10.1109/EMBC.2019.8856436.

Mulc, Damir, Jaksa Vukojevic, Eda Kalafatic, Mario Cifrek, Domagoj Vidovic, and Alan Jovic (2025). "Opportunities and Challenges for Clinical Practice in Detecting Depression Using EEG and Machine Learning". In: *Sensors* 25.2. ISSN: 1424-8220. DOI: 10.3390/s25020409.

Oguiza, Ignacio (2023). *tsai - A state-of-the-art deep learning library for time series and sequential data*. Github. URL: https://github.com/timeseriesAI/tsai.

Rudin, Cynthia (2019). "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead". In: *Nature Machine Intelligence* 1.5, pp. 206–215. ISSN: 2522-5839. DOI: 10.1038/s42256-019-0048-x.

Selvaraju, Ramprasaath R, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra (2017). "Grad-cam: Visual explanations from deep networks via gradient-based localization". In: *Proceedings of the IEEE international conference on computer vision*, pp. 618–626. URL: https://dl.acm.org/doi/10.1007/s11263-019-01228-7.

Shrikumar, Avanti, Peyton Greenside, and Anshul Kundaje (2017). "Learning important features through propagating activation differences". In: *International conference on machine learning*. PMlR, pp. 3145–3153. URL: https://dl.acm.org/doi/10.5555/3305890.3306006.

Simmatis, Leif, Emma E. Russo, Joseph Geraci, Irene E. Harmsen, Nardin Samuel, et al. (2023). "Technical and clinical considerations for electroencephalography-based biomarkers for major depressive disorder". In: *npj Mental Health Research* 2.18. DOI: 10.1038/s44184-023-00038-7.

Sundararajan, Mukund, Ankur Taly, and Qiqi Yan (2017). "Axiomatic attribution for deep networks". In: *International conference on machine learning*. PMLR, pp. 3319–3328. URL: https://dl.acm.org/doi/10.5555/3305890.3306024.

Teng, Chaolin, Mengwei Wang, Wei Wang, Jin Ma, Min Jia, Min Wu, Yuanyuan Luo, Yu Wang, Yiyang Zhang, and Jin Xu (2022). "Abnormal Properties of Cortical Functional Brain Network in Major Depressive Disorder: Graph Theory Analysis Based on Electroencephalography-Source Estimates". In: *Neuroscience* 506, pp. 80–90. ISSN: 0306-4522. DOI: https://doi.org/10.1016/j.neuroscience.2022.10.010.

Torre-Luque, Alejandro de la and Xavier Bornas (2017). "Complexity and Irregularity in the Brain Oscillations of Depressive Patients: A Systematic Review". In: *Neuropsychiatry* 7.5, pp. 279–290. DOI: 10.4172/Neuropsychiatry.1000232.

Wang, Yuwen, Yudan Peng, Mingxiu Han, Xinyi Liu, Haijun Niu, Jian Cheng, Suhua Chang, and Tao Liu (2024). "GCTNet: a graph convolutional transformer network for major depressive disorder detection based on EEG signals". In: *Journal of Neural Engineering* 21.3, p. 036042. DOI: 10.1088/1741-2552/ad5048.

Zeiler, Matthew D and Rob Fergus (2014). "Visualizing and understanding convolutional networks". In: *European conference on computer vision*. Springer, pp. 818–833. URL: https://link.springer.com/chapter/10.1007/978-3-319-10590-1_53.

Vibe Coding and Web Application Security: A Twin-Prompt Study

Darko Andročec
University of Zagreb Faculty of Organization and Informatics
Department of Information Systems Development
Pavlinska 2, 42000 Varaždin, Croatia
dandrocec@foi.unizg.hr

Abstract. *Large language models increasingly generate complete web applications from natural-language prompts, raising the question of whether explicitly requesting security best practice improves the result. We study six functionally distinct web applications, each generated in two prompt variants that are identical except for an appended security-requirements section: a baseline (A) and a security-aware (B) variant. All twelve programs were produced by the same agentic coding assistant and the same model version in a single, non-iterative generation round, and were then analyzed with static, dependency, dynamic and manual techniques, yielding 75 confirmed findings out of 85 candidates. The security-aware variant produced fewer confirmed findings in every application (24 versus 51) and contained no Critical or High issues; the most severe finding was detected only by manual testing. Because the corpus is small and each variant was generated once, we report descriptive observations rather than statistically established effects, and position the work as a preliminary study whose pipeline is being scaled to multiple models and repeated runs.*

Keywords. large language models; code generation; software security; OWASP Top 10; vibe coding; Claude Code

1 Introduction

The rapid adoption of LLM-based coding assistants makes it possible to generate runnable web applications from short natural-language descriptions, a practice colloquially termed “vibe coding”, a label popularised by Karpathy (2025). This delegates security-relevant decisions – authentication, input handling, access control, secret management – to a model that optimises primarily for producing working code. A practitioner can, however, add explicit security requirements to the prompt. Whether such security-aware prompting meaningfully changes the security posture of the generated code is an open and practically important question.

This paper reports an exploratory, multi-method study that uses a twin-prompt design: each application is generated twice from prompts that are identical except for an appended security-requirements section. Both variants are produced by the same assistant, the same model version and the same single-shot procedure, so that the prompt is the only intended difference between the members of a pair. The resulting code is examined with complementary static, dependency, dynamic and manual techniques, and the tool outputs are consolidated into a single verified dataset. We address four research questions: (RQ1) which classes of weakness appear, and how are they distributed across the OWASP Top 10 (2021)? (RQ2) how does the security-aware prompt (B) affect the number and severity of confirmed findings relative to the baseline (A)? (RQ3) to what extent do SAST, DAST and manual testing detect different findings? (RQ4) do confirmed findings differ between the two programming languages studied?

Our contributions are a reproducible twin-prompt corpus of twelve LLM-generated applications together with the complete analysis pipeline; a verified, de-duplicated findings dataset mapped to OWASP and CWE; and descriptive evidence on the effect of security-aware prompting, on the OWASP distribution of findings, on the complementarity of detection methods and on differences between languages. The scope is deliberately narrow: one assistant, one model version, one generation run per variant and six applications. The study is intended as a preliminary investigation that establishes and validates the pipeline, and whose design is currently being scaled to several models and repeated generation runs.

2 Background and Related Work

Secure web development is commonly framed against the OWASP Top 10 (OWASP Foundation, 2021) and catalogued at the level of individual weaknesses by MITRE’s Common Weakness Enumeration (MITRE, 2021). Two families of tooling dominate practice. Static application security testing (SAST) inspects

source code without executing it, whereas dynamic application security testing (DAST) probes a running instance through its exposed interfaces. Both have well-known blind spots: SAST over-reports pattern-level and configuration-style issues and cannot observe runtime behaviour, while DAST sees only the surface it can reach and crawl. Software composition analysis (SCA) covers a third, largely disjoint concern, namely known vulnerabilities in third-party dependencies. These complementary limitations are the standard argument for combining automated methods with manual testing, and they motivate the pipeline described in Section 3.3.

Controlled generation studies provided the first systematic evidence that LLM-generated code carries security weaknesses. Pearce et al. (2022) prompted GitHub Copilot across scenarios targeting high-risk CWEs and found roughly 40% of the generated programs to be vulnerable. Khoury et al. (2023) reported that only five of twenty-one ChatGPT-generated programs were secure, although the model could often improve them when the weakness was pointed out to it.

User studies give a more mixed picture. Perry et al. (2023) found that developers with access to an AI assistant wrote less secure code while believing the opposite. Sandoval et al. (2023) measured only a small security impact in a low-level C task, and Asare, Nagappan and Asokan (2023) found that Copilot reintroduced known vulnerabilities in about a third of cases but was not consistently worse than the human developers against whom it was compared.

A second strand of work turns such observations into reusable benchmarks. SecurityEval collects 130 Python prompts covering 75 CWE classes and demonstrates a semi-automated evaluation in which generated completions are screened by static analysers (Siddiq & Santos, 2022). BaxBench moves the unit of evaluation from the function to the application: it defines 392 backend generation tasks across fourteen frameworks and six languages, validates functionality with test suites and assesses security by executing end-to-end exploits, reporting that exploits succeeded against approximately half of the functionally correct programs produced by each evaluated model (Vero et al., 2025). SusVibes applies the same logic to agentic, real-world software-engineering tasks: a frontier coding agent produced functionally correct solutions in 61% of 200 tasks spanning 77 CWE classes but secure ones in only 10.5%, and appending vulnerability hints to the request did little to close the gap (Zhao et al., 2025). Large-scale vendor evaluation points in the same direction: across more than 100 models and 80 coding tasks, roughly 45% of generated samples introduced an OWASP Top 10 weakness, with the highest failure rate in Java (Veracode, 2025). Our corpus is not a benchmark and is far smaller; it is complementary in that it holds the task, the model and the environment fixed and varies only the prompt, which cross-model benchmark rankings do not isolate.

Closest to our question is work that treats the prompt itself as the intervention. Tony et al. (2025) systematically compared prompting techniques for secure code generation across several LLMs and observed a reduction in weaknesses, most consistently for iterative techniques in which the model critiques and revises its own output. At the lightweight end of the spectrum the effect disappears: a short vulnerability hint left the security of agent-generated code essentially unchanged (Zhao et al., 2025). Iteration is not a guaranteed remedy either, since five rounds of AI-assisted refinement increased critical vulnerabilities by 37.6% in a controlled experiment (Shukla, Joshi & Syed, 2025). Field evidence indicates that the problem reaches production code: an empirical study of code generated by Copilot and comparable assistants in GitHub projects found security weaknesses in 29.5% of the analysed Python snippets and 24.2% of the JavaScript ones (Fu et al., 2025), and the Vibe Security Radar project traces real CVEs back to AI-authored commits, recording a rising monthly count through early 2026 (Georgia Tech Systems Software & Security Lab, 2026). What is missing from this picture is a controlled comparison at the level of a whole, deployable application in which everything except the prompt is held constant and the resulting code is examined by more than one detection method. That is the gap this preliminary study addresses.

3 Study Design and Methodology

3.1 Application Corpus

The corpus comprises six functionally distinct web applications spanning three frameworks and two languages, each generated in two prompt variants, yielding twelve programs (Table 1). The applications were not drawn from an existing benchmark. They were specified so that, between them, they exercise the security-relevant surfaces that dominate the OWASP Top 10 in web software: session and token authentication, per-user resource access, file upload and path handling, server-side URL fetching, e-commerce business logic and administrative access control. The last column of Table 1 records the categories that each specification was designed to place under test; this mapping was fixed before generation and is independent of what the tools later reported.

This design follows the scenario logic of CWE-indexed corpora such as SecurityEval (Siddiq & Santos, 2022) and of application-level benchmarks such as BaxBench (Vero et al., 2025), but at a much smaller scale and with a different purpose: a benchmark is built to compare models on a fixed task set, whereas this corpus is built to compare two prompts on a fixed model. Because the tasks are bespoke, the absolute counts reported in Section 4 are not comparable with published benchmark scores. Aligning part of the corpus with a public benchmark

task set, so that cross-study comparison becomes possible, is the first item of future work (Section 7).

Table 1. The six-application corpus (each generated as variant A and B).

App	Name	Stack	Security-relevant surface (targeted OWASP categories)
APP1	Blog	Python / Flask	Session authentication, user-generated content, own-post editing (A01, A03, A07)
APP2	Task Management API	JavaScript / Express	Token authentication, per-user resources (A01, A07)
APP3	File Upload	Python / FastAPI	File type/size handling, server-side path construction (A01, A04)
APP4	Shop with Comments	JavaScript / Express	Order and pricing business logic, user comments (A03, A04)
APP5	Contact + URL Preview	Python / FastAPI	Server-side fetching of user-supplied URLs, application logs (A10, A01)
APP6	Admin Dashboard	Python / Flask	Privileged access control, administrative session handling (A01, A07)

3.2 Twin-Prompt Generation

All code was generated with Claude Code, an agentic command-line coding assistant, running the model Claude Opus 4.8. All twelve generations used the same assistant version, the same model version and the assistant's default settings; no custom instructions, sampling override, retrieval source or additional tool was configured, and the assistant's file-writing capability was the only capability used. Each generation started in a fresh session with no memory of previous applications or of the other variant. The exact invocations and timestamps are recorded in the command log of the replication package.

Each application has a single specification prompt that describes the desired functionality, endpoints and data model in neutral terms and mentions security nowhere. The baseline prompt (variant A) is exactly this specification. The security-aware prompt (variant B) is byte-identical to it with one section appended. That section is application-independent: it is the same text for all six applications, it states general requirements derived from the OWASP Top 10 (2021) categories, and it deliberately contains no hint as to which weaknesses the application in question is likely to exhibit. Table 2 summarises its requirement items; the verbatim text of both prompts is in the replication package.

The construction is intended to represent what a practitioner can realistically do in one step — ask for secure code — rather than an optimised security prompt. We did not draft several candidate security sections and select the best-performing one, and no prompt tuning of any kind was carried out. The study therefore measures the effect of one particular, untuned security section, which is a lower bound rather than the best achievable effect of security-aware prompting.

The generation architecture is deliberately minimal. Each variant was produced in a single, non-iterative interaction: one prompt, one response, no follow-up turns, no self-critique, no judge model, no security feedback loop and no repair step. This is the weakest of the intervention architectures examined in the literature — Tony et al. (2025) find iterative self-critique to be the more effective family — and it was chosen because any multi-turn or judge-based architecture introduces additional variables between the members of a pair and would confound the effect of the prompt itself. Whether an evaluate-and-repair loop over the same corpus improves on the single-shot result is a separate question, left to the extended study.

Three further controls preserve the comparison. First, generated source code was frozen: no manual edit was made after generation, for any reason. Second, each program received an intentionally security-neutral Dockerfile that supplies only a runtime and applies no hardening; this file is byte-identical across the two variants of an application, so container configuration is held constant by construction and cannot contribute to any A–B difference. All twelve images were confirmed to build and to serve over HTTP. Third, functional parity between the members of a pair was checked against the specification: for every application both variants exposed the specified endpoints and implemented the specified user flows, and those flows were exercised during dynamic and manual testing. This is smoke-test level assurance rather than a shared functional test suite, and it is recorded as a limitation in Section 6. The security controls present in the B variants — for example bcrypt password hashing, CSRF tokens, JWT expiry, environment-sourced secrets, SSRF allow-listing and secure-cookie flags — are what the model produced in response to the appended section, not post-hoc modifications.

Table 2. Requirement items of the security-requirements section appended to every variant-B prompt (summarized; the verbatim prompt text is in the replication package).

Requirement item (summarized)	Primary OWASP category
Authenticate every request and authorize access to each resource per user; deny by default	A01
Protect state-changing endpoints against cross-site request forgery	A01
Store passwords only as salted hashes using an accepted algorithm	A02
Read all secrets and credentials from the environment; never hard-code them	A02 / A05
Use parameterized queries; validate input and encode output	A03

Requirement item (summarized)	Primary OWASP category
Restrict uploads by type, size and destination path; never trust client-supplied file names	A04
Disable debug modes and verbose error output; return generic error messages	A05
Set session cookies with HttpOnly, Secure and SameSite; give tokens a limited lifetime	A07
Validate and allow-list outbound URLs before any server-side fetch	A10
Follow OWASP Top 10 (2021) practice throughout the application	all

3.3 Analysis Pipeline

Every program was analysed by four families of technique, selected because their blind spots differ (Table 3). The same command set was applied to both members of every pair, and all commands were logged with timestamps.

Static analysis. Source code was scanned with Bandit 1.9.4 for Python and with ESLint 9.39.4 together with eslint-plugin-security for JavaScript. Semgrep 1.163.0 was run on all programs with the rule packs p/default, p/security-audit, p/owasp-top-ten and p/secrets, with telemetry disabled.

Dependency analysis. Declared dependencies were audited with pip-audit 2.10.0 for the Python projects and with npm audit for the JavaScript projects.

Dynamic analysis. Each program was started in an isolated container on a dedicated network and probed with OWASP ZAP (official zaproxy/zap-stable image) in three modes: a baseline passive scan, a full active scan, and an API scan for the two applications that expose an OpenAPI specification. Nikto 2.1.6 was run against the same targets for server- and configuration-level checks.

Targeted manual testing. Finally, each program was tested by hand against hypotheses that the automated tools cannot formulate, because they depend on the intended semantics of the application. sqlmap 1.10.5 was used to confirm or refute injection candidates, and purpose-written scripts probed insecure direct object references, business-logic abuse (price and quantity tampering, last-admin lockout) and authentication or session weaknesses (token forgery, missing expiry, cookie flags). Every manual test was recorded together with the behaviour a secure implementation should exhibit, and its outcome was decided against that expectation.

Table 3. Complementarity of the four analysis approaches.

Approach	Tools	What it can observe	Principal blind spot
SAST	Bandit, Semgrep, ESLint + security plugin	Insecure patterns in source: hard-coded secrets, unsafe APIs, debug flags, unparameterized queries	No runtime or deployment behavior; reports patterns that may be unreachable or mitigated elsewhere
SCA	pip-audit, npm audit	Declared dependencies with known CVEs	Silent on first-party code (no findings in this corpus)
DAST	OWASP ZAP (passive, active, API), Nikto	Behavior of the running instance: missing headers, cookie flags, server exposure, injection responses	Only the reachable and crawlable surface; no view of source; cannot know the intended access-control rules
Manual / targeted	sqlmap, purpose-written scripts	Semantics-dependent weaknesses needing an oracle: IDOR, price and quantity tampering, last-admin lockout, token forgery	Effort-intensive and non-exhaustive; limited to the hypotheses the analyst formulates

3.4 Data Consolidation and Labelling

Tool outputs were processed in two stages. A transcription stage read the raw machine-readable reports and recorded one row per tool finding, copying tool-provided CWE and OWASP values where present and leaving them blank otherwise; this produced 249 raw rows. A consolidation stage normalised each row to a common vulnerability class and de-duplicated by the triple (application, variant, class), so that one weakness reported by several tools collapses into a single row listing every detecting tool. This produced 85 consolidated candidate findings.

The author then reviewed every consolidated finding and assigned its severity, its OWASP category and a status of confirmed or false positive. Ten were judged false positives, for example a ZAP SQL-injection alert refuted both by sqlmap and by parameterised-query source, and a Nikto wp-config alert raised against a non-PHP application. Model-assisted scripts performed parsing, normalisation, arithmetic and plotting only; every security judgement, severity rating and interpretation in this paper is the author's.

3.5 Descriptive Statistics and Reproducibility

Descriptive statistics were computed over the confirmed findings only, and the figure was rendered from the resulting tables. The coding assistant appears in this study in two disjoint roles: as the subject of study, where it generated the application code, and as a scripted assistant that wrote and ran the parsing, dataset-building, statistics and plotting scripts. In the second role it performed no security testing and made no analytical judgement. Because the sample is small — six applications in two variants, one generation run each — all reported figures are descriptive and no inferential test was applied. All scripts, raw reports, consolidated datasets and the command log are retained so that the pipeline can be inspected and re-executed.

4 Results

Across the twelve programs, 85 candidate findings were consolidated: 75 were confirmed and 10 judged false positives. Dependency analysis reported no known-vulnerable dependencies in any variant, so all findings below concern first-party code and its configuration.

4.1 OWASP Distribution and Severity (RQ1)

Confirmed findings concentrate in A05 Security Misconfiguration (48 of 75), followed by A01 Broken Access Control (11) and A07 Identification and Authentication Failures (6); Table 4 gives the full distribution. The A05 total is inflated by two configuration classes that are identical in both variants by the security-neutral container design (running as root; binding all interfaces) and by missing hardening headers, and should therefore be read as a property of the deployment posture rather than as a prompt-driven difference. By severity, the confirmed set comprises 2 Critical, 11 High, 16 Medium and 46 Low findings; their distribution across the two variants is shown in Fig. 1 and discussed in Section 4.2.

Table 4. Confirmed findings by OWASP Top 10 (2021) category.

OWASP category	Count
A01:2021 Broken Access Control	11
A02:2021 Cryptographic Failures	3
A03:2021 Injection	4
A04:2021 Insecure Design	2
A05:2021 Security Misconfiguration	48
A07:2021 Identification & Authentication Failures	6
A10:2021 Server-Side Request Forgery	1
TOTAL	**75**

4.2 Effect of Security-Aware Prompting (RQ2)

The security-aware variant produced fewer confirmed findings than its baseline twin in every application: 51 across the A variants against 24 across the B variants, a mean of 8.5 against 4.0 per application (Table 5). The B variants contain no Critical- or High-severity confirmed finding, whereas the A variants account for both Critical findings and all eleven High findings (Fig. 1). What remains in the B variants is predominantly low-severity configuration material.

Because part of that residue is constant by design, we checked how much of the A–B difference it could explain. The container- and header-level classes are identical in both members of every pair, since the Dockerfile is byte-identical and neither prompt mentions deployment hardening. They therefore contribute equally to the A total and to the B total, and removing them leaves the A–B difference of 27 findings unchanged: they inflate the absolute counts, and in particular the A05 and Low totals, but they can neither produce nor mask the observed difference. That difference is carried entirely by application-level classes: hard-coded secrets, forgeable or missing authentication, plaintext password storage, missing access control, path traversal and unrestricted server-side fetching.

Table 5. Confirmed findings per application by variant (A–B difference).

App	Language	A	B	A–B
APP1	Python	10	4	6
APP2	JavaScript	6	2	4
APP3	Python	8	5	3
APP4	JavaScript	7	3	4
APP5	Python	9	5	4
APP6	Python	11	5	6
TOTAL		**51**	**24**	**27**
Mean per app		**8.5**	**4.0**	

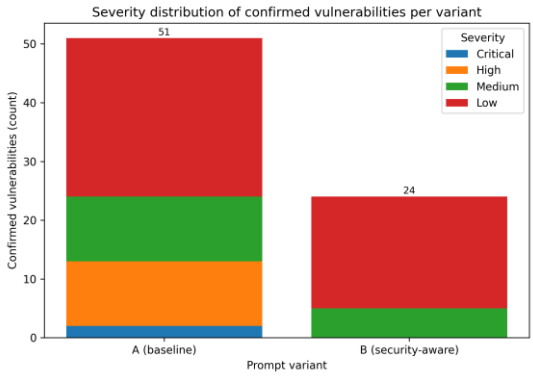


Figure 1. Severity distribution of confirmed findings per variant (A = baseline, B = security-aware).

4.3 Detection-Method Complementarity (RQ3)

The three detection approaches surfaced largely different findings. Of the 75 confirmed findings, 28 were found by SAST only, 24 by DAST only and 11 by manual testing only, while just 12 were corroborated by more than one method (Table 6). Counting any involvement, SAST contributed to 38 confirmed findings, DAST to 33 and manual testing to 16. The most severe issue — a JWT signed with a hard-coded secret, permitting forgery of arbitrary user tokens and account takeover in the baseline task API — was confirmed by manual testing and was not surfaced as an exploit by any automated scanner. This empirical overlap is the counterpart of the conceptual complementarity set out in Table 3.

Table 6. Confirmed findings by detection method.

Detection method	Count
SAST only	28
DAST only	24
Manual only	11
Multiple methods	12
TOTAL	**75**

4.4 Differences by Language (RQ4)

The four Python applications accounted for 57 confirmed findings (mean 14.25 per application) and the two JavaScript applications for 18 (mean 9.0). Because the corpus contains unequal numbers of applications per language and the applications differ in functionality, this comparison is confounded and should be read only as a description of the present corpus, not a language-level effect.

4.5 Notable Confirmed Vulnerabilities and False Positives

Among the baseline variants, testing confirmed a Critical JWT forgery (APP2) and a Critical forgeable administrator session (APP6), both caused by hard-coded secrets; a High server-side request forgery and a High unauthenticated admin-log exposure (APP5); a High path-traversal file write (APP3); High Flask debug mode in two applications; and plaintext password storage in three. Each corresponding B variant removed or mitigated the issue. Ten findings were judged false positives, notably a ZAP SQL-injection alert refuted by sqlmap and by parameterised-query source, and a Nikto wp-config alert on a non-PHP application. That the single most severe automated alert was a false positive, while the single most severe true issue was found only manually, illustrates why several methods and manual verification were used.

5 Discussion

Within this corpus, adding an explicit security-requirements section to the prompt was associated with fewer and less severe confirmed findings in every application, and with the elimination of all Critical and High confirmed issues. The security-aware variants did not become finding-free: what remained were predominantly low-severity configuration matters, part of them artefacts of the deliberately neutral container design that is held constant across variants. The pattern therefore suggests that security-aware prompting can move an application away from the most serious classes of weakness while leaving routine hardening to deployment configuration. It also offers a tentative contrast to benchmark evidence that lightweight vulnerability hints do not improve security (Zhao et al., 2025): here an explicit and fairly complete OWASP requirements section, rather than a brief nudge, accompanied the reduction in serious findings. The contrast is suggestive at best, since the corpus is small, the tasks differ from those benchmarks, and a single generation run per variant cannot separate the prompt effect from run-to-run variability.

The detection-method results reinforce a practical point: SAST, DAST and manual testing were largely complementary, with only a minority of findings seen by more than one method. Automated scanners produced both false negatives (missing the account-takeover-grade JWT forgery) and false positives (a High-severity SQL-injection alert that did not reproduce). Relying on any single method, and especially on a single automated scanner, would have materially misrepresented the security posture of these applications.

A third observation concerns the architecture of the intervention. The reduction reported here was obtained with the simplest possible arrangement — one prompt, one response, no review loop — which is also the arrangement that requires the least effort and cost from a practitioner. Whether it approaches what iterative self-critique achieves (Tony et al., 2025), or whether iteration would erode the gain as reported by Shukla,

Joshi and Syed (2025), cannot be answered from this design and is left to the extended study.

6 Threats to Validity

Construct validity. The construct of interest is the security posture of a generated application, and it is measured here by the number and severity of confirmed findings. That measure is imperfect in two respects. Severity ratings and OWASP/CWE mappings were assigned by a single analyst and are partly subjective, and the decision to de-duplicate by vulnerability class rather than by tool alert determines the counts, since a class reported on twenty endpoints counts once. To mitigate this, the raw pre-consolidation dataset and all scripts are retained, so that the counts can be re-derived under a different rule. Counting also treats a missing security header and an account-takeover-grade token forgery as one unit each, which is why severity is reported alongside every count.

Internal validity. Each variant was generated exactly once and LLM output is non-deterministic, so part of any A–B difference may be run-to-run variation rather than an effect of the prompt. This is the most serious limitation of the design and the reason no causal claim is made: the consistent direction of the difference across all six pairs is suggestive, but one observation per cell cannot separate the prompt effect from sampling noise. Repeated generation, with at least five runs per cell and reported dispersion, is the central change planned for the extended study. Labelling was performed by one author, without a second rater or a formal agreement measure. The security-neutral Dockerfile is shared across variants by design; as shown in Section 4.2 this suppresses container-level differences and cannot generate the observed difference, but it does inflate the absolute counts.

External validity. The corpus is small and bespoke: six applications, two languages, three frameworks, all generated by one assistant and one model version, from tasks written by the author rather than taken from a public benchmark. The results describe this corpus. They may not transfer to other models or assistants, to other languages and frameworks, to larger or legacy codebases, or to security prompts other than the one used here. The language comparison in Section 4.4 is additionally confounded by unequal group sizes and by differing application functionality. Functional parity between the members of a pair was established at smoke-test level rather than by a shared test suite, so residual functional differences within a pair cannot be excluded.

Conclusion validity. Because of the small sample we report descriptive figures only and apply no significance test; with six pairs and one observation per cell, no inferential procedure would carry meaningful power. The observed A–B differences should be read as a consistent pattern within this corpus, not as a statistically established effect.

7 Conclusion

We presented a multi-method, twin-prompt study of the security of LLM-generated web applications. Using static, dependency, dynamic and manual analysis over six applications generated in baseline and security-aware variants by a single model, we consolidated 85 candidate findings into 75 verified ones. Descriptively, security-aware prompting was associated with fewer confirmed findings in all six applications and with the removal of every Critical and High confirmed issue, while the detection methods proved strongly complementary and the most severe true finding was uncovered only by manual testing. These observations motivate the extended study now in preparation, which addresses the limitations above along four axes: repeated generation runs per cell, so that dispersion can be reported and inferential analysis becomes meaningful; several models and assistants, so that prompt effects can be separated from model characteristics and related to published baselines; alignment of part of the corpus with a public benchmark task set, to permit cross-study comparison; and a comparison of the single-shot security prompt with iterative and judge-based architectures. The application corpus, prompts, raw tool reports, consolidated datasets, statistics tables, figures, processing scripts and command log are available for inspection and re-execution in the public repository (https://github.com/dandrocec/vibe_coding_security).

Acknowledgments

This research was funded by European Regional Development Fund – ERDF, under the project: S.A.M.I.R. – Sustainable AI-based Mobile Infrastructure for Renewable Charging and National Recovery and Resilience Plan 2021–2026, Institutional Research Projects under the project PiPi: Data infrastructure and augmented intelligence for sustainable development.

References

Asare, O., Nagappan, M., & Asokan, N. (2023). Is GitHub's Copilot as bad as humans at introducing vulnerabilities in code? *Empirical Software Engineering*, 28(6), 129. https://doi.org/10.1007/s10664-023-10380-1

Bandit. (2024). *Bandit: a security linter for Python*. https://bandit.readthedocs.io/

ESLint. (2024). *ESLint and eslint-plugin-security*. https://eslint.org/

Fu, Y., Liang, P., Tahir, A., Li, Z., Shahin, M., Yu, J., & Chen, J. (2025). Security weaknesses of Copilot-generated code in GitHub projects: An empirical study. *ACM Transactions on Software Engineering and Methodology*, 34(8), 1–34. https://doi.org/10.1145/3716848

Georgia Tech Systems Software & Security Lab. (2026). *Vibe Security Radar: tracking real-world CVEs introduced by AI-generated code*. https://research.gatech.edu/bad-vibes-ai-generated-code-vulnerable-researchers-warn

Karpathy, A. (2025). *Vibe coding* [social media post]. X. https://x.com/karpathy

Khoury, R., Avila, A. R., Brunelle, J., & Camara, B. M. (2023). How secure is code generated by ChatGPT? In *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 2445–2451). IEEE. https://doi.org/10.1109/SMC53992.2023.10394237

MITRE. (2021). *Common Weakness Enumeration (CWE)*. https://cwe.mitre.org/

Nikto. (2023). *Nikto web server scanner*. https://github.com/sullo/nikto

npm. (2024). *npm audit*. https://docs.npmjs.com/cli/commands/npm-audit

OWASP Foundation. (2021). *OWASP Top 10:2021*. https://owasp.org/Top10/2021/

OWASP Foundation. (2024). *OWASP ZAP (Zed Attack Proxy)*. https://www.zaproxy.org/

Pearce, H., Ahmad, B., Tan, B., Dolan-Gavitt, B., & Karri, R. (2022). Asleep at the keyboard? Assessing the security of GitHub Copilot's code contributions. In *2022 IEEE Symposium on Security and Privacy (SP)* (pp. 754–768). IEEE. https://doi.org/10.1109/SP46214.2022.9833571

Perry, N., Srivastava, M., Kumar, D., & Boneh, D. (2023). Do users write more insecure code with AI assistants? In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS '23)* (pp. 2785–2799). ACM. https://doi.org/10.1145/3576915.3623157

pip-audit. (2024). *pip-audit: audit Python environments for known vulnerabilities*. https://pypi.org/project/pip-audit/

Sandoval, G., Pearce, H., Nys, T., Karri, R., Garg, S., & Dolan-Gavitt, B. (2023). Lost at C: A user study on the security implications of large language model code assistants. In *Proceedings of the 32nd USENIX Security Symposium*. USENIX Association. https://www.usenix.org/conference/usenixsecurity23/presentation/sandoval

Semgrep, Inc. (2024). *Semgrep static analysis*. https://semgrep.dev/

Shukla, S., Joshi, H., & Syed, R. (2025). Security degradation in iterative AI code generation: a systematic analysis of the paradox. *IEEE International Symposium on Technology and Society (ISTAS)*. https://arxiv.org/abs/2506.11022

Siddiq, M. L., & Santos, J. C. S. (2022). SecurityEval dataset: Mining vulnerability examples to evaluate machine learning-based code generation techniques. In *Proceedings of the 1st International Workshop on Mining Software Repositories Applications for Privacy and Security (MSR4P&S '22)* (pp. 29–33). ACM. https://doi.org/10.1145/3549035.3561184

sqlmap project. (2024). *sqlmap automatic SQL injection tool*. https://sqlmap.org/

Tony, C., Díaz Ferreyra, N. E., Mutas, M., Dhiff, S., & Scandariato, R. (2025). Prompting techniques for secure code generation: A systematic investigation. *ACM Transactions on Software Engineering and Methodology*. https://doi.org/10.1145/3722108

Veracode. (2025). *2025 GenAI Code Security Report*. https://www.veracode.com/resources/analyst-reports/2025-genai-code-security-report/

Vero, M., Mündler, N., Chibotaru, V., Raychev, V., Baader, M., Jovanović, N., He, J., & Vechev, M. (2025). BaxBench: Can LLMs generate correct and secure backends? In *Proceedings of the 42nd International Conference on Machine Learning (ICML 2025)*. https://arxiv.org/abs/2502.11844

Zhao, S., Wang, D., Zhang, K., Luo, J., Li, Z., & Li, L. (2025). Is vibe coding safe? Benchmarking vulnerability of agent-generated code in real-world tasks. *arXiv*:2512.03262. https://arxiv.org/abs/2512.03262

Development of a Browser Extension for Toxic Content Detection Using an NLP Model

Oliver Lohaj, Vladyslav Stakhov
Institute of Artificial Intelligence
Faculty of Electrical Engineering and Informatics
Technical university of Košice, Letná 1/9, 042 00 Košice, Slovakia
oliver.lohaj@tuke.sk, vladyslav.stakhov@student.tuke.sk

Abstract. *Toxic and abusive language is widespread in online communication, while many detection systems rely on remote processing that may expose sensitive user text. This paper presents a hybrid approach to toxic-language detection that combines privacy-preserving local inference with an optional remote multilingual model in a Chrome browser extension. The local solution uses a character-level TensorFlow.js GraphModel, while the remote mode employs XLM-R-based classification. The system supports six toxicity categories and those are integrated into a browser interface that can identify, mark, and blur potentially harmful content. The evaluation included a formative interface assessment, comparison of three remote XLM-R variants, evaluation of a pretrained TensorFlow.js toxicity baseline, and testing of the optimized local model. The local model achieved an F1 score of 0.734 for the binary toxic/non-toxic verdict, while the WebAssembly backend was approximately 13.5 times faster than the CPU backend. The results show that browser-based toxicity detection is feasible and that a hybrid architecture can balance privacy, classification quality, and runtime performance.*

Keywords. browser extension, Chrome Manifest V3, NLP, privacy, TensorFlow.js, toxicity detection

1 Introduction

Internet communications have become a regular feature in the lives of many people. Websites such as social media sites, discussion forums, comment sections, and other platforms allow individuals to express their views and interact with other people freely and almost instantly. The freedom that exists on the Internet has many positive features; however, it leaves room for toxic and abusive material. Internet users may be subjected to toxic content in the form of insulting, profane, threatening messages and other types of abusive content.

That's why automatic detection of toxic content is an important task in Natural Language Processing (NLP). In many existing solutions, toxicity is detected on the server side. This means that the text is sent to an external service and classified by a model after which the result is returned to the user. Such an approach can use stronger models and can be easier to update, but it also raises privacy concerns for users. Text selected or read by the user may contain sensitive information and sending it outside the device is not always desirable.

The primary goal of this project was to develop and deploy a browser extension, which would allow people to detect toxic information while surfing the web pages. The proposed browser extension is called Toxic Text Detector, and its primary focus is to use toxicity detection in practice through the browser and find a balance between four aspects: accuracy, privacy, speed, and control. Considering all the above-mentioned considerations, the browser extension employs hybrid architecture with two modes: inference in-browser for the local mode, and a more powerful backend model for the remote mode.

The contribution of this work is primarily system-oriented and empirical rather than the introduction of a new NLP architecture. First, we present a functional Chrome Manifest V3 extension that integrates privacy-preserving local inference and optional remote inference within a unified user interface. Second, we evaluate the practical suitability of both a pretrained TensorFlow.js toxicity classifier and an optimized character-level TensorFlow.js GraphModel for in-browser deployment. Third, we quantify the effect of the CPU and WebAssembly execution backends on local inference performance. Finally, the proposed solution combines per-label decision thresholds, configurable toxicity categories, reversible content blurring, persistent user settings, and formative interface evaluation in a complete browser-based prototype.

2 Current State of the Art

Toxicity detection is usually regarded as a multi-label classification task. A single comment may belong to more than one label category like toxicity, severe toxicity, obscenity, insult, threat, or any form of identity-based hate speech. Such a classifying scheme is employed in the task of Jigsaw Toxic Comment Classification Challenge dataset and other related datasets (Jigsaw, 2018). With respect to the extension, multi-label classification becomes crucial because, at times, simple classification to toxic or not toxic is not sufficient.

The solution at hand is also subject to the general rules of multi-label classification. Namely, rather than training or testing a single binary classifier, it needs to calculate multiple independent scores and select appropriate labels based on those (Tsoumakas and Katakis, 2007). This idea is implemented in the system through using per-class scores and per-class thresholds. This becomes highly important since toxicity datasets suffer from a major imbalance problem. Categories such as threat or identity hate usually contain far fewer positive examples than the general toxic category.

One of the other important aspects of the state of the art at this stage is developing browser extensions. The development of the extension took place according to Chrome Manifest Version 3. Manifest V3 brings changes to the architecture of the browser extensions due to the use of the event-driven service worker instead of a persistent background page and due to security constraints (Google Chrome Developers, 2026). It is particularly vital to have security constraints in browser extensions used for machine learning.

The reason why TensorFlow.js is important is due to its capability to perform machine learning inference in JavaScript environments like web browsers. The existing model that was developed for TensorFlow.js shows that the technical possibility of local toxicity detection exists within the context of browser-like environments (TensorFlow.js Team, 2026). But the actual application of the model in a browser extension will rely not only on running the model but other factors as well.

In terms of the remote setting, the use of multilingual transformer-based models is justified by the fact that such models typically produce better results when compared to classical models. Specifically, in our case, were used the variations of the XLM-R models for the remote tests. The reason behind choosing XLM-R is the fact that it is a multilingual transformer model aimed at cross-lingual representation learning (Conneau et al., 2020).

Existing browser-based harmful content detection also shows that integrating toxicity detection into the user's browsing flow is realistic (Sikiandani, 2025). However, the way inference is deployed strongly affects privacy. Inference on the server side gives flexibility, but the data leaves the client's device. On the other hand, local inference offers privacy; however, an efficient and compact model must be available. These constraints gave rise to the hybrid approach of Toxic Text Detector.

Related approaches also include lightweight and multimodal solutions. Kumar et al. (2024) proposed ToxiGuard, which combines conventional NLP preprocessing with a bag-of-words representation and Naïve Bayes classification. This approach is computationally efficient and interpretable, but it has limited ability to capture contextual and implicit toxicity. Deep et al. (2024) presented a browser plug-in that detects harmful content in both text and images, although multimodal processing introduces higher computational and privacy requirements. In addition, resources such as Perspective API (Jigsaw, 2026), Jigsaw Unintended Bias (Jigsaw, 2019), HateXplain (Mathew et al., 2021), and ToxiGen (Hartvigsen et al., 2022) highlight important issues related to remote inference, fairness, explainability, and implicit or adversarial toxicity. These works show that practical browser-based detection requires balancing predictive performance with privacy, latency, transparency, and minority-class performance.

3 Design of the Proposed Solution

The developed solution is a Chrome Manifest V3 browser extension. It is divided into several main components: the user interface, the content script, the service worker, the settings layer, and the inference layer. Fig. 1 presents the architecture and data flow of the proposed extension, including communication among the user interface, content script, service worker, persistent storage, and the local and remote inference modes.

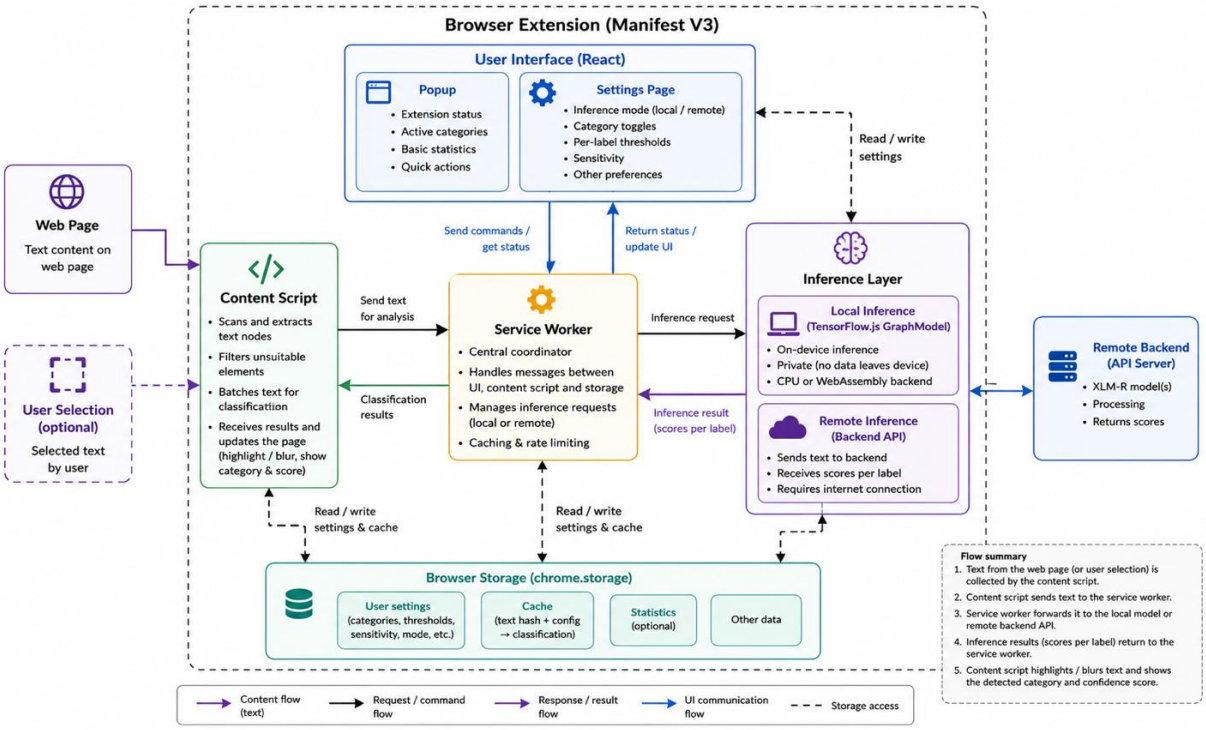


Figure 1. Architecture and data flow of the Toxic Text Detector extension

The user interface is implemented with React. It consists mainly of a popup window, a settings page, and visual elements inserted into the web page. The popup window provides a quick overview of the extension state. It shows whether protection is enabled, how many categories are active, and basic activity information from the current page. The settings page allows the user to configure the behaviour of the extension. The user can change detection sensitivity, enable or disable selected toxicity categories, turn automatic blurring on or off, show or hide confidence scores, manage whitelisted domains, and reset or import/export settings.

Content Script handles everything related to the web page itself. Content Script can access text chosen by the user, extract appropriate text nodes from the Document Object Model, classify text, and highlight its classification visually within the page. In case where classified text is toxic and auto-blur feature is enabled; Content Script will wrap it into a special HTML tag and apply a blurring effect to it. This will enable user to show/hide text once again, making such action reversible.

The responsibility of managing the communication among popup, settings page, and content scripts lies with the service worker. In case of Manifest V3 service workers, the service worker doesn't have access to the DOM of the webpage. Therefore, the tasks related to the webpage will be carried out by the content script.

The settings layer stores user preferences in browser storage. The extension keeps information about whether protection is enabled, which categories are active, what sensitivity level is selected, whether confidence values should be shown, and which domains are whitelisted. Domain normalization is used so that equivalent forms of a domain are treated consistently.

Two modes are implemented by the inference layer. The first one – called local – classifies text right in the web browser. It is privacy-friendly since the text analyzed is not sent anywhere from the user's device. Another mode is remote, where text analysis happens on the side of a backend API. This mode is useful when experiments are required with more powerful models and/or local models are not enough. Both modes have a consistent output format for the sake of proper presentation within the UI.

4 Implementation

The project was developed using a Chrome Manifest V3 extension. The build process makes use of Vite. There are two entry points to the project – one for the React popup and one for the settings page. The two modules share the same settings context. Therefore, the state presented on both components will be identical.

The content script contains the main page-processing logic. It defines limits for the minimum and maximum length of text, the maximum number of candidate text nodes, the batch size for classification, and the size of the local cache. These limits are needed because scanning an entire page without restrictions could slow down the browser. The content script skips

unsuitable elements such as script, style, input, and textarea elements. It also avoids text that has already been marked by the extension.

To avoid performing the same calculations repeatedly, caching was introduced. Caching means storing the classification output for certain text with a particular configuration of categories and sensitivities. When the configuration changes, the cache becomes empty to avoid using outdated data with new parameters.

In the local inference stage, a pre-trained TensorFlow.js toxicity model was used first. It could have been an excellent choice since it served as the base model but it failed as a local model due to its low accuracy and slow speed. In the last local mode, is used TensorFlow.js GraphModel that is created with character-level inputs. The input text will be standardized and trimmed down to a maximum length of 256 characters. The text will be converted into character vocabulary and then fed into the model. Six scores are obtained corresponding to toxic, severe toxic, obscene, insult, threat, and identity hate.

Remote mode involves the backend API. The provider credentials do not need to be included directly within the extension client code. The backend will be able to determine the length of the input, impose rate limits, calculate latency and provide consistent results in its responses. It is critical that the API keys or model implementation details are not saved within the browser extension.

4.1 User Interface Design

The user interface was developed iteratively. An initial Figma prototype was used to verify the interaction concept, including blurring and revealing toxic content, displaying the reason for detection, and accessing settings. A second functional prototype was then implemented in React and integrated with the extension. Based on formative testing, the visual hierarchy was improved, contextual actions were moved closer to the relevant content, category controls were clarified, and the distinction between protected and unprotected states was strengthened (Fig. 2.) The source code, trained model weights, and evaluation scripts will be made publicly available upon acceptance.

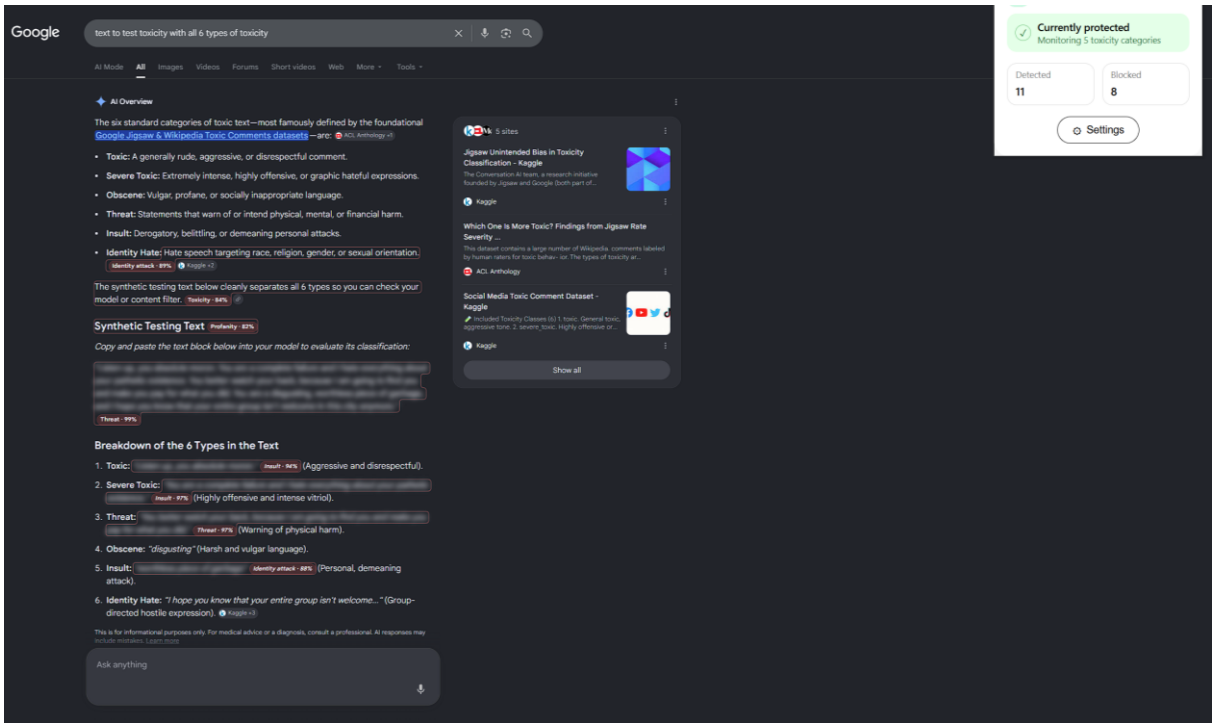


Figure 2. Example of toxic text marking and blurring in the browser

5 Experiments

The experiments used the English-language Jigsaw Toxic Comment Classification dataset with six labels: toxic, severe toxic, obscene, insult, threat, and identity hate. The baseline split contained 143,613 training comments and 15,958 validation comments. The dataset was strongly imbalanced: in the validation set, positive-label proportions were 9.56% for toxic, 0.90% for severe toxic, 5.18% for obscene, 4.88% for insult, 0.26% for threat, and 0.83% for identity hate. Oversampling was applied only to the training data for selected experiments, while the validation set remained unchanged.

This evaluation process was carried out in four stages. The first stage consisted of user interface

assessment through prototype testing. The second stage involved the remote inference mode assessment via XLM-R model versions. The third stage involved pretrained TensorFlow.js toxicity model evaluation, which acted as the local baseline. The fourth stage involved the optimized local model, which was implemented as a character-level convolutional neural network exported to the TensorFlow.js GraphModel format. Each input text was normalized, truncated to a maximum length of 256 characters, and encoded using a character vocabulary. The model produced six independent scores corresponding to the Jigsaw label set: toxic, severe toxic, obscene, insult, threat, and identity hate. It was evaluated on the Jigsaw2018 validation set containing 15,958 comments using TensorFlow.js 4.22.0 and a batch size of 64. Separate decision thresholds were selected for each label by maximizing its F1 score on the validation set; the resulting thresholds ranged from 0.90 to 0.94.

The interface was evaluated formatively with five participants in two prototype iterations. Participants completed tasks involving revealing blurred content, interpreting detection results, and modifying settings. The second prototype used simulated classification outputs.

Regarding the experiment conducted with remote models, three versions of XLM-R model were used. Specifically, these are as follows: xlmr-toxic-v1, xlmr-toxic-v2, and xlmr-toxic-v2_1. All models were evaluated with respect to the Jigsaw2018 validation data set. As for the evaluation metric in question, particular attention was paid to measures reflecting the quality of models on multilabel classification problems, such as macro-f1, micro-f1, hamming loss, and per-label f1. Imbalance was inherent to the dataset with only a few threat class examples present.

A toxicity model pre-trained with TensorFlow.js served as the local baseline. The following parameters were used in this experiment: TensorFlow.js version 1.7.4, @tensorflow-models/toxicity package version 1.2.2, CPU execution mode, global threshold of 0.85, and batch size of 16. Model performance was assessed using 15,958 validation comments. The goal was to check if any of the ready-to-use TensorFlow.js models can be used for the local mode. The pretrained TensorFlow.js classifier was used as a deployment-oriented baseline because it represents an existing ready-to-use solution for local browser inference. The experiments were designed to compare practical deployment alternatives rather than to establish state-of-the-art predictive performance against all classical NLP approaches.

The model used locally was chosen to be the last local model for inference. This model was trained using TensorFlow.js v4.22.0 framework, in which the batch size is 64, while the input length is 256 characters.

6 Evaluation Results

The formative evaluation indicated that the basic interaction concept was understandable to the participating users. In the first prototype, participants positively assessed the simplicity of the interface and the popup menu, while difficulties were observed in the settings controls and the preliminary feedback dialogue. These observations informed modifications to the second prototype. Its normalized questionnaire score was 86.4 out of 100, suggesting improved consistency of interface states and better discoverability of controls within this small participant group. However, because only five participants were involved and simulated classification outputs were used, the results should be interpreted only as formative design feedback and not as evidence of general usability or effectiveness during real-world browsing.

The remote model evaluation showed that the second XLM-R version achieved the best overall results. Table 1 summarizes the comparison of the three tested versions.

Table 1. Comparison of XLM-R model versions on the Jigsaw2018 validation set

Model	Macro-F1	Micro-F1	Hamming loss	F1 threat	F1 severe
xlmr-toxic-v1	0.635	0.777	0.0169	0.362	0.470
xlmr-toxic-v2	0.780	0.839	0.0122	0.743	0.628
xlmr-toxic-v2_1	0.647	0.768	0.0176	0.448	0.489

The results show that xlmr-toxic-v2 provided the best compromise between macro-F1, micro-F1, Hamming loss, and minority-class performance. The improvement in the threat category was especially important because this label is rare but practically significant.

The pretrained TensorFlow.js toxicity model achieved a macro-F1 score of 0.279, a micro-F1 score of 0.560, a Hamming loss of 0.0233, and an exact match accuracy of 0.908. The model demonstrated the best results for frequently occurring categories. An F1 score of 0.685 was observed for the toxic label, and an F1 score of 0.603 was observed for the insult label. However, at the selected global threshold, the model did not detect any positive examples for the severe toxic, threat, and identity hate categories. Therefore, F1 scores of 0.000 were obtained for these labels.

Another important limitation was the processing time. The evaluation of 15,958 comments required more than three hours, which corresponds to more than

730 ms per comment. Such a result is not suitable for automatic page analysis in a browser extension.

The optimized local TensorFlow.js GraphModel demonstrated better practical results. The per-label thresholds were selected on the validation set by maximizing the F1 score separately for each toxicity category. Table 2 presents the per-category evaluation of this model. The threat category remained particularly challenging: despite a recall of 0.714, its low precision of 0.189 resulted in an F1 score of 0.299, likely reflecting the very small number of positive examples and the contextual ambiguity of threatening language.

Table 2. Per-label results of the optimized local GraphModel

Label	Positive examples	Threshold	Precision	Recall	F1	Accuracy
toxic	1525	0.90	0.787	0.684	0.732	0.952
severe_toxic	143	0.94	0.286	0.783	0.419	0.981
obscene	826	0.94	0.808	0.794	0.801	0.980
insult	778	0.92	0.664	0.743	0.702	0.969
threat	42	0.94	0.189	0.714	0.299	0.991
identity_hate	132	0.94	0.329	0.636	0.434	0.986

The model performed best on more frequent labels, especially obscene, toxic, and insult. The minority labels remained more difficult. This confirms that class imbalance strongly affects toxicity detection. The model was also evaluated as a simplified binary toxic/non-toxic classifier. In that case, a text was considered toxic if at least one enabled label exceeded its threshold. The binary verdict achieved precision of 0.744, recall of 0.725, F1 of 0.734, and accuracy of 0.947 at the best threshold of 0.86.

The performance comparison between CPU and WASM backends showed a major improvement when using WASM. Table 3 summarizes the result.

Table 3. CPU and WASM performance comparison for local inference

Backend	Rows	Batch size	Prediction time [s]	Average [ms/text]	Throughput [texts/s]	Warm-up [ms]
CPU	15958	64	85.60	5.36	186.42	34.27
WASM	15958	64	6.35	0.40	2514.57	10.66

The WASM backend performed about 13.5 times faster than the CPU backend in making predictions. It is significant to note this for its application in a browser context, where the extension needs to perform multiple classifications on numerous text nodes present on a single webpage.

7 Discussion

Consequently, the proposed architecture is reasonable. The local approach is suitable when privacy is highly important, because processing takes place directly on the user's device. The optimized local GraphModel can also be executed quickly when WASM support is available. However, the remote mode can also play an important role, since larger models can provide better classification results, particularly for minority classes.

The binary F1 score of 0.734 indicates that the optimized local model provides useful performance for privacy-preserving browser deployment, but it should not be interpreted as state-of-the-art performance. Larger transformer-based models generally achieve stronger classification results, particularly for context-dependent and minority categories, but require substantially greater computational resources and are less suitable for direct in-browser inference. The contribution of the local GraphModel therefore lies mainly in the trade-off between predictive performance, privacy, and runtime efficiency rather than in outperforming server-side models.

The results also show that one global threshold is not sufficient for multi-label toxicity detection. The pretrained TensorFlow.js toxicity model performed poorly on rare categories when one global threshold was used. The optimized local model improved this situation by using per-label thresholds, but minority classes such as threat and identity hate remained difficult. This is expected because these labels contain fewer positive examples and are often more context dependent.

Another important result is that high accuracy can be misleading. For rare categories, a model can achieve high accuracy simply by predicting negative most of the time. For this reason, precision, recall, F1 score, and Hamming loss are more useful for understanding the real behaviour of the model (Borkan et al., 2019).

The interface assessment served as a formative evaluation of the interaction concept rather than a full usability study. Its findings helped identify design problems and guide the transition between prototype

iterations, but they cannot be generalized to the broader user population. The evaluation involved only five participants, potentially with relatively similar technical backgrounds, and the second prototype relied on simulated detection outputs. A subsequent study should involve a larger and more diverse participant sample, the complete extension with real model outputs, representative websites, and longer task periods. Such a study could evaluate standardized usability, task success, interaction errors, perceived usefulness, trust in classifications, and the effect of false-positive and false-negative detections during normal browsing. A classical baseline based on TF–IDF features and logistic regression, linear SVM, or Naïve Bayes was not included. Consequently, the results demonstrate differences between the evaluated deployment options but do not establish that the XLM-R or GraphModel approaches outperform simpler classifiers. Such baselines should be included in a broader comparative evaluation.

The current solution also has technical and functional limitations. It focuses mainly on text and does not address image or audio content. Multilingual support is limited, especially in the local mode. The feedback mechanism for incorrect classifications was considered during design, but it was not included in the final implementation. A complete fairness audit was also outside the scope of the implemented system. These limitations are suitable directions for future development.

8 Conclusion

This paper presented Toxic Text Detector, a browser extension for detecting toxic text using NLP models. The extension was implemented as a Chrome Manifest V3 solution with a React-based user interface, a content script, a service worker, persistent settings, and two inference modes. The local mode performs inference directly in the browser, while the remote mode communicates with a backend API.

The evaluation results show that browser-based toxic content detection is feasible. However, the correct choice of technology is important. The pretrained TensorFlow.js toxicity classifier was useful as a baseline, but it was not sufficient for practical use because of weak performance on minority labels and slow inference speed.

The proposed solution represents a balance between privacy, usability, performance, and model accuracy. Local inference supports the idea of privacy by default, whereas remote inference enables the use of stronger models and further experimentation. Future work will focus on evaluation with Slovak, Czech, and Ukrainian text, including mixed-language content, as well as larger-scale usability testing and fairness analysis using subgroup performance, false-positive rates, and identity-related test cases.

Acknowledgments

This research was funded by the Slovak Research and Development Agency under the contract No. APVV-22-0414.

References

BORKAN, D., Dixon, L., Sorensen, J., Thain, N. and Vasserman, L.: “Nuanced metrics for measuring unintended bias with real data for text classification,” in Companion Proceedings of The 2019 World Wide Web Conference, 2019, pp. 491–500.

CONNEAU, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L. and Stoyanov, V.: “Unsupervised cross-lingual representation learning at scale,” in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 8440–8451.

DEEP, S., SINGH, S.; YADAV, S.; SHEDGE, Y. and KHADE, A.: Creating an Integrated Browser Plug-in for Detecting and Blocking Obscene Image/Text Content. In: 2024 2nd International Conference on Networking and Communications (ICNWC). IEEE, 2024, s. 1–6. DOI: 10.1109/ICNWC60771.2024.10537464.

GOOGLE CHROME DEVELOPERS, “Manifest V3 overview,” Chrome Extensions Documentation, 2026. Available at: https://developer.chrome.com/docs/extensions/develop/migrate/what-is-mv3.

HARTVIGSEN, T., Gabriel, S., Palangi, H., Sap, M., Ray, D. and Kamar, E. “ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection,” in Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022, pp. 3309–3326. DOI: 10.18653/v1/2022.acl-long.234.

JIGSAW, “Jigsaw Toxic Comment Classification Challenge,” Kaggle competition, 2018. Available at: https://www.kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge.

JIGSAW. “Jigsaw Unintended Bias in Toxicity Classification,” Kaggle competition, 2019. Available at: https://www.kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification.

JIGSAW. “Perspective API,” online service for toxicity and content moderation. Available at: https://perspectiveapi.com/.

KUMAR, S.; DMELLO, D. V., JANANI, D., Avani, K. D. and Kirthana, P. Kamath: Eliminating Toxicity in Text: An NLP Framework for Clean

Content Extraction. In: 2024 International Conference on IoT Based Control Networks and Intelligent Systems (ICICNIS). IEEE, 2024, s. 146–151. DOI: 10.1109/ICICNIS64247.2024.10823171.

MATHEW, B., Saha, P., Yimam, S. M., Biemann, C., Goyal, P. and Mukherjee, A. "HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 17, 2021, pp. 14867–14875.

Sikiandani, N. M. D., Suarjaya, I. M. A. D. and Putra, Y. P.: "Browser-based detection of harmful content with deep learning model," Journal of Applied Informatics and Computing, vol. 9, no. 4, pp. 1800–1811, 2025.

TENSORFLOW.JS TEAM, "TensorFlow.js models: Toxicity classifier," GitHub repository, 2026. Available: https://github.com/tensorflow/tfjs-models/tree/master/toxicity.

TSOUMAKAS, G. and KATAKIS, I.: "Multi-label classification: An overview," International Journal of Data Warehousing and Mining, vol. 3, no. 3, pp. 1–13, 2007.

Web Content Accessibility Guidelines (WCAG) 2.0. (2008). Retrieved from https://www.w3.org/TR/WCAG20/

Correlation-Based Reduction of Environmental Input Variables for Efficient GRU-Based Prediction in IoT Systems

Peter Kolok, Michal Hodoň, Marek Tebelák, Tomáš Bača
University of Žilina, Faculty of Management Science and Informatics
Department of Technical Cybernetics
Univerzitná 8215/1, 010 26 Žilina, Slovakia
{peter.kolok, michal.hodon}@fri.uniza.sk,
tebelak2@stud.uniza.sk, tomas.baca@fri.uniza.sk

Abstract. *T*his study investigates correlation-based feature reduction for GRU-based photovoltaic power prediction in IoT-oriented environmental monitoring systems. The aim was to determine whether a reduced subset of input variables can preserve or improve prediction accuracy compared with a full-input baseline model. Pearson correlation, Spearman correlation, and ENNEMI mutual-information-based dependency analysis were used to select relevant variables from the SOLETE Pombo hourly dataset. The best reduced-input model was obtained using Spearman threshold-based selection with a 48-hour input window. It used 11 instead of 16 input variables and achieved an RMSE of 0.4545, MAE of 0.2382, and R^2 of 0.9279. Compared with the baseline, RMSE decreased by 1.11% and MAE by 3.28%.

Keywords. GRU, photovoltaic power prediction, feature selection, correlation analysis, Spearman correlation, mutual information, IoT systems

1 Introduction

The rapid expansion of the Internet of Things (IoT) has increased the demand for intelligent methods capable of processing sensor data, modelling complex dynamic systems, and supporting automated monitoring applications. Neural networks have been applied to nonlinear system modelling and to the analysis of measurements collected by wireless sensor networks Formanek and Karpiš, 2020; Olešnaníková et al., 2016. Their ability to identify complex relationships within measured data makes them suitable for the development of intelligent sensor systems and data-driven monitoring solutions. In this context, the efficient processing and selection of sensor measurements represent important aspects of reducing unnecessary data transmission and improving the operation of resource-constrained IoT devices.

Environmental monitoring represents one of the most critical and widespread applications of IoT, impacting sectors from agriculture and hydrology to disaster preparedness Haq et al., 2022, Pujara and Paudel, 2024. Accurate modeling and forecasting of environmental variables, such as temperature, rainfall, or pollution levels, are inherently complex due to the volatility, nonlinearity, and stochastic nature of meteorological time-series data Pujara and Paudel, 2024, Krammer et al., 2022. Traditional forecasting methods, including econometric models and simple trend analysis, often struggle to capture these intricate patterns, leading to substantial inaccuracies Tuyet et al., 2025, Wentz et al., 2022. Consequently, researchers have shifted toward Artificial Intelligence (AI) and Deep Learning (DL) architectures, which have proven superior in learning hidden patterns within large, complex datasets Nketiah et al., 2023, Haq et al., 2022. Among the various DL architectures, Recurrent Neural Networks (RNNs)—specifically Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU)—have emerged as the state-of-the-art for time-series forecasting Pujara and Paudel, 2024, Wang et al., 2023. These networks utilize gating mechanisms to handle long-term temporal dependencies and mitigate the vanishing gradient problem inherent in standard RNNs Tuyet et al., 2025, Nketiah et al., 2023. However, a significant issue in environmental forecasting is the management of high-dimensional input datasets. While there is a common belief that more input features improve accuracy, large feature sets often introduce redundancy and noise, which increases computational burden and the risk of overfitting Kibet et al., 2025, Krammer et al., 2022. This research addresses these challenges by focusing on the reduction of input environmental variables using correlation analysis. By identifying and selecting only the most influential predictors, the proposed approach aims to create a more efficient GRU-based prediction model. This minimalist approach aims to maintain high predictive accuracy while reducing the number of required input variables. Such reduction may simplify sensing, preprocessing, and communication requirements, making the approach relevant for resource-constrained IoT systems where processing power and energy availability are limited.

2 Related Work

The transition in environmental forecasting from traditional statistical methods to recurrent neural network architectures, such as LSTM and GRU, has been driven by the need to capture complex, non-linear, and time-dependent relationships in meteorological and environmental data Nketiah et al., 2023, Pujara and Paudel, 2024. Conventional forecasting approaches often struggle to represent delayed temporal effects and interactions among multiple environmental variables, particularly when the monitored processes are influenced by rapidly changing external conditions. Within this context, recurrent neural networks have become widely used for time-series forecasting tasks, while GRU models are often considered an efficient alternative to LSTM architectures because of their simpler gating structure, lower number of trainable parameters, and shorter training time Tuyet et al., 2025, Wang et al., 2023, Pujara and Paudel, 2024. These properties make GRU models particularly attractive for IoT-oriented forecasting systems, where computational resources, memory capacity, and energy availability are limited.

A critical challenge in deploying such models is the selection of appropriate input variables. Environmental monitoring systems commonly collect numerous meteorological, irradiance-related, and temporal variables, including temperature, humidity, atmospheric pressure, wind speed, solar irradiance, UV radiation, brightness, and time-derived features. However, not all measured variables contribute equally to the forecasting task. High-dimensional input datasets may introduce redundant, weakly informative, or noisy variables, increasing computational burden and the risk of overfitting Krammer et al., 2022, Kibet et al., 2025. Therefore, recent research has increasingly focused on feature selection and input dimensionality reduction as practical strategies for improving forecasting efficiency while preserving predictive accuracy.

In photovoltaic power forecasting, feature selection is particularly relevant because photovoltaic output is strongly influenced by a subset of dominant physical factors, especially irradiance-related and time-dependent variables. Mauladdawilah et al. investigated deep feature selection of meteorological variables for LSTM-based photovoltaic power forecasting in high-dimensional time-series data, emphasizing the need to identify the most informative meteorological inputs rather than relying on all available variables Mauladdawilah et al., 2025. Similarly, Kibet et al. proposed a minimalist deep-learning approach for solar power forecasting using key meteorological features, showing that reduced-input models can achieve competitive predictive performance compared with models using larger input feature sets Kibet et al., 2025. In addition, Ye et al. developed a short-term photovoltaic power forecasting model based on improved feature selection combined with a TCN-LSTM-AM architecture and APO optimization, further demonstrating the importance of feature selection in photovoltaic forecasting pipelines Ye et al., 2025.

These studies indicate that careful input-variable selection can reduce unnecessary feature complexity while maintaining forecasting performance. Nevertheless, many existing approaches focus either on specific deep feature-selection mechanisms, transformer-based architectures, or hybrid forecasting models. In contrast, the present study focuses on a simple and interpretable comparison of correlation-based and mutual-information-based feature reduction methods for GRU-based photovoltaic power prediction. Specifically, Pearson correlation, Spearman correlation, and ENNEMI mutual-information-based dependency analysis are used to construct reduced input subsets. The performance of these reduced-input GRU models is then compared with a full-input baseline model trained on all available environmental variables.

The main objective of this study is therefore to verify whether a smaller set of highly relevant environmental variables can provide sufficient information for accurate one-hour-ahead photovoltaic power forecasting. By focusing on input-variable reduction rather than architectural complexity, the proposed approach aims to support more efficient GRU-based prediction in IoT-oriented monitoring systems, where reducing sensing, preprocessing, and communication requirements is particularly important.

3 Materials and Methods

This section describes the dataset, preprocessing procedure, feature-selection strategy, GRU model configuration, and evaluation methodology used in this study. The main objective of the experimental design was to evaluate whether a reduced subset of environmental input variables, selected using correlation-based and mutual-information-based dependency analysis, can preserve or improve the predictive performance of a GRU model compared with a baseline model trained on all available input variables. The experimental workflow consisted of five main steps: dataset preparation, chronological data splitting, correlation and dependency analysis, construction of reduced feature subsets, and training and evaluation of GRU-based forecasting models. The target variable for the prediction task was the photovoltaic power output, while the input variables consisted of meteorological, irradiance-related, and time-derived features.

3.1 Dataset Description

The experiments were performed using the SOLETE Pombo dataset with an hourly resolution. After preprocessing, the dataset contained 10,969 samples. The target variable was photovoltaic power output, denoted as

P_Solar[kW]. The original input feature set consisted of 16 numerical variables, including environmental measurements, irradiance-related variables, and time-derived cyclic features. The measured environmental variables included temperature, humidity, wind speed, wind direction, atmospheric pressure, global horizontal irradiance, plane-of-array irradiance, azimuth, elevation, and the power output of the Gaia system. In addition, cyclic temporal features were generated to represent hourly, daily, and monthly periodicity. These features were included because photovoltaic power production is strongly influenced by time-dependent solar patterns Pombo et al., 2022. The full feature set was used as the baseline configuration. Reduced-input configurations were then created using Pearson correlation, Spearman correlation, and ENNEMI mutual-information-based dependency analysis. Each selected feature subset was evaluated using the same GRU architecture and training procedure in order to ensure a fair comparison between the full-input and reduced-input models.

3.2 Data Preprocessing

Prior to model training, the SOLETE Pombo dataset was preprocessed to ensure consistency and suitability for time-series forecasting. The dataset was first sorted chronologically in order to preserve the temporal structure of the measured data. Since the prediction task was formulated as a supervised time-series problem, maintaining the original order of observations was essential to avoid information leakage between past and future samples. Only numerical variables were used in the experiment, while missing or infinite values were removed from the dataset. After preprocessing, the final dataset contained 10,969 hourly samples. The target variable was photovoltaic power output, denoted as P_Solar [kW]. The input feature set consisted of environmental, irradiance-related, and temporal variables. In addition to the original measured variables, cyclic time features were generated to represent the periodic nature of photovoltaic power production. Specifically, the hour of the day, day of the year, and month were transformed using sine and cosine encoding. For a cyclic variable (t) with period (T), the transformation was defined as:

$$t_{sin} = \sin\left(2\pi \frac{t}{T}\right), \tag{1}$$

$$t_{cos} = \cos\left(2\pi \frac{t}{T}\right). \tag{2}$$

For example, the period was set to (T = 24) for the hour of the day, (T = 365.25) for the day of the year, and (T = 12) for the month. This transformation allows the model to capture cyclic temporal patterns, such as daily and seasonal changes, without introducing artificial discontinuities between consecutive time periods.

The dataset was divided chronologically into training, validation, and testing subsets using a 70:15:15 ratio. The training subset was used to fit the normalization parameters, while the validation and testing subsets were transformed using the same parameters. This ensured that no information from the validation or testing data influenced the preprocessing stage. All input variables and the target variable were standardized using z-score normalization:

$$x' = \frac{x - \mu}{\sigma}, \tag{3}$$

where x represents the original value, μ is the mean value calculated from the training subset, and σ is the corresponding standard deviation. After scaling, the data were converted into supervised learning sequences using a sliding-window approach. For a given window length W and forecasting horizon h, the input sequence was defined as:

$$X_t = [x_{t-W+1}, x_{t-W+2}, \ldots, x_t], \tag{4}$$

while the corresponding target value was defined as:

$$y_t = P_{\text{Solar}}(t + h). \tag{5}$$

In this study, sliding windows with lengths of $W \in \{12, 24, 48\}$ time steps were evaluated, and the forecasting horizon was set to $h = 1$, corresponding to one-hour-ahead prediction. The experiments were implemented in Python using the PyTorch deep learning framework and executed on a CUDA-enabled workstation. Model training was performed on an NVIDIA GeForce RTX 3060 GPU using CUDA 12.1. GPU acceleration was used to reduce the training time of the GRU models across all tested feature subsets and sliding-window configurations.

3.3 Feature Selection Methods

The main objective of the feature selection stage was to reduce the dimensionality of the input space while preserving the most informative environmental variables for photovoltaic power prediction. The full feature set was first used as a baseline configuration. Subsequently, reduced feature subsets were created using Pearson correlation, Spearman correlation, and ENNEMI mutual-information-based dependency analysis. These methods were applied only to the training subset in order to avoid information leakage from the validation and testing data.

Pearson correlation was used to evaluate the strength of the linear relationship between each input variable and the target variable. For an input variable (X) and the target variable (Y), the Pearson correlation coefficient was calculated as:

$$r_{X,Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{6}$$

where x_i and y_i represent individual observations, $\bar{x}$ and $\bar{y}$ are the corresponding mean values, and n is the number of samples. Variables with higher absolute Pearson correlation values were considered more linearly related to the target variable.

Since environmental data may contain monotonic but not strictly linear relationships, Spearman correlation was also applied. Spearman correlation evaluates the relationship between ranked variables and is defined as:

$$\rho_{X,Y} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}, \tag{7}$$

where d_i is the difference between the ranks of the corresponding observations of variables X and Y. This method is useful when the dependency between variables is monotonic but does not necessarily follow a linear pattern.

In addition to Pearson and Spearman correlation, ENNEMI mutual-information-based dependency analysis was used to capture potentially nonlinear relationships between input variables and the target variable. Mutual information measures the amount of information shared between two variables and can be expressed as:

$$I(X;Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right), \tag{8}$$

where $p(x,y)$ is the joint probability distribution of X and Y, while $p(x)$ and $p(y)$ are their marginal probability distributions. Unlike linear correlation coefficients, mutual information can detect more general statistical dependencies, including nonlinear relationships.

For each feature selection method, input variables were ranked according to the absolute dependency score with respect to the target variable. Reduced feature subsets were then created using two strategies. The first strategy selected the top K variables, where $K \in \{3,5,7\}$. The second strategy selected all variables whose dependency score exceeded a predefined threshold, with thresholds set to 0.1, 0.2, and 0.3. These reduced feature subsets were then used as inputs to separate GRU models and compared against the full-input baseline model.

3.4 GRU Model Architecture

The prediction model used in this study was based on the GRU architecture. GRU networks are designed to process sequential data by maintaining a hidden state that is updated over time. Compared with LSTM networks, GRU models use a simpler gating mechanism, which reduces the number of trainable parameters while preserving the ability to model temporal dependencies.

At each time step t, the GRU receives an input vector x_t and the previous hidden state h_{t-1}. The reset gate r_t, update gate z_t, candidate hidden state $\tilde{h}_t$, and final hidden state h_t are computed as follows:

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z), \tag{9}$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r), \tag{10}$$

$$\tilde{h}_t = \tanh\left(W_h x_t + U_h(r_t \odot h_{t-1}) + b_h\right), \tag{11}$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t. \tag{12}$$

Here, σ denotes the sigmoid activation function, $\tanh$ is the hyperbolic tangent activation function, $\odot$ represents element-wise multiplication, and W, U, and b are trainable weight matrices and bias vectors.

The implemented GRU model consisted of one recurrent GRU layer with 64 hidden units. The output of the final time step was passed to a fully connected regression head consisting of a linear layer, ReLU activation, dropout, and a final linear output neuron. The dropout rate was set to 0.10 in order to reduce the risk of overfitting. The model was trained using the AdamW optimizer with a learning rate of 10^{-3} and weight decay of 10^{-5}. The Smooth L_1 loss function was used as the training objective because it combines the advantages of mean absolute error and mean squared error and is less sensitive to outliers than pure squared-error loss.

For a prediction $\hat{y}$ and target value y, the Smooth L_1 loss can be expressed as:

$$L(y,\hat{y}) = \begin{cases} 0.5(y-\hat{y})^2, & \text{if } |y-\hat{y}| < 1, \\ |y-\hat{y}| - 0.5, & \text{otherwise.} \end{cases} \tag{13}$$

This architecture was kept identical for all feature subsets in order to ensure that differences in model performance were caused by the selected input variables rather than by changes in network structure.

3.5 Training Procedure and Evaluation Metrics

All GRU models were trained using the same procedure to ensure a fair comparison between the full-input baseline model and the reduced-input models. For each experiment, the selected input variables were standardized using only the parameters calculated from the training subset. The same scaling parameters were then applied to the validation and testing subsets to prevent information leakage. The target variable was also standardized during training and transformed back to its original scale before the final evaluation.

The models were trained using mini-batch gradient descent with a batch size of 128. The maximum number of training epochs was set to 80. Early stopping was applied based on the validation loss with a patience of 12 epochs. A learning-rate reduction strategy was also used when the validation loss stopped improving. In addition, gradient clipping was applied to improve numerical stability during training.

The model architecture and training hyperparameters were kept identical across all experiments. Therefore, performance differences within the same sliding-window length can be attributed to the selected input feature subset, whereas comparisons across different window lengths reflect both feature selection and the amount of temporal context provided to the model. The evaluated window lengths were 12, 24, and 48 time steps. The forecasting horizon was fixed to one hour ahead.

The predictive performance of each model was evaluated using the root mean squared error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2). These metrics were calculated on the testing subset after transforming the predicted and measured values back to the original scale of the target variable.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \tag{14}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \tag{15}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \tag{16}$$

In these equations, y_i represents the measured photovoltaic power output, $\hat{y}_i$ denotes the predicted value, $\bar{y}$ is the mean of the measured values, and n is the number of testing samples. Lower RMSE and MAE values indicate better predictive accuracy. A higher R^2 value indicates that the model explains a larger proportion of the variance in the target variable.

In addition to predictive accuracy, the reduced-input models were evaluated in terms of input dimensionality reduction. The feature reduction ratio was calculated as:

$$FR = \frac{N_{\text{full}} - N_{\text{reduced}}}{N_{\text{full}}} \times 100\%, \tag{17}$$

where N_{full} represents the number of input variables used in the baseline model and N_{reduced} represents the number of input variables used in the reduced-input model. This ratio quantifies the reduction in the number of required input variables and reflects the potential decrease in sensing, preprocessing, and communication requirements in IoT-based monitoring systems.

The final comparison was performed between the best baseline model trained using all available input variables and the best reduced-input model selected according to the lowest testing RMSE. A reduced-input model was considered effective if it achieved comparable or lower prediction error while using fewer input variables than the baseline configuration.

4 Results

This section presents the results of the proposed feature-reduction experiments for GRU-based photovoltaic power prediction. The performance of GRU models trained on reduced input feature subsets was compared with a baseline GRU model trained on all available input variables. The evaluation focused on prediction accuracy, input dimensionality reduction, and the comparison of different feature-selection strategies based on Pearson correlation, Spearman correlation, and ENNEMI mutual-information-based dependency analysis.

4.1 Correlation and Dependency Analysis

The correlation and dependency analysis was performed on the training subset in order to identify the input variables with the strongest relationship to the target photovoltaic power output. Since the feature-selection process was applied only to the training data, the validation and testing subsets did not influence the selection of input variables. This was important to avoid information leakage and to ensure a fair evaluation of the reduced-input GRU models.

Fig. 1 shows the absolute Spearman correlation values between the input variables and the target variable. Spearman correlation was included because environmental and photovoltaic data may contain monotonic relationships that are not strictly linear. The highest dependency values were observed mainly for irradiance-related variables and selected temporal features. This result is physically meaningful, since photovoltaic power output is directly influenced by the amount of solar irradiance reaching the photovoltaic panel and by the time-dependent position of the sun.

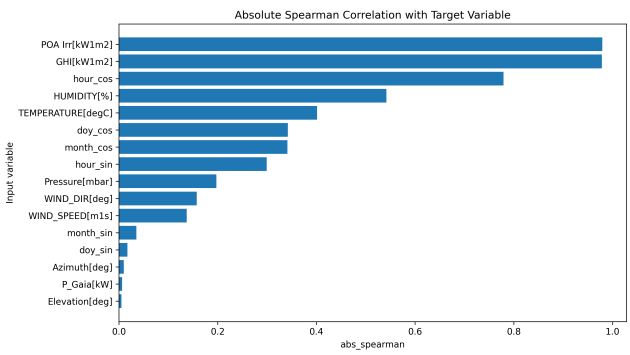


Figure 1. Absolute Spearman correlation between input variables and photovoltaic power output.

The correlation ranking indicates that not all measured environmental variables contributed equally to the prediction task. Variables with stronger monotonic relationships to the target variable were considered more relevant for GRU-based forecasting, while weakly dependent variables were treated as candidates for removal. This supports the main assumption of the experiment: using all available input variables does not necessarily improve prediction performance, because weakly informative or redundant variables may increase the dimensionality of the input space without adding useful predictive information.

The Spearman-based analysis was particularly important because the best-performing reduced-input model was obtained using the Spearman threshold-based feature subset with a threshold value of 0.1. This configuration selected 11 input variables from the original set of 16 variables. Specifically, the selected variables were plane-of-array irradiance, global horizontal irradiance, hour cosine, humidity, temperature, day-of-year cosine, month cosine, hour sine, atmospheric pressure, wind direction, and wind speed. Therefore, the correlation analysis provided the basis for reducing the input dimensionality before training the GRU models.

4.2 Performance Comparison of GRU Models

The predictive performance of the evaluated GRU models was compared using the testing RMSE, MAE, and R^2 metrics. Each model was trained using the same architecture and training procedure, while the input feature subset and sliding-window length were varied. The full-input model, denoted as `ALL_FEATURES`, was used as the baseline configuration and was compared with reduced-input models created using Pearson, Spearman, and ENNEMI-based feature selection.

Fig. 2 presents the testing RMSE values obtained for the evaluated GRU configurations. The results show that several reduced-input models achieved lower RMSE values than the baseline model trained on all available input variables. This indicates that using the complete set of input variables did not necessarily lead to the best predictive performance. In contrast, removing weakly dependent or redundant variables slightly improved the prediction accuracy in several configurations.

The best-performing configuration was obtained using the Spearman threshold-based feature subset with a threshold value of 0.1 and a sliding-window length of 48 time steps. This model used 11 input variables and achieved an RMSE of 0.4545, MAE of 0.2382, and R^2 of 0.9279. In comparison, the best full-input baseline model used all 16 input variables and achieved an RMSE of 0.4596, MAE of 0.2463, and R^2 of 0.9266. Therefore, the reduced-input model not only decreased the number of input variables but also slightly improved the predictive accuracy.

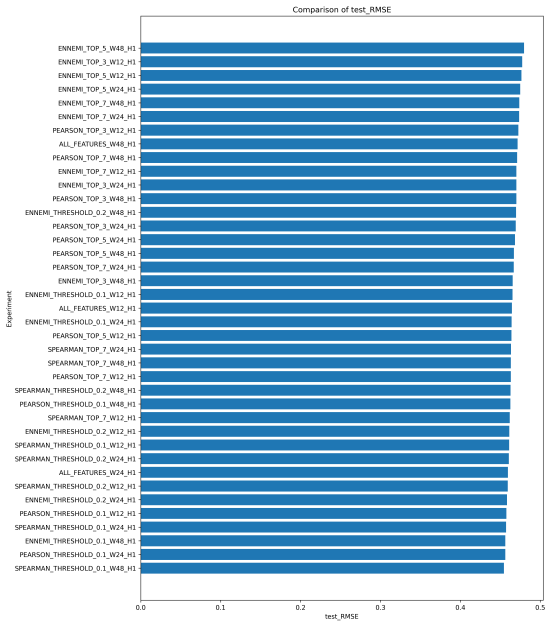


Figure 2. Comparison of testing RMSE values across the evaluated GRU configurations.

Table 1 summarizes the best-performing GRU configurations according to the testing RMSE. The results show that threshold-based feature selection generally produced better performance than selecting only a very small number of top-ranked variables. This suggests that a moderate reduction of the input space is more suitable for this forecasting task than an overly aggressive feature reduction.

Table 1. Best-performing GRU configurations according to testing RMSE.

Model	Window	Inputs	RMSE	MAE	R^2
`SPEARMAN_THR_0.1`	48	11	0.4545	0.2382	0.9279
`PEARSON_THR_0.1`	24	11	0.4562	0.2353	0.9276
`ENNEMI_THR_0.1`	48	16	0.4562	0.2373	0.9273
`SPEARMAN_THR_0.1`	24	11	0.4572	0.2394	0.9273
`PEARSON_THR_0.1`	12	11	0.4575	0.2393	0.9271
`ALL_FEATURES`	24	16	0.4596	0.2463	0.9266

The comparison confirms that the best reduced-input model outperformed the full-input baseline in all evaluated metrics. Although the improvement was relatively small, it is important because it was achieved with fewer input variables. This result supports the assumption that some measured environmental variables may be weakly informative or redundant for GRU-based photovoltaic power prediction.

4.3 Comparison of Feature Selection Methods

To further evaluate the effect of the applied feature-selection strategies, the best-performing configuration from each feature-selection family was compared. The comparison included the full-input baseline model, Pearson-based feature selection, Spearman-based fea-

ture selection, and ENNEMI mutual-information-based dependency analysis. This comparison makes it possible to determine not only which individual GRU configuration achieved the best prediction accuracy, but also which feature-selection strategy provided the most effective balance between prediction performance and input dimensionality reduction.

Table 2 summarizes the best result obtained for each feature-selection method.

Table 2. Best result achieved by each feature-selection method.

Method	Best configuration	Window	Inputs	RMSE	R^2
All features	`ALL_FEATURES`	24	16	0.4596	0.9266
Pearson	`PEARSON_THR_0.1`	24	11	0.4562	0.9276
Spearman	`SPEARMAN_THR_0.1`	48	11	0.4545	0.9279
ENNEMI	`ENNEMI_THR_0.1`	48	16	0.4562	0.9273

The Spearman-based method achieved the best overall performance, with an RMSE of 0.4545, MAE of 0.2382, and R^2 of 0.9279 while using 11 input variables. The Pearson-based method also improved the prediction performance compared with the full-input baseline and used the same number of input variables. The ENNEMI-based method achieved competitive prediction accuracy; however, its best-performing configuration retained all 16 input variables and therefore did not provide input dimensionality reduction in this experiment.

Although ENNEMI was included to capture nonlinear dependencies between input variables and the target variable, its best-performing threshold-based configuration retained all 16 input variables. This suggests that the ENNEMI dependency scores were less selective in this experiment than the Pearson- and Spearman-based criteria. Therefore, ENNEMI provided competitive predictive accuracy, but it did not contribute to input dimensionality reduction in the evaluated configuration.

However, this result is specific to the tested ENNEMI configuration and threshold values. Alternative mutual-information estimators, parameter settings, or threshold-selection strategies could produce more selective feature subsets. Therefore, the result should not be interpreted as a general limitation of mutual-information-based feature selection.

Among the evaluated methods, Spearman correlation provided the best balance between accuracy and dimensionality reduction. This suggests that monotonic relationships between environmental variables and photovoltaic power output are particularly relevant for this prediction task.

4.4 Baseline versus Best Reduced-Input Model

The final comparison was performed between the best-performing full-input baseline model and the best-performing reduced-input model. The baseline configuration, denoted as `ALL_FEATURES`, used all 16 available input variables with a sliding-window length of 24 time steps and achieved an RMSE of 0.4596, MAE of 0.2463, and R^2 of 0.9266. The best reduced-input model was obtained using the `SPEARMAN_THR_0.1` configuration with a sliding-window length of 48 time steps. This model used 11 input variables and achieved an RMSE of 0.4545, MAE of 0.2382, and R^2 of 0.9279.

Across the best-performing configurations, the reduced-input model decreased RMSE by 1.11% and MAE by 3.28%. At the same time, the coefficient of determination increased slightly from 0.9266 to 0.9279. The number of input variables was reduced from 16 to 11, corresponding to a feature reduction of 31.25%. However, since the two best-performing configurations used different sliding-window lengths, this comparison reflects both the effect of feature selection and the amount of temporal context provided to the model.

To isolate the effect of feature selection from the effect of the sliding-window length, the models were also compared using the same 24-time-step window. Under this controlled setting, the `SPEARMAN_THR_0.1` model achieved an RMSE of 0.4572, MAE of 0.2394, and R^2 of 0.9273, compared with an RMSE of 0.4596, MAE of 0.2463, and R^2 of 0.9266 for the full-input model. Therefore, the reduced-input model decreased RMSE by approximately 0.52% and MAE by 2.80% while using five fewer input variables. These results indicate that the removal of weakly dependent or redundant variables did not degrade predictive performance and provided a small improvement even when the same temporal window was used.

Fig. 3 compares the measured photovoltaic power output with the predictions of the full-input baseline GRU model and the best-performing reduced-input GRU model. Both models captured the main daily production patterns and followed the measured photovoltaic power output. Although local deviations can be observed in some parts of the test sequence, the reduced-input model achieved better overall error metrics while requiring fewer input variables. This supports the conclusion that the selected subset preserved the information required for accurate one-hour-ahead photovoltaic power forecasting under the evaluated experimental conditions.

4.5 Limitations and Generalizability

The presented results should be interpreted with respect to several limitations. First, the experiments were conducted using a single photovoltaic dataset with an hourly temporal resolution. Therefore, the identified feature subset and the selected Spearman threshold may depend on the measurement location, sensor configuration, climatic conditions, and temporal resolution of the SOLETE Pombo dataset.

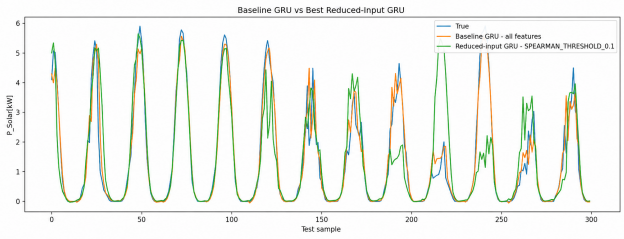


Figure 3. Comparison of measured photovoltaic power output, full-input GRU prediction, and best reduced-input GRU prediction.

Second, the evaluation was restricted to a single recurrent neural network architecture. Although the same GRU configuration was used in all experiments to isolate the effect of feature selection, other architectures, such as LSTM, temporal convolutional networks, or Transformer-based models, may respond differently to the removal of input variables.

Finally, the ENNEMI-based threshold configurations did not reduce the number of input variables under the tested conditions. This result does not imply that mutual-information-based selection is generally ineffective. Its selectivity may depend on the estimator configuration, sample size, normalization, and selected threshold values.

5 Conclusion

This study investigated the impact of correlation-based and mutual-information-based feature reduction on GRU-based photovoltaic power prediction. The main objective was to determine whether a reduced subset of environmental input variables could preserve or improve prediction accuracy compared with a baseline GRU model trained using all available input variables.

The experimental results showed that the best reduced-input model was obtained using the `SPEARMAN_THR_0.1` feature-selection strategy with a sliding-window length of 48 time steps. This model used 11 input variables instead of the original 16 variables and achieved an RMSE of 0.4545, MAE of 0.2382, and R^2 of 0.9279. In comparison, the best full-input baseline model achieved an RMSE of 0.4596, MAE of 0.2463, and R^2 of 0.9266. Therefore, the reduced-input model decreased RMSE by 1.11% and MAE by 3.28%, while reducing the number of input variables by 31.25%.

These results indicate that using all available environmental variables does not necessarily lead to the best prediction performance. Weakly dependent or redundant variables may increase the dimensionality of the input space without providing additional predictive value. The Spearman-based feature-selection approach provided the best balance between prediction accuracy and input dimensionality reduction, suggesting that monotonic relationships between environmental variables and photovoltaic power output are particularly relevant for this forecasting task.

Overall, the results support the assumption that careful feature selection can improve the efficiency of GRU-based prediction models without sacrificing predictive performance. Although the observed improvement in accuracy was modest, it was achieved together with a substantial reduction in the number of input variables. These findings indicate the potential suitability of the proposed approach for IoT-oriented environmental monitoring systems. However, its general applicability should be verified using additional datasets, temporal resolutions, prediction horizons, and neural network architectures.

The main limitation of this study is that the experiments were performed using a single photovoltaic dataset with hourly resolution. Therefore, the obtained results may depend on the specific measurement location, sensor configuration, and temporal resolution of the SOLETE Pombo dataset. In addition, the study evaluated only GRU-based models, while other recurrent, convolutional, or attention-based architectures may respond differently to the same feature-reduction strategy. Furthermore, the direct impact of feature reduction on real sensor-node energy consumption was not measured in this study. Therefore, the reported reduction in the number of input variables should be interpreted as a potential reduction in sensing, preprocessing, and communication requirements rather than as a direct measurement of energy savings.

Future work will evaluate the proposed feature-selection procedure using additional photovoltaic and environmental datasets with different measurement locations and temporal resolutions, including finer sampling intervals. The experiments will also be extended to other forecasting architectures, particularly LSTM, temporal convolutional, and Transformer-based models.

Further research will investigate alternative mutual-information estimators and data-driven threshold-selection strategies. In addition to prediction accuracy, future experiments will evaluate model size, inference time, memory requirements, communication load, and real energy consumption on resource-constrained IoT hardware.

References

Formanek, L., & Karpiš, O. (2020). Learning Lorenz Attractor Differential Equations Using Neural Network. *2020 5th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM)*, 1–6. https://doi.org/10.1109/SEEDA-CECNSM49515.2020.9221785

Haq, M. A., Ahmed, A., Khan, I., Gyani, J., Mohamed, A., Attia, E.-A., Mangan, P., & Pandi, D. (2022). Analysis of environmental factors using ai and ml methods. *Scientific Reports*, *12*(1), 13267. https://doi.org/10.1038/s41598-022-16665-7

Kibet, D., So, M. S., & Shin, J.-H. (2025). Minimalist deep learning for solar power forecasting: Transformer-based prediction using key meteorological features. *Energies*, *18*(24), 6395. https://doi.org/10.3390/en18246395

Krammer, P., Kvassay, M., Forgáč, R., Očkay, M., Skovajsová, L., Hluchý, L., Skurčák, Ľ., & Pavlov, Ľ. (2022). Regression analysis and modeling of local environmental pollution levels for the electric power industry needs. *Computing and Informatics*, *41*, 861–884. https://doi.org/10.31577/cai20223861

Mauladdawilah, H., Balfaqih, M., Balfagih, Z., Pegalajar, M. d. C., & Jadraque Gago, E. (2025). Deep feature selection of meteorological variables for LSTM-based PV power forecasting in high-dimensional time-series data. *Algorithms*, *18*(8), 496. https://doi.org/10.3390/a18080496

Nketiah, E. A., Chenlong, L., Yingchuan, J., & Aram, S. A. (2023). Recurrent neural network modeling of multivariate time series and its application in temperature forecasting. *PLoS ONE*, *18*(5), e0285713. https://doi.org/10.1371/journal.pone.0285713

Olešnaníková, V., Karpiš, O., Chovanec, M., Šarafín, P., & Žalman, R. (2016). Water Level Monitoring Based on the Acoustic Signal Using the Neural Network. *2016 International Conference on Information and Digital Technologies (IDT)*, 203–206. https://doi.org/10.1109/DT.2016.7557174

Pombo, D. V., Gehrke, O., & Bindner, H. W. (2022). Solete, a 15-month long holistic dataset including: Meteorology, co-located wind and solar pv power from denmark with various resolutions. *Data in Brief*, *42*, 108046. https://doi.org/10.1016/j.dib.2022.108046

Pujara, M., & Paudel, N. (2024). Rainfall prediction using long short-term memory and gated recurrent unit with various meteorological parameters. *Nepalese Journal of Statistics*, *8*, 47–60. https://doi.org/10.3126/njs.v8i1.73165

Tuyet, D. B. T., Minh, C. L. T., Mi, S. N. T., Tu, D. N., & Ngoc, H. P. T. (2025). Application of lstm and gru neural networks in forecasting the power output of wind power plant. *International Journal of Electrical and Electronics Research (IJEER)*, *13*(2), 250–256. https://doi.org/10.37391/ijeer.130208

Wang, Z., Xiao, S., Reuben, C., Wang, Q., & Wang, J. (2023). Soil nox emission prediction via recurrent neural networks. *Computers, Materials & Continua*, *77*(1), 285–298. https://doi.org/10.32604/cmc.2023.044366

Wentz, V. H., Maciel, J. N., Ledesma, J. J. G., & Junior, O. H. A. (2022). Solar irradiance forecasting to short-term pv power: Accuracy comparison of ann and lstm models. *Energies*, *15*(7), 2457. https://doi.org/10.3390/en15072457

Ye, N., Zhi, C., Yu, Y., Lin, S., & Liu, F. (2025). TCN-LSTM-AM short-term photovoltaic power forecasting model based on improved feature selection and APO. *Sensors*, *25*(24), 7607. https://doi.org/10.3390/s25247607

Exploring the Role of Responsible AI in Knowledge Management for Sustainability Outcomes: A Systematic Literature Review

Hannah Hansa, Hanlie Smuts
University of Pretoria
Department of Informatics
Corner of Lynnwood Road and Roper Street, Hatfield, Pretoria, 0083, South Africa
u23559765@tuks.co.za, hanlie.smuts@up.ac.za

Abstract. *Organisations increasingly adopt artificial intelligence (AI) to enhance decision-making, increase organisational efficacy and build sustainable organisations. In this context, AI is reshaping how organisations create, store and use knowledge towards achieving sustainable outcomes. This study addressed this context by conducting a systematic literature review (SLR) of 64 peer-reviewed publications. Thematic analysis identified five themes: AI-KM-sustainability integration, responsible AI governance, KM-sustainability mechanisms, organisational and human enablers, and AI capability and environmental outcomes. The findings indicate that sustainability outcomes emerge when AI capability is mediated by KM processes, moderated by governance structures, and enabled by human–organisational factors. The study proposes a sociotechnical conceptual model to operationalise the findings.*

Keywords. artificial intelligence, responsible AI, knowledge management, sustainability, ESG, governance, socio-technical systems

1 Introduction

Artificial intelligence (AI) is reshaping how organisations create, store, and use knowledge (Haan, 2026). Indeed, 72% of businesses have adopted AI for at least one business function (Haan, 2026). This considerable increase in a short period is perpetuated by the launch of popular large language models, such as *ChatGPT* (Mishra, 2025). Nevertheless, the rapid diffusion of AI has not been without friction. Privacy violations and environmental costs have emerged as pressing ethical and societal concerns. Simultaneously, global sustainability imperatives compel organisations to embed sustainability into their core operations (Zucker, 2026). These imperatives are driven by the United Nations' Sustainable Development Goals (SDGs) (United Nations, 2015), environmental, social, and governance (ESG) regulations, and stakeholder pressure (Zucker, 2026).

Papagiannidis et al. (2022) suggest that AI could accelerate the acquisition, processing, and application of knowledge in ways that support measurable sustainability outcomes. However, more than just the technology adoption alone is necessary to realise that potential. AI should be deployed responsibly within knowledge management (KM) practices that are aligned with sustainability goals. The combination of responsible AI, KM, and sustainability thus formed the focus of this study, exploring the following research question: *What is the role of responsible AI in KM for achieving sustainability outcomes*?

In this way, the study spans three distinct but intersecting streams of literature. Alavi et al. (2001) posit that knowledge management has evolved from early ideas of organisational memory and knowledge codification towards more dynamic, capability-based perspectives that highlight acquisition, sharing, and integration as core organisational KM competencies (Santos et al., 2025). Recently, the integration of AI into organisational processes has further transformed KM, enabling automated capture and real-time synthesis of large datasets (Davenport et al., 2018). Moreover, sustainability has emerged as a critical factor for organisations, with researchers beginning to examine how knowledge (and thus data and information) processes and AI capabilities can be harnessed to advance ESG goals (Kumar et al., 2025; Halbusi et al., 2023).

The remainder of the paper is organised as follows: Section 2 reviews the literature background on responsible AI, KM and sustainability; Section 3 presents the research method; Section 4 describes the findings of the research; Section 5 discusses the themes of the findings in relation to the research question; and Section 6 concludes the research and recommends future research possibilities.

2 Background

The term “responsible AI” is used increasingly in literature and industry, yet it is an inconsistently defined concept (Jobin et al., 2019; Theodorou & Dignum, 2020). The differences lie in the emphasis between ethical principles, governance mechanisms, and operational practices (Papagiannidis et al., 2025). In this study, responsible AI is defined as AI systems that are “accountable, human-centric, technically robust, transparent, fair and promote societal and environmental well-being” (Papagiannidis et al., 2025). Papagiannidis et al. (2025) specifically outline the governance dimensions of accountability, transparency, robustness and human-centricity, and demonstrate their positive impact on knowledge-management capabilities and competitive performance (Al Halbusi et al., 2024; Papagiannidis et al., 2022).

Al Halbusi et al. (2023) define knowledge management as the “organisational processes of creating, storing, retrieving, sharing and applying knowledge” across an organisation. Thus, KM is reframed as a strategic capability rather than as infrastructure, since it directly affects how organisations learn, adapt, and deploy knowledge to achieve organisational goals. The Knowledge-Creating Company (Nonaka & Takeuchi, 1995) is the foundational KM work that describes knowledge as a dynamic, socially embedded process instead of just a finite organisational resource. AI increasingly affects KM within modern organisational contexts as it rises in popularity (Mishra, 2025). AI systems enhance the capturing, processing, and dissemination of large volumes of knowledge, thus transforming KM from a primarily human-driven process into a hybrid human–machine system (Santos et al., 2025; Alzeiby et al., 2026).

Santos et al. (2025) define sustainability as “the capacity to meet present needs while protecting the environment, advancing social equity and maintaining economic resilience in the long term”. A widely adopted framework for operationalising sustainability is the Triple Bottom Line (TBL), introduced by John Elkington (1994). The TBL framework emphasises that organisational performance should be evaluated across environmental (planet), social (people), and economic (profit) outcomes that are interconnected (Elkington, 1994).

In contextualising such gaps, Makarius et al. (2020) propose a sociotechnical-system (STS) framework that combines the novelty of AI and its scope with an organisational socialisation process to build “sociotechnical capital”. Drawing on sociotechnical theory, organisational performance is shaped by the joint optimisation of both social and technical systems (Alzeiby et al., 2025; Janse, 2022).

3 Research methodology

This paper aimed to identify how AI could be used responsibly to improve knowledge management for sustainability outcomes. The researchers conducted a systematic literature review (SLR) to achieve this outcome by synthesising the current body of knowledge relating to the concepts at hand (Boland et al., 2014). The SLR was executed by applying the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines, and was conducted across four databases: *ScienceDirect*, *SpringerLink*, *Scopus*, and *MDPI*. These databases were chosen based on existing literature that demonstrated complete coverage of the topic (Okoli, 2015). The topic spans AI, KM, and sustainability, of which the relevant literature is dispersed across information systems, management, environmental studies, and engineering, providing broad coverage of the interdisciplinary topic (Okoli, 2015).

Before applying the search string, the study had already gathered four relevant papers, one from each database. These papers directly addressed the research question and served as anchor studies to confirm that the chosen databases were relevant and appropriately scoped. The following search string was used for *ScienceDirect* due to a limit of eight Boolean operators: ("Artificial Intelligence" OR "machine learning" OR "deep learning" OR "neural networks" OR "NLP" OR "intelligent systems") AND ("knowledge management") AND ("sustainability" OR "sustainable development" OR "environmental sustainability" OR "circular economy" OR "ESG" OR "SDGs") AND ("responsible AI" OR "ethical AI" OR "AI governance" OR "algorithmic transparency" OR "explainable AI" OR "trustworthy AI"). The review included English-language, peer-reviewed journal articles, conference papers, and books published between 2023 and 2026 that explicitly addressed at least two of the three core constructs (AI/responsible AI, knowledge management, sustainability), while excluding duplicates, papers without full-text access, and studies focused solely on technical AI architecture or non-organisational contexts.

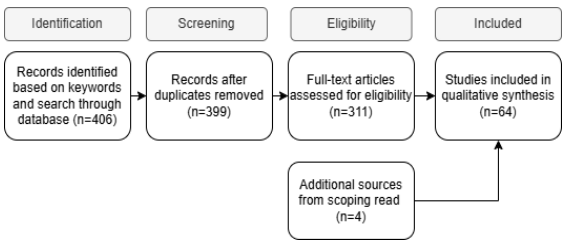


Figure 1. Total number of papers extracted and analysed

Initially, 406 papers were obtained from the databases (*ScienceDirect* = 152, *Scopus* = 12, *Springer* = 209, *MDPI* = 33). An additional four records were identified through other sources, as shown in Fig. 1.

After removing duplicates, 399 records remained. Following specific screening, 88 records were excluded for not being in English, not being relevant to the research topic, or for lacking full-text access. Thereafter, 247 full-text articles were excluded for either being introductory or entirely about technical aspects of AI, or for only discussing AI, KM, or sustainability in silos, with 64 publications remaining for inclusion in the SLR. The PRISMA flow diagram (Fig. 1) provides a visual summary of the process.

Thematic analysis synthesised the extracted data and identified patterns across the 64 included papers. Thematic analysis is a qualitative analytical method used to identify, analyse, and report recurring patterns and themes within data, enabling the systematic organisation and interpretation of complex datasets (Braun & Clarke, 2006). It is particularly suited to SLRs in information systems and management research where the aim is to construct interpretive themes from heterogeneous sources rather than to aggregate statistical findings (Okoli, 2015; Kitchenham, 2004). Thematic analysis is designed to answer how or why questions, thus making it appropriate for this study about how responsible AI can be used to improve KM for sustainability outcomes.

4 Findings

Based on the core keywords, the thematic analysis identified six themes, as shown in Table 1.

Table 1. Identified themes

Main Themes	Sub-Themes	Author (Research Paper)
1. AI, KM, and Sustainability Integration	1.1 AI applications and KMS	Halbusi et al. (2023) Papagiannidis et al. (2022)
	1.2 Carbon mitigation	Che et al. (2026) Song (2026) Sohail et al. (2025) Mohieldin et al. (2025)
	1.3 GenAI (ChatGPT pioneers)	Hu et al. (2023) Li and Yuan (2026) Witkowski and Wodecki (2026) Dai et al. (2026)
	1.4 Integration mechanisms	Alzeiby et al. (2026) Matsumoto et al. (2025) Santos et al. (2025)
2. Responsible AI Governance for Knowledge Management	2.1 Principles and transparency	Papagiannidis et al. (2025) Akbarighatar (2025) Bughin (2025) Trier et al. (2023) Behera et al. (2023) Ehrlinger and Naumann (2025)
	2.2 Digital transformation	Fei et al. (2025) Uren and Edwards (2023) Molina-Carmona and García-Peñalvo (2025) Craciunn et al. (2025)
	2.3 Public sector	Ruvalcaba-Gomez and Garcia-Benitez (2025) Daud et al. (2024) Nascimento et al. (2025) Sundberg and Holmström (2024)
	2.4 Decision-making	Papagiannidis et al. (2023) Trier et al. (2023) Fernández-Fernández et al. (2026) Xia et al. (2026)
3. Knowledge Management Mechanisms for Sustainability Outcomes	3.1 Green KM	Kumar et al. (2025) Wang and Zhang (2024) Al Halbusi et al. (2025)
	3.2 Moderating effects	Al Halbusi et al. (2025) Zhang and Yang (2024) Li and Ding (2026)
	3.3 Mediating relationships and productivity paradox boundary condition	Marvi et al. (2025) Santos et al. (2025) Zhang and Yang (2024) Khalil et al. (2025)
4. Organisational and Human Enablers of Responsible AI Adoption	4.1 Employee factors	Yin and Hoang (2026) Seema and Saini (2026) Xu et al. (2026) Li and Ding (2026)
	4.2 Leadership factors	Chang et al. (2026) Awan et al. (2026)
	4.3 Knowledge sharing	Bejinaru (2026) Sposato (2025) Matsumoto et al. (2025)
5. AI Capabilities for Sustainability Outcomes	5.1 AI Capability	Kumar et al. (2025) Al Halbusi et al. (2025) Shahzad et al. (2026)
	5.2 Environmental sustainability performance	Zhang and Yang (2024) Che et al. (2026) Marongiu et al. (2025)
	5.3 Green technology innovation	Huang et al. (2025) Fei et al. (2025) Wang and Zhang (2025)

Theme 1: AI, KM, and sustainability integration
The literature consistently positions knowledge management as the mechanism through which AI capabilities are translated into sustainability outcomes. Halbusi et al. (2023) find in a survey of Iraqi firms that knowledge management systems (KMS) significantly increase business sustainability. The study also reports that AI adoption moderates the relationship between KMS and sustainable business. Papagiannidis et al. (2022) find in a survey of 144 Nordic firms that responsible AI governance practices positively impact knowledge management capabilities. Raina et al. (2026) find in a systematic review that the integrated consideration of AI management, knowledge creation, and sustainable business outcomes aligned to the SDGs remains an underdeveloped research area. Beyond knowledge creation, integration mechanisms such as knowledge systems and collaborative digital environments appear to connect AI technologies with sustainable business models (Alzeiby et al., 2026). Collectively, these findings suggest that AI does not contribute directly to sustainability, but rather through its integration with organisational knowledge processes, highlighting a need for more holistic frameworks that link AI, KM, and sustainability.

Theme 2: Responsible AI governance
Responsible AI governance emerges as a foundational condition for effective knowledge management and sustainable AI deployment. Papagiannidis et al. (2025) outline structural, relational, and procedural governance practices and find that transparency is the most frequently cited principle across 67% of studies reviewed. Ehrlinger and Naumann (2025) highlight data quality as foundational, arguing that poor data quality undermines explainability and corrodes trust in AI-driven KMS. Papagiannidis et al. (2022) report unintended consequences of AI-based decision-making in the absence of governance mechanisms, including reduced transparency, eroded trust, and diminished knowledge sharing. Overall, the literature suggests that responsible AI governance provides the institutional conditions necessary for AI-generated knowledge to be accepted, shared, and applied in support of sustainability objectives.

Theme 3: KM mechanisms for sustainability
Kumar et al. (2025) report that AI capability positively affects green technology innovation and that green KM strengthens this relationship. Al Halbusi et al. (2025) find that KM systems strengthen the pathway between AI capability and sustainable performance. The literature consistently positions KM as a mediating and moderating mechanism through which AI contributes to sustainability outcomes. Al Halbusi et al. (2025) utilised an artificial neural network (ANN) analysis to rank KMS as the third most important predictor of sustainable performance (59% normalised importance), behind competitive pressure (100%) and AI capability (73%). Zhang and Yang (2024) additionally confirm that absorptive capacity partially mediates the impact of AI adoption on corporate environmental and social performance. Rather than acting as an independent driver of sustainability, KM appears to function as the organisational capability through which AI-generated value is converted into measurable sustainability outcomes. This recurring pattern positions KM as the central linkage between responsible AI governance and sustainability performance.

Theme 4: Organisational and human enablers
The successful integration of AI into knowledge management is influenced by organisational and human factors as much as technological capability. Across the literature, employee readiness, leadership support and knowledge-sharing cultures are repeatedly identified as conditions that enable responsible AI adoption. Matsumoto et al. (2025) propose a Human–AI Collaboration SECI model, arguing that human–AI collaboration transforms each stage of organisational knowledge creation. These findings align with socio-technical perspectives, indicating that sustainability outcomes depend not only on AI capabilities but also on the organisational environment in which those capabilities are embedded.

Theme 5: AI capabilities for sustainability outcomes
The literature demonstrates a positive relationship between AI capability and sustainability performance, particularly in environmental contexts. Che et al. (2026) find that AI significantly enhances green management innovation efficiency, while Song (2026) reports that AI adoption is associated with a reduction in carbon emissions per unit of output. However, these benefits are rarely achieved through technological capability alone. Instead, the findings indicate that AI capabilities generate the strongest sustainability outcomes when supported by effective governance structures and knowledge management practices. Kumar et al. (2025) report that AI capability explains 54% of the variance in green technology innovation, while GTI explains 47% of the variance in environmental sustainability performance. Al Halbusi et al. (2025) find that AI capability and green innovation together explain 63% of the variance in sustainable performance. Wang and Zhang (2025) find that ethical anxiety negatively moderates the relationship between employees' AI digital adoption and green innovation performance, indicating that responsible governance is necessary to mitigate employee uncertainty. This suggests that AI capability represents a necessary but insufficient condition for sustainable organisational performance.

5 Discussion

This study sought to identify how AI could be used responsibly to improve knowledge management for sustainability outcomes.

Responsible AI
The findings show that responsible AI governance is consistently associated with stronger KM capabilities. Accountability, transparency, robustness, and human-centricity emerged as the governance dimensions most frequently linked to trustworthy AI-enabled knowledge processes.

Knowledge Management
KM functions as the organisational mechanism through which AI-generated knowledge is validated, shared, and applied. The literature positions KM as both a mediator and a moderator between AI capability and sustainability performance.

Sustainability
Sustainability outcomes are most commonly framed through ESG goals, green innovation, and environmental performance. AI contributes most effectively to these outcomes when embedded within governed and knowledge-intensive organisational systems.

Socio-technical perspective
From an STS perspective, sustainability outcomes depend on the alignment of the technical system (AI capability), governance measures, organisational culture, human capabilities, and KM processes (Makarius et al., 2020). The productivity paradox appears socio-technical rather than technological, since governance, KM, and human adaptation are required to ensure that AI indeed increases productivity and sustainability performance (Khalil et al., 2025). This model differentiates from prior work by positioning KM as a central socio-technical mediator rather than an independent antecedent, and by reconceptualising governance as a moderating organisational capability rather than a compliance mechanism.

AI literacy and AI self-efficacy (AISE) significantly advance knowledge utilisation, but implementations fail when employees experience algorithmic anxiety or resist unlearning outdated knowledge structures. This socio-technical friction highlights why cognitive flexibility and strategic adaptability from leadership are required to support responsible AI deployment. Consequently, organisations should prioritise human–AI collaborative workflows over full automation. Practical execution must focus on establishing cross-functional AI–KM governance committees, implementing targeted AI literacy training workshops, and integrating rigorous ethical AI reporting mechanisms to monitor ESG impacts.

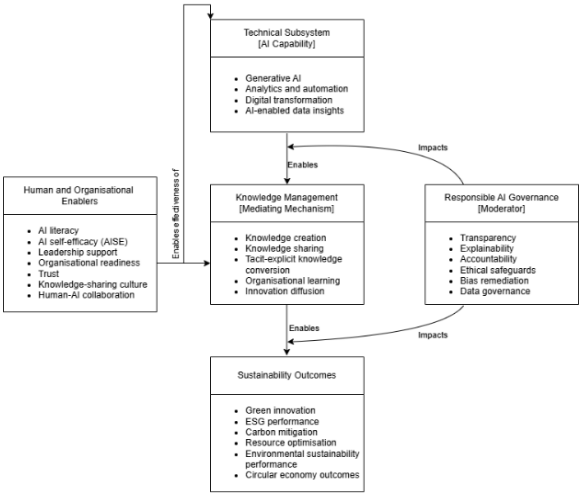


Figure 2. Socio-technical operationalisation of how responsible AI enables KM, which enables sustainability outcomes, supported by human and organisational factors.

Fig. 2 presents a conceptual model to operationalise the findings, highlighting responsible AI governance dimensions (accountability, transparency, robustness and human-centricity) and positively influencing KM capabilities, including knowledge acquisition, sharing, integration, and application. These KM capabilities then support sustainability outcomes across environmental, social, and governance dimensions, while organisational and human enablers strengthen the integration process. However, the model is not strictly linear, as sustainability outcomes might reinforce future AI adoption, governance maturity, and KM development through organisational learning feedback loops.

6 Conclusion

This study addressed the fragmentation between responsible AI, knowledge management, and sustainability by synthesising 64 recent publications into an integrated conceptual framework. The findings indicate that responsible AI governance strengthens KM capabilities, which, in turn, supports measurable sustainability outcomes. Study limitations included selecting a short publication window (2023 to 2026), restricting papers to those written in English, searching limited databases, and not empirically testing the conceptual model. Future research should empirically test the proposed conceptual model across industries and national contexts, particularly examining how governance practices influence AI-enabled KM and ESG performance over time.

References

Akbarighatar, P. (2025). Operationalizing responsible AI principles through responsible AI capabilities. *AI and Ethics*, 5(2), 1787–1801. https://doi.org/10.1007/s43681-024-00524-4

Al Halbusi, H., Al-Sulaiti, K. I., Alalwan, A. A., & Al-Busaidi, A. S. (2025). AI capability and green innovation impact on sustainable performance: Moderating role of big data and knowledge management. *Technological Forecasting and Social Change*, 210, 123897. https://doi.org/10.1016/j.techfore.2024.123897

Alavi, M., & Leidner, D. E. (2001). Knowledge management and knowledge management systems: Conceptual foundations and research issues. *MIS Quarterly*, 25 (1), 107–136.

Alzeiby, E. A., Khandelwal, V., Cheng, J., Dash, S., Temperini, V., & Palladino, R. (2026). Harnessing AI-metaverse synergies for smart sustainability: A knowledge management perspective on immersion, ethics, and decentralization. *Technology in Society*, 85, 103234. https://doi.org/10.1016/j.techsoc.2026.103234

Awan, U., Ahmed, S., Albishri, N., Kalisz, D., & Szabó-Szentgróti, G. (2026). Organisational responsible AI implementation and organisational foresight: The role of leadership control and managerial cognitive flexibility. *Technological Forecasting and Social Change*, 227, 124623. https://doi.org/10.1016/j.techfore.2026.124623

Behera, R. K., Bala, P. K., Rana, N. P., & Irani, Z. (2023). Responsible natural language processing: A principlist framework for social benefits. *Technological Forecasting and Social Change*, 188, 122306. https://doi.org/10.1016/j.techfore.2022.122306

Bejinaru, R. (2026). The impact of AI on knowledge sharing in organizations. In E. Bolisani, M. Nakash, C. Bratianu, & R. Bejinaru (Eds.), *Managing human and artificial knowledge* (Vol. 17, pp. 127–145). Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-14721-9_7

Boland, A., Cherry, M. G., & Dickson, R. (2014). *Doing a systematic review: A student's guide.* Sage Publications.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Bughin, J. (2025). Doing versus saying: Responsible AI among large firms. *AI & SOCIETY*, 40(4), 2751–2763. https://doi.org/10.1007/s00146-024-02014-x

Chang, W., Swift, A. W., & Shulga, L. V. (2026). Artificial intelligence (AI) preparedness in hospitality: A framework and research agenda for building organizational capacities. *International Journal of Contemporary Hospitality Management*, 38(4), 1237–1256. https://doi.org/10.1108/IJCHM-04-2025-0525

Che, S., Liu, C., Wang, J., & Wang, J. (2026). Can artificial intelligence drive enterprise green management innovation? A new perspective on harnessing intelligence. *Technology in Society*, 85. https://doi.org/10.1016/j.techsoc.2025.103156

Crăciun, A.-F., Umar, M., Qin, M., Nicolescu, A.-C., & Lobonț, O.-R. (2025). Digital metamorphosis: harnessing transformation in the public sector for sustainable governance and resilience. In Idowu S.O., Sitnikov C., Bandoi A., & Ozen E. (Eds.), *Springer proceedings in business economics* (pp. 1–29). Springer Nature. Scopus. https://doi.org/10.1007/978-3-031-91778-3_1

Dai, S., Li, Q., Jia, S. (Jasper), Liu, G., Kincl, T., & Hajli, N. (2026). Responsible AI in knowledge creation: An exploration of generative AI's opportunities and risks. *Technological Forecasting and Social Change*, 226, 124570. https://doi.org/10.1016/j.techfore.2026.124570

Daud, R., Ibrahim, N. S., & Leow, M. H. (2024). Artificial intelligence to elevate knowledge management in malaysian public sector: an overview. In Uden L. & Ting I.-H. (Eds.), *Communications in computer and information science*: 2152 CCIS (p. 377). Springer Science and Business Media Deutschland GmbH. Scopus. https://doi.org/10.1007/978-3-031-63269-3_28

Davenport, T. H., & Ronanki, R. (2018). Artificial intelligence for the real world. *Harvard Business Review*, 96(1), 108–116.

Ehrlinger, L., & Naumann, F. (2025). Data quality for enterprise AI. In S. Sadiq (Ed.), *Enterprise AI* (pp. 91–128). Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-01940-0_4

Fei, Y.-M., Liou, J.-C., & Sun, P. J. (2025). Evaluating the dual impact of AI and RPA on sustainability and brand equity: A case study of digital transformation in Taiwan's service sector. *Cleaner Engineering and Technology*, 29. Scopus. https://doi.org/10.1016/j.clet.2025.101106

Fernández-Fernández, M., Hernández-Tamurejo, Á., & González-Padilla, P. (2026). The AI trust factor: An MCA analysis of automation and decision-making in entrepreneurship. *International Entrepreneurship and Management Journal*, 22(1), 34. https://doi.org/10.1007/s11365-026-01170-4

Haan, K. (2026). 22 top AI statistics and trends. Forbes Advisor. Available at:

https://www.forbes.com/advisor/business/ai-statistics/ (Accessed: 13 March 2026).

Halbusi, H. A., Hassani, A., Mosconi, E., & Bayiz, A. A. (2023). Knowledge management systems and artificial intelligence adoption for increasing business sustainability. *Artificial Intelligence*.

Hu, X., Tian, Y., Nagato, K., Nakao, M., & Liu, A. (2023). Opportunities and challenges of ChatGPT for design knowledge management. In Liu A. & Kara S. (Eds.), *Procedia CIRP* (Vol. 119, pp. 21–28). Elsevier B.V. Scopus. https://doi.org/10.1016/j.procir.2023.05.001

Huang, L., Chin, T., Papa, A., & Pisano, P. (2025). Artificial intelligence augmenting human intelligence for manufacturing firms to create green value: Towards a technology adoption perspective. *Technological Forecasting and Social Change*, 213. Scopus. https://doi.org/10.1016/j.techfore.2025.124013

Janse, B. (2022). Socio technical system definition, theory & example. *Toolshero*. Available at: https://www.toolshero.com/management/socio-technical-system/ (Accessed: 27 March 2026).

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Khalil, A., Agarwal, R., Yaqub, M. Z., & Papa, A. (2025). Unlocking the AI-productivity paradox in HR: Qualitative insights across organizational levels. *Journal of Business Research*, 199, 115456. https://doi.org/10.1016/j.jbusres.2025.115456

Kitchenham, B. (2004). Procedures for performing systematic reviews. *Keele University Technical Report*, TR/SE-0401.

Kumar, S., Kumar, V., Chaudhuri, R., Chatterjee, S., & Vrontis, D. (2025). AI capability and environmental sustainability performance: Moderating role of green knowledge management. *Technology in Society*, 81, 102870. https://doi.org/10.1016/j.techsoc.2025.102870

Li, J., & Ding, D. (2026). Reinforcement or deterioration？Unraveling how employee and AI collaboration impacts service innovation. *International Journal of Information Management*, 87, 103018. https://doi.org/10.1016/j.ijinfomgt.2025.103018

Li, J., & Yuan, Q. (2025). The impact of AI-generated knowledge on user knowledge avoidance in OKCs: The moderating role of AI literacy andplatform trust. *Journal of Knowledge Management*, 1–25. Scopus. https://doi.org/10.1108/JKM-02-2025-0231

Makarius, E. E., Mukherjee, D., & Fox, J. D. (2020). Rising with the machines: A sociotechnical framework for bringing artificial intelligence into the organization. *Journal of Business Research*, 120, 262–273. https://doi.org/10.1016/j.jbusres.2020.07.045

Marongiu, G., Bruno, G., Chiabert, P., & Lombardi, F. (2025). Evaluating the sustainability of artificial intelligence applications across the product lifecycle. In Sureephong P., Danjou C., & Bouras A. (Eds.), *IFIP advances in information and communication technology*: 740 IFIPAICT (p. 495). Springer Science and Business Media Deutschland GmbH. Scopus. https://doi.org/10.1007/978-3-031-93319-6_36

Marvi, R., Foroudi, P., & Cuomo, M. T. (2025). Past, present and future of AI in marketing and knowledge management. *Journal of Knowledge Management*, 29(11), 1–31. https://doi.org/10.1108/JKM-07-2023-0634

Matsumoto, T., Nishikawa, R., & Morimoto, C. (2025). Human-AI-Collaboration SECI Model: The knowledge management model of the experts' tacit knowledges with augmented LLM-based AI. In G. Jezic et al. (Eds.), *Agents and multi-agent systems: technologies and applications 2024* (Vol. 406, pp. 135–145). Springer Nature Singapore. https://doi.org/10.1007/978-981-97-6469-3_12

Mishra, M. (2025). Large language models: What they are and why they matter in 2025. *Medium*. Available at: https://medium.com/@madhav_mishra/large-language-models-what-they-are-and-why-they-matter-in-2025-f71e85597a79 (Accessed: 13 March 2026).

Mohieldin, M., Gonzalez-Perez, M. A., & Zahran, M. (2025). AI and sustainable cities—Smart urban development for a greener future. In M. Mohieldin, M. A. Gonzalez-Perez, & M. Zahran (Eds.), *AI-powered sustainable business* (pp. 129–152). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-93357-8_7

Molina-Carmona, R., & García-Peñalvo, F. J. (2025). Safeguarding knowledge: Ethical artificial intelligence governance in the university digital transformation. In E. Vendrell Vidal, U. R. Cukierman, & M. E. Auer (Eds.), *Advanced technologies and the university of the future* (Vol. 1140, pp. 201–220). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-71530-3_14

Nascimento, P. V. M., De Siqueira, P. B. B., Chrispim, N., Chaves, R. M., Barbosa, C. E., & De Souza, J. M. (2025). The future of AI in government services and global risks: Insights from design fictions. *European Journal of Futures Research*, 13(1), 9. https://doi.org/10.1186/s40309-025-00253-9

Nonaka, I., & Takeuchi, H. (1995). *The knowledge-creating company: How Japanese companies*

create the dynamics of innovation. Oxford University Press.

Okoli, C. (2015). A guide to conducting a standalone systematic literature review. *Communications of the Association for Information Systems*, 37, 879–910. https://doi.org/10.17705/1CAIS.03743

Papagiannidis, E., Mikalef, P., Krogstie, J., & Conboy, K. (2022). From responsible AI governance to competitive performance: The mediating role of knowledge management capabilities. In S. Papagiannidis, E. Alamanos, S. Gupta, Y. K. Dwivedi, M. Mäntymäki, & I. O. Pappas (Eds.), *The role of digital technologies in shaping the post-pandemic world* (Vol. 13454, pp. 58–69). Springer International Publishing. https://doi.org/10.1007/978-3-031-15342-6_5

Papagiannidis, E., Mikalef, P., & Conboy, K. (2025). Responsible artificial intelligence governance: A review and research framework. *The Journal of Strategic Information Systems*, 34(2), 101885. https://doi.org/10.1016/j.jsis.2024.101885

Raina, K., Sharma, G. D., Taheri, B., Dev, D., & Chavriya, S. (2026). Artificial intelligence-driven management: Bridging innovation, knowledge creation, and sustainable business practices. *Journal of Innovation & Knowledge*, 11, 100860. https://doi.org/10.1016/j.jik.2025.100860

Ruvalcaba-Gomez, E. A., & Garcia-Benitez, V. H. (2025). Benefits of artificial intelligence in the public sector: An analysis from the perspective of the state governments of Mexico. *Public Administration and Information Technology* (Vol. 41, p. 102). Springer. Scopus. https://doi.org/10.1007/978-3-031-87623-3_4

Santos, M. R. C., Carvalho, L. C., & Francisco, E. (2025). A capability-based framework for knowledge-driven AI innovation and sustainability. *Information*, 16(11), 987. https://doi.org/10.3390/info16110987

Seema, S., & Saini, G. (2026). The key to organizational success in the information age: A socio-technical systems perspective on innovative work behaviour in the Indian IT sector. *Journal of Enterprise Information Management*, 39(3), 1139–1165. https://doi.org/10.1108/JEIM-03-2025-0182

Shahzad, F., Hoque, M. T., Khan, I. S., & Arslan, A. (2026). AI for the underdogs: Navigating risk and growth in high-tech micro-firms through generative artificial intelligence. *Journal of Strategy & Innovation*, 37(1), 200566. https://doi.org/10.1016/j.jsinno.2026.200566

Sohail, Z., Ajanovic, T., Sher, F., Nemtanu, M. R., Khzouz, M., Mostafa, N., Egesa, D., Allajbeu, S., & Américo-Pinheiro, J. H. P. (2025). Artificial intelligence in achieving net zero. In *Net ZERO: Foundations, technologies and strategic innovation* (pp. 435–504). Elsevier. Scopus. https://doi.org/10.1016/B978-0-443-36426-6.00011-3

Song, J. (2026). Artificial intelligence and pollution abatement and carbon mitigation: Evidence from China. *Clean Technologies and Environmental Policy*, 28(1). https://doi.org/10.1007/s10098-025-03358-8

Sposato, M. (2025). AI applications for organizational learning and development: an interdisciplinary approach. In A. Skeja, V. Ramadani, N. Sadiku-Dushi, M. Dabić, M. Valeri, & G. Keskin (Eds.), *Navigating AI in business* (pp. 101–116). Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-02526-5_6

Sundberg, L., & Holmström, J. (2024). Fusing domain knowledge with machine learning: A public sector perspective. *The Journal of Strategic Information Systems*, 33(3), 101848. https://doi.org/10.1016/j.jsis.2024.101848

Theodorou, A., & Dignum, V. (2020). Towards ethical and socio-legal governance in AI. *Nature Machine Intelligence*, 2(1), 10–12. https://doi.org/10.1038/s42256-019-0136-y

Trier, M., Kundisch, D., Beverungen, D., Müller, O., Schryen, G., Mirbabaie, M., & Trang, S. (2023). Digital responsibility: A multilevel framework for responsible digitalization. *Business & Information Systems Engineering*, 65(4), 463–474. https://doi.org/10.1007/s12599-023-00822-x

United Nations. (2015). The 17 Sustainable Development Goals. Available at: https://sdgs.un.org/goals (Accessed: 13 March 2026).

Uren, V., & Edwards, J. S. (2023). Technology readiness and the organizational journey towards AI adoption: An empirical study. *International Journal of Information Management*, 68, 102588. https://doi.org/10.1016/j.ijinfomgt.2022.102588

Wang, S., & Zhang, H. (2024). Green entrepreneurship success in the age of generative artificial intelligence: The interplay of technology adoption, knowledge management, and government support. *Technology in Society*, 79, 102744. https://doi.org/10.1016/j.techsoc.2024.102744

Wang, S., & Zhang, H. (2025). Artificial intelligence digital employees and sustainable innovation in online retail: The mediating role of ambidextrous green innovation and the moderating role of ethical anxiety. *Journal of Retailing and Consumer Services*, 84. Scopus. https://doi.org/10.1016/j.jretconser.2025.104235

Witkowski, A., & Wodecki, A. (2025). Where does AI play a major role in the new product development and product management process? *Management*

Review Quarterly. https://doi.org/10.1007/s11301-025-00533-5

Witkowski, A., & Wodecki, A. (2026). A prescriptive generative AI maturity model for new product development processes. *Journal of Manufacturing Technology Management*, 37(9), 40–63. https://doi.org/10.1108/JMTM-09-2025-0884

Xia, H., Chen, H., Zhang, J. Z., & Kamal, M. M. (2026). Exploring the impact of responsible AI governance on corporate performance: A quasi-natural experiment. *Technological Forecasting and Social Change*, 223, 124425. https://doi.org/10.1016/j.techfore.2025.124425

Xu, Y., Xie, P., Naeem, R. M., Almugren, I., Hameed, Z., & Agarwal, S. (2026). Responsible AI and employee service innovation behavior: A sequential mediation model of AI self-efficacy and AI crafting. *Technological Forecasting and Social Change*, 224, 124470. https://doi.org/10.1016/j.techfore.2025.124470

Yin, J., & Hoang, K. D. (2026). AI-enabled knowledge renewal: The role of leaders' AI attitudes and unlearning in enhancing employees' creative performance. *Journal of Knowledge Management*, 30(1), 211–230. https://doi.org/10.1108/JKM-02-2025-0209

Zhang, C., & Yang, J. (2024). Artificial intelligence and corporate ESG performance. *International Review of Economics & Finance*, 96, 103713. https://doi.org/10.1016/j.iref.2024.103713

Zucker, A. (2026). 5 sustainability trends shaping corporate responsibility in 2025. Available at: https://www.workiva.com/blog/sustainability-trends-shaping-corporate-responsibility-2025 (Accessed: 13 March 2026).

Preserving Human Agency in AI-Driven Interactions: The Role of User Experience Design

Hanlie Smuts, Blossom Alloziem, Catherine Dyason, Keiara Govender
University of Pretoria
Department of Informatics
Corner of Lynnwood Road and Roper Street, Hatfield, Pretoria, 0083, South Africa
hanlie.smuts@up.ac.za,{u22604244, u23573806, u23689197}@tuks.co.za

Abstract. *Artificial intelligence (AI) is a foundational component of information systems, influencing decisions across all industries and social platforms. Product and software solution developers devote considerable attention to improving algorithmic performance, with less focus on how customers, communities, and users experience, understand, and influence AI systems. Therefore, this paper seeks to explain how user experience (UX) design can be applied to preserve human agency in AI-driven interactions. The authors analyzed 129 papers using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) and derived four UX design considerations: trust, transparency and explainability, inclusion and accessibility, and embedded ethics as key dimensions of responsible AI design preserving human agency. These considerations shift the focus of AI design from technological capability toward human empowerment and agency. Preserving human agency requires more than technical accuracy; it requires systems that support meaningful human participation and oversight.*

Keywords. Artificial Intelligence, User Experience, Explainable AI, Participatory Design, Community Empowerment, AI Ethics

1 Introduction

Artificial intelligence (AI) is no longer an emerging technology. AI is embedded in the infrastructure of everyday life, and AI-driven systems increasingly influence decision-making in education, employment, healthcare, financial services, and other sectors. (Mishra, 2024). This integration of AI technologies shapes opportunities, allocates resources, facilitates interactions between individuals and institutions, and consequently, has become more than just a technological tool (Naikar et al., 2026). Sophisticated AI systems are perceived as more than just a tool for business operations efficacy. It is regarded as a key strategic asset with the potential to produce never-before-experienced outcomes in business, such as hyper-personalization at scale, real-time decision-making, hidden customer needs prediction, and inimitable solution design (Dzreke, 2025). AI systems also provide new opportunities for innovation and value creation beyond realizing operational efficiency (Ali et al., 2024).

In addition, organizations can make more informed decisions and respond more effectively in dynamic environments due to the ability of AI systems to process vast amounts of data (big data), identify complex patterns, and generate actionable insights - in some instances, autonomously (Li & Tian, 2026; Nagaboyina, 2024). AI technologies have the potential to augment human expertise, enhance creativity and support productivity (Smuts & Smith, 2021).

Despite these advances in AI systems, concerns have emerged about fairness, transparency, accountability, and inclusion (Akinrinola et al., 2024; Panarese et al., 2025). Stakeholders (communities, customers, users, and employees) are impacted by opaque algorithmic decisions, creating challenges for engendering trust and establishing democratic oversight (Panarese et al., 2025). Furthermore, stakeholders most impacted by AI remain underrepresented in the design, development and governance of such systems (Kaminski & Malgieri, 2024). Consequently, AI systems risk reinforcing existing inequalities, while simultaneously obscuring the mechanisms through which such inequalities are reproduced (Bircan & Özbilgin, 2025).

The socio-technical nature of user experience (UX) is associated with the usability and usefulness of interface design, mediating relationships between people (stakeholders) and intelligent systems (Cha & Wong, 2025). Through decisions about interaction design, accessibility, explainability, and participation, UX practitioners influence whether AI systems empower users or diminish their agency (Bennett et al., 2023). Therefore, the research question investigated in this paper asked: "*How can user experience design be applied to preserve human agency in AI-driven interactions?*" The findings of this study could assist

AI solution designers in developing AI-driven systems that enhance user autonomy, support informed decision-making, and maintain meaningful human control throughout the interactive process.

Section 2 presents the background of the study, followed by the research methodology in Section 3 and the findings in Section 4. Section 5 presents the study contributions, and Section 6 concludes the paper.

2 Background

2.1 Human agency and explainability in AI systems

The rapid adoption of AI has transformed the relationship between humans and technology (Naikar et al., 2026). Such automated systems increasingly influence decisions and recommendations traditionally made by humans. Scholars report that when stakeholders cannot understand or challenge algorithmic outputs, their capacity for informed decision-making is reduced (Grimmelikhuijsen & Meijer, 2022; Heaton et al., 2023; Masoudi, 2025). Human agency is the individual capacity to make independent choices, take purposeful actions, and shape their own lives and environment. It contrasts with the idea that humans are just passive reactors to natural forces or societal rules; rather, it embodies the power to think, decide, and enact change (Welzel & Inglehart, 2010). Human agency in AI contexts refers to stakeholders' ability to understand, influence, and contest decisions that affect them (Holter & El-Assady, 2024).

Scholars recognize trust as a prerequisite for successful human-AI interaction (Glikson & Woolley, 2020; Thinnakkakath, 2026). Accordingly, research in explainable AI (XAI) has focused on developing methods that make algorithmic reasoning more interpretable for stakeholders to understand how AI systems function, why decisions are made, and what limitations exist. Effective explanations should be understandable, actionable, and contextually relevant to stakeholders' knowledge and circumstances (Hannibal & Lindner, 2023; Subramanian et al., 2024; Wang et al., 2025). Hannibal and Lindner (2023) elaborate that AI systems should not proffer superficial explanations that create the appearance of openness yet do not enable true understanding. In this context, UX design is vital to ensuring that transparency mechanisms indeed empower users rather than merely legitimize opaque systems (Gray et al., 2018).

Data represent another key resource related to XAI (Durango et al., 2025; Meier, 2025). Carroll et al. (2020) define the people- and purpose-oriented CARE principles (**C**ollective benefit, **A**uthority to control, **R**esponsibility, and **E**thics), which reflect the important role of data in advancing governance and human agency. These principles position stakeholders as data stewards rather than just data sources to attain personal data sovereignty and governance (Baraku et al., 2025). Data sovereignty should be designed to grant stakeholders practical control over their distributed personal data, especially regarding its usage in AI training (Baraku et al., 2025).

2.2 Inclusion and participatory design

The transition from algorithms to human agency is necessary to ensure that AI systems promote equality, accountability, and human dignity (Papagiannidis et al., 2025). Therefore, AI systems should not reproduce social inequalities when they are developed within unequal social contexts (Lainjo, 2023). Such development could cause marginalized groups to experience barriers related to digital access, language support, accessibility, and representation in training datasets. Older adults, rural communities, low-income populations, minority language-speakers, and individuals with disabilities are vulnerable to exclusion (Cunningham et al., 2025; Panarese et al., 2025).

Kang et al. (2026) propose that UX design principles provide a useful framework for addressing these challenges, as inclusion should not be treated as a compliance requirement but rather as a foundational design principle that benefits all stakeholders (Ntoa et al., 2024). Therefore, within the AI context, accessibility extends beyond interface usability to include accessible explanations, adaptive interaction models, and equitable access to algorithmic services (Panarese et al., 2025). Rather than treating stakeholders as passive recipients of technology, UX design positions them as active contributors and co-creators (Olsson, 2004). Hence, participatory design ensures that stakeholders are directly involved in the design process by participating in defining problems, selecting datasets, and evaluating outcomes (Bødker et al., 2022).

2.3 Human agency and oversight models

The tension between human agency and automation in organizations suggests that organizations should balance their productivity gains with professional (human) accountability (Taş et al., 2026). Therefore, oversight routines vary across different task types as transactional work offers limited intervention points, whereas collaborative and expert (knowledge) work provide more opportunities to assert human agency (Krakowski, 2025). In this context, where humans maintain meaningful control in human-AI interaction, the socio-technical nature of AI-enabled work is emphasized (Taş et al., 2026).

To operationalize human agency, human oversight models have been proposed. Based on research over decades, scholars applied these models to human-AI interaction. The Human-in-the-Loop (HITL) model suggest direct human involvement in the AI decision-making process. Human review, approve or modify AI outputs before actions are implemented (EU, 2018;

Kabata & Thaldar, 2024; Singh & Szajnfarber, 2026). HITL is typically applied in high-risk environments such as healthcare, criminal justice and finance (Kabata & Thaldar, 2024).

The Human-on-the-Loop (HOTL) model offers lower level of intervention as in this instance, AI systems operate autonomously while a human supervisor continuously monitors its performance. The human operator can intervene or override decisions when necessary (EU, 2018; Maadi et al., 2021; Singh & Szajnfarber, 2026). HOTL is typically applied in semi-autonomous environments where human supervision is essential for safety and accountability, e.g., autopilot in commercial aviation or in autonomous vehicles.

In the Human-in-Control (HIC) model, humans retain ultimate authority over the deployment, objectives, operation, and governance of AI systems rather than being based on individual decisions (EU, 2018; Singh & Szajnfarber, 2026). Applying the HIC model, humans maintain strategic control (as opposed to operational monitoring only) as they determine when AI should be used, define the operational boundaries, establish governance mechanisms, and remain accountable for system outcomes (Taş et al., 2026).

Taş et al. (2026) present the concept of *episodic oversight sequencing* which suggests that human oversight should not be confined to a single approval point, but rather distributed across multiple workflow stages.

3 Research methodology

This paper aimed to investigate how UX design could be applied to preserve human agency in AI-driven interactions. The study achieved this aim by applying the analysis process suggested by Donthu et al. (2021), consisting of four steps: defining the aims and scope of the study; choosing the techniques for analysis; collecting the data for analysis and executing the analysis; and reporting the findings. The researchers applied the enrichment technique and created a network analysis (Donthu et al., 2021).

The study executed a keyword search based on the research objectives and to collect the data for the analysis in *Scopus*, a large literature repository of peer-reviewed abstract and citation data, applying the following keywords: (TITLE-ABS-KEY (human agency) AND TITLE-ABS-KEY (artificial intelligence) AND TITLE-ABS-KEY (human centered design)) AND (LIMIT-TO (LANGUAGE , "English")). We specifically chose these keywords as we followed a focused strategy centered explicitly on human agency and UX as opposed to broader frameworks such as autonomy, human control, user agency and human-centered AI. The keyword search yielded a result set of 160 papers. The authors screened the papers and removed one duplicate, 18 papers unrelated to the research topic (e.g., assisted building design, media innovation, biomedical visualization, children's fables, agriculture study, or animals), and 11 papers that indeed related to the subject area but did not explicitly deal with human agency. Fig. 1 depicts the PRISMA reporting that resulted in 129 papers from the corpus for analysis.

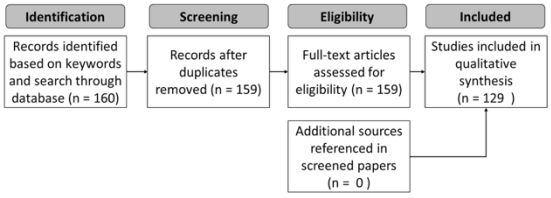


Figure 1. PRISMA reporting

Fig. 2 demonstrates a gradual increase in the number of included studies over time, from a single study in 2014 and 2017–2018 to 13 studies in 2024. The highest number of studies was published in 2025 (n = 54), followed by 2026 (n = 34). Overall, the findings indicate a substantial growth in research activity in recent years, with the majority of included studies published between 2024 and 2026. It must be noted that, as a cross-sectional study that was completed at a snapshot in time, the citations for 2026 will still increase.

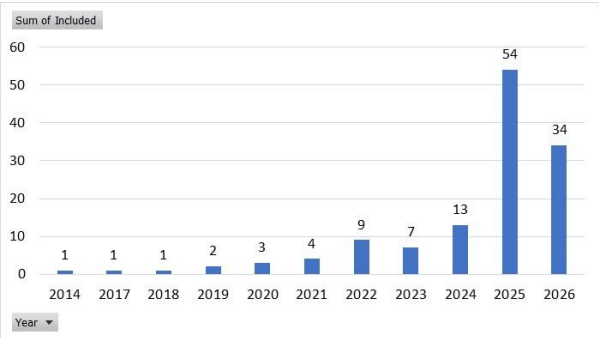


Figure 2. Year of publication

Table 1 presents the distribution of the 129 publications according to publication type. Journal articles constituted the largest proportion of the corpus, accounting for 59 publications (46%). Conference papers represented the second-largest category, with 51 publications (40%). Book chapters accounted for 12 publications (9%), and books contributed five publications (4%). These publication types were less prevalent, yet they provided valuable theoretical perspectives and comprehensive discussions that complemented the journal and conference literature. Editorials constituted the smallest category, with only two publications (2%).

Table 1. Paper type

Publication type	n	%

Article	59	46%
Book	5	4%
Book chapter	12	9%
Conference paper	51	40%
Editorial	2	2%
Grand Total	129	100%

4 Data analysis

The bibliometric information in the 129 papers was downloaded as a .RIS file and imported into *VOSviewer* version 1.6.20 (https://app.vosviewer.com/). *VOSviewer* constructed and visualized the co-occurrence data of the corpus, depicted in Fig. 3. The co-occurrence network consisted of four clusters extracted from 5695 terms, with 120 terms meeting the threshold of a minimum of ten occurrences per term.

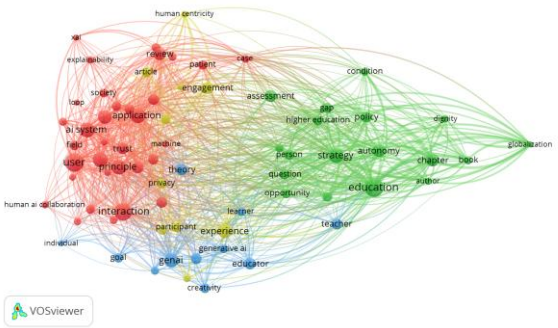


Figure 3. UX, Human agency, and AI cluster analysis

Table 2. Cluster analysis description

	Items (n)	Derived cluster label	Cluster description
1	33	Trust	Trust must be built through consistent accountability mechanisms and genuine stakeholder responsiveness.
2	18	Transparency and Explainability	Stakeholders must be able to understand, contest, and shape the AI systems that affect their lives, and not merely receive their outputs.
3	11	Inclusion and Accessibility	Inclusive design should consider who the design excludes and what is required to reach them.
4	10	Embedded Ethics	Ethics must be a structural feature of the design process, governance, and iteration cycle.

Table 2 presents the four clusters in more detail, highlighting the cluster number, the number of themes in a cluster, and a detailed list of the themes. A cluster label and a short description were derived for each cluster from the cluster themes.

5 Results

This paper investigated the role of UX design in preserving human agency in AI-driven interactions. AI systems consist of algorithms, datasets, predictive models, and automated decision-making processes. Stakeholders bring lived experiences, cultural knowledge, values, needs, and expectations. UX may function as the socio-technical interface through which these domains interact. Based on the bibliographic coupling of the corpus, the clusters identified in Table 2 are described in more detail in the following sections and visualized in Fig. 4.

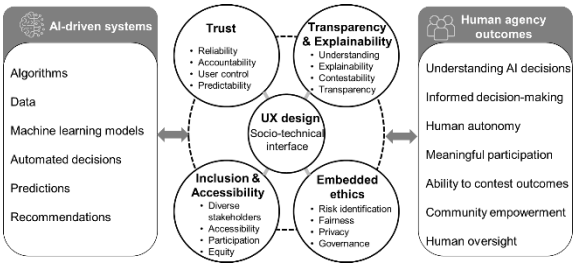


Figure 4. UX principles for preserving human agency in AI-driven interactions

Fig. 4 illustrates the relationship between AI-driven systems, user experience (UX) design, and human agency outcomes. On the left, AI-driven systems are depicted as algorithms, data, machine learning models, automated decisions, and predictive recommendations that influence user interactions. At the centre of the visualization is User Experience (UX) Design, which serves as a socio-technical interface connecting AI systems with human outcomes. Four key UX design principles surround the central component. Together, these design principles shape positive human agency outcomes, including informed decision-making, human autonomy, meaningful participation, community empowerment, and human oversight of AI-enabled decisions.

5.1 Trust

Trust emerged as a key requirement for preserving human agency in AI-driven interactions. Stakeholders are more willing to engage with AI systems when they perceive them as reliable, predictable, and aligned with their interests (Bach et al., 2024; Glikson & Woolley, 2020). UX engenders trust through interaction

mechanisms that enable stakeholders to understand, monitor, and influence AI system behaviour, e.g., when farmers use an AI crop advisory system but can compare AI recommendations with historical local farming practices. In the context of trust, UX design emphasizes the importance of communicating system capabilities, limitations, confidence levels, and uncertainty to support appropriate reliance rather than merely accepting algorithmic outputs (Stimson & Raper, 2025). This status is achieved by creating interfaces that provide feedback, support user control, and enable system predictability across contexts (Nam & Ki Chan, 2025; Renata et al., 2023). Thus, UX design encourages informed reliance on AI systems, while providing opportunities for human judgment and intervention (Doh et al., 2025).

5.2 Transparency and explainability

Stakeholders can understand and contest AI-driven systems through a transparency and explainability lens. UX design holds that explanations should be meaningful, contextual, and tailored to the needs of different stakeholders, e.g., an AI welfare allocation system provides citizens with clear explanations of eligibility decisions and appeal processes (Bozkurt & Sharma, 2024). Transparency contributes to trust when stakeholders understand why a recommendation or decision has been generated and what factors influence the outcome (Zerilli et al., 2022). However, transparency alone is not sufficient if explanations are overly technical or merely create the illusion of openness without supporting genuine understanding, e.g., when a bank loan is denied by an AI system without providing the customer with reasons, nor the process for improving or appealing the decision (Rudin, 2019; Xie et al., 2020). UX design ensures multi-layered, visual representations and interactive feedback mechanisms (Bozkurt & Sharma, 2024).

5.3 Inclusion and accessibility

The inclusion and accessibility cluster highlights the importance of ensuring that AI systems serve diverse stakeholders rather than reinforcing social inequalities through underrepresented training data, e.g., using AI systems trained predominantly on data from one demographic group in healthcare (Murphy & Largacha-Martínez, 2022; Panarese et al., 2025). Underrepresentation might also be included in training data based on minority language speakers, differently abled people, rural populations, and historically marginalized communities. For example, AI services through mobile phones and short message service (SMS) are unavailable to stakeholders without broadband internet access, or with indigenous language interfaces that do not enable participation by stakeholders who are typically excluded from mainstream AI systems (Cunningham et al., 2025). Inclusive UX design extends beyond usability and addresses inequities through accessible interface, adaptive interaction models, and culturally sensitive design practices (Jacucci et al., 2014; Nkwo et al., 2025). Hence, by treating stakeholders as co-creators rather than passive users, UX design supports more equitable outcomes and reduces the likelihood of exclusion, ensuring the incorporation of local contexts, values, and knowledge structures into AI systems (Ahrweiler et al., 2025; Stimson & Raper, 2025).

5.4 Embedded ethics

Embedded ethics imply that ethical considerations should be integrated into the design and governance of AI systems rather than applied retrospectively (after deployment), e.g., stakeholders participate in defining acceptable uses of their data before AI development begins (Bryson, 2018; Sanderson et al., 2023). UX designers operationalize principles, such as fairness, accountability, contestability, privacy, and human-centred values, throughout the system lifecycle by translating them into practical design features that stakeholders can experience and evaluate (Papagiannidis et al., 2025). This outcome is achieved through interface design that makes responsibility visible and actionable, identifies risks, and evaluates impacts throughout development (Lindgren et al., 2025; Young et al., 2019). UX design supports human agency and embedded ethics by ensuring that AI systems remain accountable to the stakeholders they serve by balancing automation with human oversight, fairness, and user control (Bryson, 2018; Mittelstadt et al., 2016).

6 Conclusion

The objective of this paper was to understand how UX design could be applied to maintain human agency in AI-driven interactions. The study analyzed 129 papers by applying the PRISMA method. Subsequently, a conceptual framework was derived consisting of four UX design considerations: trust, transparency and explainability, inclusion and accessibility, and embedded ethics, as key responsible AI design dimensions for preserving human agency. These considerations elevate AI system design beyond technological capability towards human empowerment. It emphasizes systems that enable meaningful participation, informed decision-making, and effective human oversight, i.e., human agency.

Methodological decisions such as restricting the review to Scopus-indexed publications and applying predefined search terms, were addressed in the methodology section. However, these limitations also provide opportunities for future research. Future empirical studies should validate the proposed conceptual framework across discrete AI system domains. In addition, future research could investigate the experiences of diverse and underrepresented user

groups to ensure that AI systems support agency inclusively and equitably.

References

Ahrweiler, P., Späth, E., Siqueiros García, J. M., Capellas, B. L., & Wurster, D. (2025). Inclusive Technology Co-design for Participatory AI. In P. Ahrweiler (Ed.), *Participatory Artificial Intelligence in Public Social Services: From Bias to Fairness in Assessing Beneficiaries* (pp. 35-62). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-71678-2_2

Akinrinola, O., Okoye, C. C., Ofodile, O. C., & Ugochukwu, C. E. (2024). Navigating and reviewing ethical dilemmas in AI development: Strategies for transparency, fairness, and accountability. *GSC Advanced Research and Reviews*, *18*(3), 050-058. https://doi.org/https://doi.org/10.30574/gscarr.2024.18.3.0088

Ali, M., Khan, T. I., Khattak, M. N., & Şener, İ. (2024). Synergizing AI and business: Maximizing innovation, creativity, decision precision, and operational efficiency in high-tech enterprises. *Journal of Open Innovation: Technology, Market, and Complexity*, *10*(3), 100352. https://doi.org/https://doi.org/10.1016/j.joitmc.2024.100352

Bach, T. A., Khan, A., Hallock, H., Beltrão, G., & Sousa, S. (2024). A Systematic Literature Review of User Trust in AI-Enabled Systems: An HCI Perspective. *International Journal of Human–Computer Interaction*, *40*(5), 1251-1266. https://doi.org/10.1080/10447318.2022.2138826

Baraku, V., Paraskakis, I., Veloudis, S., & Yadav, P. (2025). *Extending Personal Data Sovereignty by Enabling Governance of AI Training on Personal Data* Working Conference on Virtual Enterprises, https://doi.org/10.1007/978-3-032-05673-3_2

Bennett, D., Metatla, O., Roudaut, A., & Mekler, E. D. (2023). *How does HCI understand human agency and autonomy?* Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, https://doi.org/10.1145/3544548.3580651

Bircan, T., & Özbilgin, M. F. (2025). Unmasking inequalities of the code: Disentangling the nexus of AI and inequality. *Technological Forecasting and Social Change*, *211*, 123925. https://doi.org/https://doi.org/10.1016/j.techfore.2024.123925

Bødker, S., Dindler, C., Iversen, O. S., & Smith, R. C. (2022). What is participatory design? In *Participatory design* (pp. 5-13). Springer. https://doi.org/https://doi.org/10.1007/978-3-031-02235-7_2

Bozkurt, A., & Sharma, R. (2024). Trust, Credibility and Transparency in Human-AI Interaction: Why We Need Explainable and Trustworthy AI and Why We Need It Now. *19*, i-ix. https://doi.org/10.5281/zenodo.14599168

Bryson, J. J. (2018). Patiency is not a virtue: the design of intelligent systems and systems of ethics [Article]. *Ethics and Information Technology*, *20*(1), 15-26. https://doi.org/10.1007/s10676-018-9448-6

Carroll, S. R., Garba, I., Figueroa-Rodríguez, O. L., Holbrook, J., Lovett, R., Materechera, S., Parsons, M., Raseroka, K., Rodriguez-Lonebear, D., & Rowe, R. (2020). The CARE principles for indigenous data governance. *Data Science Journal*, *19*(43), 1-12. https://doi.org/https://doi.org/10.5334/dsj-2020-043

Cha, I., & Wong, R. Y. (2025). *Understanding Socio-technical Factors Configuring AI Non-Use in UX Work Practices* Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, https://doi.org/10.1145/3706598.3713140

Cunningham, J. L., Adjagbodjou, A., Basoah, J., Jawara, J., Kadoma, K., & Lewis, A. (2025). *Toward Responsible ASR for African American English Speakers: A Scoping Review of Bias and Equity in Speech Technology* [Conference paper]. Proceedings of the 8th AAAI/ACM Conference on AI, Ethics, and Society (AIES 2025), Madrid, Span. https://doi.org/10.1609/aies.v8i1.36580

Doh, W., Goh, Y., & Kim, S. H. (2025). *Beyond Overreliance: The Human-AI-System Concordance (HASC) Matrix and the Cognitive Dynamics of AI-Assisted Decision-Making* [Conference paper]. Proceedings of the Human Factors and Ergonomics Society, https://doi.org/10.1177/10711813251358240

Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W. M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research*, *133*, 285-296. https://doi.org/https://doi.org/10.1016/j.jbusres.2021.04.070

Durango, I., Gallud, J. A., & Penichet, V. M. R. (2025). The data dance: choreographing seamless partnerships between humans, data, and GenAI [Article]. *International Journal of Data Science and Analytics*, *20*(4), 3613-3640. https://doi.org/10.1007/s41060-024-00684-w

Dzreke, S. S. (2025). The competitive advantage of AI in business: A strategic imperative. *International Journal for Multidisciplinary Research*, *7*(4), 50400. https://doi.org/https://doi.org/10.36948/ijfmr.2025.v07i04.50400

EU. (2018). European Commission Ethics guidelines for trustworthy AI 2019 https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai accessed June 2026.

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, *14*(2), 627-660. https://doi.org/10.5465/annals.2018.0057

Gray, C. M., Kou, Y., Battles, B., Hoggatt, J., & Toombs, A. L. (2018). *The dark (patterns) side of UX design* Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, https://doi.org/10.1145/3173574.3174108

Grimmelikhuijsen, S., & Meijer, A. (2022). Legitimacy of algorithmic decision-making: Six threats and the need for a calibrated institutional response. *Perspectives on Public Management and Governance*, *5*(3), 232-242. https://doi.org/https://doi.org/10.1093/ppmgov/gvac008

Hannibal, G., & Lindner, F. (2023). *Towards a Questions-Centered Approach to Explainable Human-Robot Interaction* [Conference paper]. Frontiers in Artificial Intelligence and Applications, https://doi/org/10.3233/FAIA220641

Heaton, D., Clos, J., Nichele, E., & Fischer, J. E. (2023). *The social impact of decision-making algorithms: reviewing*

the influence of agency, responsibility and accountability on trust and blame Proceedings of the First International Symposium on Trustworthy Autonomous Systems, https://doi.org/10.1145/3597512.3599706

Holter, S., & El-Assady, M. (2024). Deconstructing Human-AI Collaboration: Agency, Interaction, and Adaptation [Article]. *Computer Graphics Forum*, *43*(3), Article e15107. https://doi.org/10.1111/cgf.15107

Jacucci, G., Spagnolli, A., Freeman, J., & Gamberini, L. (2014). Symbiotic interaction: A critical definition and comparison to other human-computer paradigms [Article]. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8820*, 3-20. https://doi.org/10.1007/978-3-319-13500-7_1

Kabata, F., & Thaldar, D. (2024). Human in the loop requirement and AI healthcare applications in low-resource settings: A narrative review. *South African Journal of Bioethics and Law*, *17*(2), 70-73. https://doi.org/https://hdl.handle.net/10520/ejc-m_sajbl_v17_n2_a6

Kaminski, M. E., & Malgieri, G. (2024). Impacted stakeholder participation in AI and data governance. *Forthcoming on Yale Journal of Law and Technology (2024-2025), U of Colorado Law Legal Studies Research Paper*(24-23). https://doi.org/https://ssrn.com/abstract=4836460

Kang, E., Choi, S., Li, X., & Hwang, H. (2026). *Exploring Factors Influencing Perceived Usefulness and Adoption Intention for Financial AI Services Across User Experience Levels* [Conference paper]. Communications in Computer and Information Science, https://doi.org/10.1007/978-3-032-12767-9_14

Krakowski, S. (2025). Human-AI agency in the age of generative AI. *Information and Organization*, *35*(1), 100560. https://doi.org/https://doi.org/10.1016/j.infoandorg.2025.100560

Lainjo, B. (2023). The Global Social Dynamics and Inequalities of Artificial Intelligence. *International Journal of Innovation Scientific Research and Review*, *5*(8), 4966-4974. https://doi.org/10.13140/RG.2.2.20078.79689

Li, H., & Tian, F. (2026). Advancing Decision-Making through AI-Human Collaboration: A Systematic Review and Conceptual Framework. *Group Decision and Negotiation*, *35*(2), 26. https://doi.org/10.1007/s10726-026-09980-1

Lindgren, H., Lindvall, K., & Richter-Sundberg, L. (2025). Responsible design of an AI system for health behavior change-an ethics perspective on the participatory design process of the STAR-C digital coach. *Front Digit Health*, *7*, 1436347. https://doi.org/10.3389/fdgth.2025.1436347

Maadi, M., Akbarzadeh Khorshidi, H., & Aickelin, U. (2021). A Review on Human–AI Interaction in Machine Learning and Insights for Medical Applications. *International Journal of Environmental Research and Public Health*, *18*(4), 2121. https://doi.org/https://doi.org/10.3390/ijerph18042121

Masoudi, M. (2025). Algorithmic Governance, Data-Driven Decision Making, and the Transformation of Democratic Accountability in Contemporary States. *Advanced Journal of Management, Humanity and Social Science*, *2*(1), 10-22. https://doi.org/10.5281/zenodo.18009536

Meier, R. (2025). Balancing Minds and Data: The Privacy Dilemma of LLMs and Anthropomorphism in LLMs [Article]. *Journal of Social Computing*, *6*(3), 173-183. https://doi.org/10.23919/JSC.2025.0014

Mishra, P. (2024). Artificial Intelligence in Everyday Life. *Journal of Innovations in Computer Science And Trends in IT*, *1*(2), 3048-4707. https://doi.org/https://bciit.ac.in/pdf/eJournal/Issue%202/3Formatted_PreetiAlok.pdf

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2), 1-21. https://doi.org/10.1177/2053951716679679

Murphy, J. W., & Largacha-Martínez, C. (2022). Decolonization of AI: a Crucial Blind Spot [Article]. *Philosophy and Technology*, *35*(4), Article 102. https://doi.org/10.1007/s13347-022-00588-2

Nagaboyina, P. K. (2024). AI-Driven Big Data Analytics: Transforming Decision-Making and Strategic Insights. *INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS*, *11*, 786-791. https://doi.org/https://www.researchgate.net/publication/395464648_AI-Driven_Big_Data_Analytics_Transforming_Decision-Making_and_Strategic_Insights

Naikar, N., Hoffman, R., Roth, E. M., Klein, G., Militello, L. G., & Dominguez, C. (2026). Should we Make AI More Tool-like or Teammate-Like? *Journal of Cognitive Engineering and Decision Making*, *20*(2), 137-157. https://doi.org/https://doi.org/10.1177/15553434251346904

Nam, S., & Ki Chan, C. C. (2025). *Towards Designing User Interfaces for Optimized Human-AI Communication and Supervisory Control in Software Engineering* [Conference paper]. Proceedings - 29th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, SNPD 2025-Summer, https://doi.org/10.1109/SNPD65828.2025.11252932

Nkwo, M., Ikwunne, T., Adejoro, C., & Anuyah, O. (2025). Exploring Usability of AI Systems in the Global South—Toward Responsible Human-Centered AI for Sustainable Cities and Communities. In *Human - Computer Interaction Series* (Vol. Part F9478, pp. 125-165). https://doi.org/10.1007/978-3-031-88326-2_5

Ntoa, S., Margetis, G., Valakou, A., Makri, F., Dimitriou, N., Karakostas, I., Kokkinis, G., Apostolakis, K. C., Tzovaras, D., & Stephanidis, C. (2024). *A Mixed-Methods Approach for the Evaluation of Situational Awareness and User Experience with Augmented Reality Technologies* [Conference paper]. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), https://doi.org/10.1007/978-3-031-61569-6_13

Olsson, E. (2004). What active users and designers contribute in the design process. *Interacting with Computers*, *16*(2), 377-401. https://doi.org/10.1016/j.intcom.2004.01.001

Panarese, P., Grasso, M. M., & Solinas, C. (2025). Algorithmic bias, fairness, and inclusivity: A multilevel framework for justice-oriented AI. *AI & SOCIETY*, *41*, 1-23. https://doi.org/https://doi.org/10.1007/s00146-025-02451-2

Papagiannidis, E., Mikalef, P., & Conboy, K. (2025). Responsible artificial intelligence governance: A review and research framework. *The Journal of Strategic Information Systems*, *34*(2), 101885. https://doi.org/https://doi.org/10.1016/j.jsis.2024.101885

Renata, V., Lee, J. D., & Kamaraj, A. V. (2023). *Trading Degrees of Freedom: User Control in Pens, GPT, and Automated Vehicles* [Conference paper]. Proceedings of the Human Factors and Ergonomics Society, https://doi.org/10.1177/21695067231192270

Rudin, C. (2019). Stop explaining black box machine learning models for high-stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206-215. https://doi.org/10.1038/s42256-019-0048-x

Sanderson, C., Douglas, D., Lu, Q., Schleiger, E., Whittle, J., Lacey, J., Newnham, G., Hajkowicz, S., Robinson, C., & Hansen, D. (2023). AI Ethics Principles in Practice: Perspectives of Designers and Developers [Article]. *IEEE Transactions on Technology and Society*, *4*(2), 171-187. https://doi.org/10.1109/TTS.2023.3257303

Singh, A., & Szajnfarber, Z. (2026). Architecting Human-AI Systems for Effective Collaboration and Oversight: Making Sense of Human/AI-in/on/Over/Under/Along-the-Loop. *Systems Engineering*, *29*(2), 337-353. https://doi.org/https://doi.org/10.1002/sys.70024

Smuts, H., & Smith, A. (2021). Collaboration of Human and Machine for Knowledge Work: An Organisational Transformation Framework for Data-driven Decision-making. In Z. W. Y. Lee, T. K. H. Chan, & C. M. K. Cheung (Eds.), *Information Technology in Organisations and Societies: Multidisciplinary Perspectives from AI to Technostress* (pp. 0). Emerald Publishing Limited. https://doi.org/10.1108/978-1-83909-812-320211002

Stimson, C. E., & Raper, R. (2025, 2025//). *Participatory AI: A Method for Integrating Inclusive and Ethical Design Considerations into Autonomous System Development* Towards Autonomous Robotic Systems, Cham. https://doi.org/10.1007/978-3-031-72059-8_13

Subramanian, H. V., Canfield, C., & Shank, D. B. (2024). Designing explainable AI to improve human-AI team performance: A medical stakeholder-driven scoping review [Review]. *Artificial Intelligence in Medicine*, *149*, Article 102780. https://doi.org/10.1016/j.artmed.2024.102780

Taş, E., Memmert, L., & Bittner, E. (2026). Episodic oversight in generative AI workflows: A nine-step protocol for preserving human agency (OP-9). *Electronic Markets*, *36*(1), 58. https://doi.org/https://doi.org/10.1007/s12525-026-00915-x

Thinnakkakath, G. (2026). The Future of Digital Trust: Designing Transparent and Ethical Information Ecosystems. In *Media Literacy and the Politics of Digital Misinformation* (pp. 1-27). https://doi.org/10.4018/979-8-3373-6566-4.ch001

Wang, L., Anyi, C. L., Xu, K., Liu, Y., Arriaga, R. I., & Goel, A. K. (2025). *Explainable AI for Daily Scenarios from End-Users' Perspective: Non-Use, Concerns, and Ideal Design* [Conference paper]. DIS 2025 - Proceedings of the 2025 ACM Designing Interactive Systems Conference, https://doi.org/10.1145/3715336.3735796

Welzel, C., & Inglehart, R. (2010). Agency, values, and well-being: A human development model. *Social indicators research*, *97*(1), 43-63. https://doi.org/https://doi.org/10.1007/s11205-009-9557-z

Xie, Y., Chu, M. T., Li, G., & Li, Y. (2020). Explainable artificial intelligence for designers: A human-centered perspective. *International Journal of Human-Computer Interaction*, *36*(16), 1487-1499. https://doi.org/10.1080/10447318.2020.1801171

Young, M., Katell, M., Krafft, P. M., & Dailey, D. (2019). Practical fairness: Dilemmas in the everyday work of machine learning application development. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1-23. https://doi.org/10.1145/3359236

Zerilli, J., Bhatt, U., & Weller, A. (2022). How transparency modulates trust in artificial intelligence. *Patterns (N Y)*, *3*(4), 100455. https://doi.org/10.1016/j.patter.2022.100455

A Comparative Study of SNN Learning Paradigms in Hybrid Graph-Spiking Neural Networks for Biomedical Data Analysis

Dominik Čambál, Monika Herchlová, Juraj Ďuďák
Slovak University of Technology in Bratislava
Faculty of Materials Science and Technology in Trnava
Jána Bottu 25, 917 24 Trnava, Slovakia
{dominik.cambal, monika.herchlova, juraj.dudak}@stuba.com

Abstract. *This study evaluates three spiking neural network (SNN) paradigms within a hybrid Graph-SNN framework for 12-lead ECG arrhythmia classification under severe class imbalance (*$10.3 : 1$*). We benchmark backpropagation through time (BPTT), ANN-to-SNN conversion, and unsupervised STDP on the INCART database across paired partitions. While conversion and BPTT achieve comparable ranking discrimination, conversion yields vastly superior calibration, enabling* 0.933 *minority beat recall and* 97.3% *accuracy. STDP fails to form a usable classifier. Requiring half the training time of BPTT, conversion offers the most effective quality-per-unit-compute trade-off.*

Keywords. graph neural networks, spiking neural networks, ECG classification, SNN learning paradigms

1 Introduction

Electrocardiogram (ECG) signal classification exhibits a dual structure; signals are sequential in time and relational across leads. A standard 12-lead ECG does not represent independent measurements, but 12 geometric projections of a single cardiac dipole in a three-dimensional space, constrained by electrode placement and cardiac physiology (Kligfield et al., 2007).

This structure is particularly evident in the precordial leads (V1-V6), which form a spatial trajectory across the anterior chest wall. As their position shifts relative to cardiac anatomy, so does their perspective on the depolarization wavefront, resulting in systematic morphological variations such as R-wave progression and QRS polarity changes. These patterns do not emerge independently at each lead, but arise from a shared underlying process observed from multiple points of view (Bulíková, 2014). Treating leads as independent channels, therefore, discards structural information that is intrinsic to the signal.

Conventional deep learning approaches only address one dimension of the structure. Convolutional neural networks, whether applied per-lead or over multiple stacked lead inputs, are effective local feature extractors, but without memory the classification depends only on the current receptive field and not on the temporal context accumulated across the beat window (Hannun et al., 2019). Recurrent architectures (e.g. LSTM) address the temporal dimension by maintaining a hidden state across timesteps, but represent leads (inputs) as a flat feature vector, collapsing their inter-lead relational structure.

Graph neural networks (GNNs) applied to a 12-lead ECG, where leads are modelled as nodes and their pairwise statistical dependencies as edges, allow message-passing over the lead topology at each timestep. That way, the GNNs extract inter-lead spatial features at each beat sample, while a downstream stateful temporal module integrates those features across the beat window.

The spatial-relational decomposition processing by a GNN along with a temporal integration by a stateful architecture fits the duality of multi-lead ECGs well. A hybrid GCN-GRU for the detection of atrial fibrillation was demonstrated to outperform an equivalent GRU applied to concatenated lead features by approximately six percentage points of F1, directly attributing the gain to the graph-structured spatial module (Ma and Xia, 2023).

The motivation for instantiating the temporal module as a spiking neural network (SNN) rather than an LSTM brings a two-fold value.

First, SNNs encode information at the rate or timing of discrete binary spike events and enable a sparse, asynchronous computation (Eshraghian et al., 2023) which has been shown to yield great reductions in energy on neuromorphic hardware per inference compared to GPU-based deployment in the Von Neumann sense (Malcolm and Casco-Rodriguez, 2023). The primary clinical context for arrhythmia detection, in continuous monitoring battery-powered wearable devices over days to weeks, directly impacts this efficiency argument (Chu et al., 2022).

Second, the temporal structure of ECG morphology

is naturally aligned with the integrate-and-fire neuron model. The beat features are present as a sequence of transient events in which timing relationships carry diagnostic information with an encoding consistent with the membrane potential dynamics of leaky integrate-and-fire neurons (LIF). Prior work on GNN-SNN hybridisation has addressed the spatial module. Zhu et al. (Zhu et al., 2022) introduced a spiking graph convolutional network that combines GCN-based spatial aggregation with spike-based neural processing, yielding competitive results on benchmark datasets. Wang and Jiang (Wang and Jiang, 2022) embedded spiking dynamics directly into the graph attention mechanism, reporting energy reductions over continuous-valued GATs. Both architectures treat the graph as static and target the spatial component only; neither addresses the problem of integrating a GNN encoder with a spiking temporal integrator over time-varying graph signals, which is the structure required for beat-level ECG classification. Moreover, no existing study compares multiple SNN learning paradigms within a fixed GNN-SNN hybrid framework on a class-imbalanced biomedical task.

All three principal training paradigms; backpropagation through time with surrogate gradients, ANN-to-SNN conversion via threshold-balanced shadow training, and unsupervised spike-timing-dependent plasticity (STDP); differ in computational complexity and cost, efficiency, biological plausibility, i.e. in reliance on supervisory signals, and overall optimisation dynamics (Yi et al., 2023). Their behavior under the pronounced class imbalance typical of arrhythmia datasets, such as the 10:1 Normal-to-Ventricular ratio in the INCART database, is not sufficiently understood.

This work addresses these gaps by implementing a hybrid GNN-SNN architectures in which a graph convolutional encoder processes the 12-lead ECG at each time step, and a spiking temporal module performs sequence integration. The temporal component is trained using three distinct paradigms; firstly, end-to-end BPTT with surrogate gradients (SNNTorch), secondly, ANN-to-SNN conversion via threshold-balanced shadow training (SpikingJelly), and lastly, unsupervised STDP (BindsNET). All variants are evaluated on binary arrhythmia classification (Normal vs Ventricular beats) using the St. Petersburg INCART 12-lead database, consequently, they are compared to a GNN-LSTM baseline, trained under identical conditions.

The following parts of the paper are organised as follows: Section 2 presents the theoretical foundations and methodology, including dataset construction. Section 3 describes the experimental setup and results. Sections 4 and 5 discuss the findings and outline directions for further work.

2 Methodology

In this section, we describe key concepts related to graph neural networks, recurrent and spiking neural networks and the three major learning paradigms of spiking neural networks.

2.1 Graph Neural Networks

Graph neural networks process signals defined on graph structures by iteratively updating node embeddings through neighborhood aggregation. In the 12-lead ECG setting, each lead is represented as a node and inter-lead dependencies as edges, allowing spatial information to be exchanged before temporal integration.

At layer k, the spatial message-passing mechanism computes messages from neighboring nodes $u \in \mathcal{N}(v)$, aggregates them, and updates the hidden representation of node v:

$$m_{u\to v}^{(k)} = \text{Message}^{(k)}(h_u^{(k-1)}, h_v^{(k-1)}, e_{u,v}), \quad (1)$$

$$m_v^{(k)} = \text{Aggregate}^{(k)}(\{m_{u\to v}^{(k)} : u \in \mathcal{N}(v)\}), \quad (2)$$

$$h_v^{(k)} = \text{Update}^{(k)}(h_v^{(k-1)}, m_v^{(k)}). \quad (3)$$

In this work, we use graph convolutional layers because they provide an efficient form of degree-normalized neighborhood aggregation, making them suitable for repeated processing of ECG lead graphs across timesteps.

2.1.1 Recurrent Neural Networks and Backpropagation Through Time

Recurrent neural networks maintain a hidden state across timesteps, allowing predictions to depend on both the current input and previously observed sequence context:

$$\mathbf{h}[t] = \tanh(\mathbf{W}\mathbf{x}[t] + \mathbf{U}\mathbf{h}[t-1]). \quad (4)$$

Gated variants such as LSTMs and GRUs mitigate the vanishing gradient problem by regulating information flow through internal gates. These models are usually trained by backpropagation through time (BPTT), where the recurrent computation is unrolled over the sequence and gradients are propagated through shared temporal parameters (Zhang et al., 2023). The repeated product of temporal Jacobians explains why long-sequence training may still suffer from vanishing or exploding gradients, motivating alternative stateful temporal modules such as spiking neural networks.

2.1.2 Spiking Neural Networks

Spiking neural networks (SNNs) are inspired by biological neurons. Unlike conventional neural networks, their neurons communicate through time-dependent spikes, represented as spike trains of ones and zeroes.

Several spiking neuron models exist, differing in biological plausibility and computational efficiency.

A commonly used model is the Leaky Integrate-and-Fire (LIF) neuron. Its internal state, called the membrane potential $U[t]$, accumulates input and decays over time. When it crosses a threshold θ, a spike is emitted and the membrane potential is reset:

$$U[t] = \beta U[t-1] + WX[t] - S_{out}[t-1]\theta. \quad (5)$$

Here, β is the decay rate, W are synaptic weights, $X[t]$ are inputs, and $S_{out}[t-1]$ denotes whether a spike occurred in the previous time step. If the decay is removed ($\beta = 1$), the model becomes the simpler Integrate-and-Fire (IF) model.

SNNs are motivated mainly by sparsity and energy efficiency. Information may be encoded in the rate or timing of spikes, allowing asynchronous and parallel computation. This makes SNNs particularly suitable for neuromorphic hardware, where memory and computation are integrated locally and energy consumption can be significantly reduced (Malcolm and Casco-Rodriguez, 2023).

However, SNNs still face important challenges, especially in training. Standard backpropagation is difficult to apply because spike generation is non-differentiable, leading to the so-called dead neuron problem. As a result, SNNs often underperform conventional neural networks on large benchmark datasets, although high accuracy has been reported on simpler datasets such as MNIST and CIFAR-10 (Eshraghian et al., 2023; Yi et al., 2023).

2.2 SNN Learning Paradigms

Training conventional neural networks relies on backpropagation and gradient descent, which require continuous and differentiable activation functions. In SNNs, however, the activation function is usually the Heaviside step function. Since it is not differentiable in the classical sense, gradients become either zero or undefined, which prevents direct application of standard backpropagation and causes the dead neuron problem (Eshraghian et al., 2023).

In practice, several approaches are used to train SNNs, including surrogate gradients, ANN-to-SNN conversion, and more biologically plausible local learning rules. These methods are often heuristic rather than fully rigorous.

2.2.1 Surrogate Gradient

The surrogate gradient method replaces the derivative of the non-differentiable spike function during training with the derivative of a smooth approximation, such as a sigmoid, arctangent, or triangular function (Eshraghian et al., 2023). The forward pass still uses spikes, but the backward pass uses the surrogate derivative.

This allows BPTT to be applied to SNNs by treating the membrane potential similarly to the hidden state in RNNs. For the LIF model, the temporal derivative includes the reset mechanism:

$$\frac{\partial U[k+1]}{\partial U[k]} = \beta - \theta \frac{\partial S[k]}{\partial U[k]}, \quad (6)$$

where the non-differentiable term is replaced by a surrogate gradient during training.

Surrogate gradients are currently considered the de facto standard for training SNNs because they can achieve sufficient accuracy. Their disadvantages include biological implausibility, increased memory and computational cost, and the need to tune surrogate functions, thresholds, and time constants carefully (Ayasi et al., 2025; Zenke and Vogels, 2023).

2.2.2 Shadow Training

Shadow training is based on training a conventional artificial neural network and then converting it into an SNN. Neuron activations are transformed into spikes using rate or temporal coding. The main advantage is that efficient ANN training methods can be used, which often improves accuracy on complex datasets. However, the conversion itself is approximate and may reduce the energy-efficiency benefits of SNNs, while also lacking biological plausibility (Eshraghian et al., 2023).

2.2.3 Spike-timing-dependent Plasticity

Spike-timing-dependent plasticity (STDP) is an unsupervised learning rule inspired by biological synaptic plasticity and the Hebbian principle. It updates synaptic weights according to the relative timing of presynaptic and postsynaptic spikes. If the presynaptic spike occurs shortly before the postsynaptic spike, the synapse is strengthened; if it occurs after, the synapse is weakened.

This mechanism is often described by

$$\Delta W = \begin{cases} A_+ e^{\Delta t/\tau_+}, & t_{post} > t_{pre}, \\ -A_- e^{-\Delta t/\tau_-}, & t_{post} < t_{pre}. \end{cases} \quad (7)$$

Here, ΔW is the weight update, A_+ and A_- determine the magnitudes of potentiation and depression, and τ_+ and τ_- are time constants. STDP is biologically plausible and local, but on its own it is usually insufficient for competitive performance and is often combined with supervisory signals or architectural modifications (Eshraghian et al., 2023; Hebb, 1949).

2.3 Dataset Description and Class Imbalance

We utilized the St. Petersburg INCART 12-lead Arrhythmia Database (Tihonenko et al., 2007), which

comprises 75 annotated 30-minute recordings extracted from 32 Holter records at a sampling rate of 257 Hz. To construct our dataset, we combined data from the first five recordings. Because the first two recordings contain exclusively normal (`N`) and ventricular ectopic (`V`) beats, all other beat classes were omitted to formulate a binary classification problem.

The resulting dataset exhibits severe class imbalance, containing 10,967 `N` beats and 1,066 `V` beats (an imbalance ratio of approximately 10.3:1).

2.4 Preprocessing and Data Partitioning

During preprocessing, each lead within every beat window was independently normalized using z-score normalization:

$$z = \frac{x - \mu}{\sigma} \tag{8}$$

where μ and σ denote the mean and standard deviation of the specific lead window. Independent lead scaling eliminates amplitude variance across leads and subjects, forcing models to learn structural waveform morphology. Inter-lead graph edges were established based on spatial correlation, retaining connections with a correlation coefficient exceeding 0.5.

The preprocessed beat dataset was partitioned into training, validation, and test sets using an 70:10:20 ratio. Data partitioning was performed on a beat-wise basis rather than a patient-wise basis. This design decision was made to establish a controlled intra-dataset comparison of SNN learning paradigms under identical sample dynamics rather than a clinical generalization study. We discuss the implications of beat-wise splitting regarding patient generalization in Section 6.

2.5 Training Setup, Optimization, and Dynamic Thresholding

All models were trained using an original pipeline designed to produce the initial benchmarks and an updated pipeline designed to handle class imbalance, prevent overfitting, and optimize decision boundaries. Presented in this study are only the results of the updated pipeline, although the results from both were comparable.

To prevent gradient bias toward the majority class during training, mini-batches were drawn using PyTorch's `WeightedRandomSampler`. This forces an approximately balanced 50:50 distribution of `N` and `V` beats in every training batch without altering the underlying loss function formulation. Validation and test sets maintained their original 10.3:1 natural class distribution to accurately evaluate real-world performance.

Models were optimized using the Adam optimizer with an initial learning rate of $\eta = 10^{-3}$. A dynamic learning rate scheduler reduced η by half (factor of 0.5) whenever validation performance plateaued for 3 consecutive epochs.

To prevent overfitting, early stopping was implemented based on the Precision-Recall Area Under the Curve (PR-AUC) evaluated on the un-resampled validation set. PR-AUC was chosen over standard accuracy or ROC-AUC due to its sensitivity under severe class imbalance. Training was halted if validation PR-AUC failed to improve for 8 consecutive epochs, and the model weights corresponding to the highest validation PR-AUC were restored for testing.

Gradient-based architectures (Base GNN-LSTM, GNN-SNN BPTT, and ANN-to-SNN conversion) were allocated a maximum budget of 30 epochs. In contrast, the unsupervised STDP model was capped at 10 epochs. This difference was dictated by computational complexity and learning dynamics:

1. **Computational Cost:** Due to BindsNet's unbatched, sample-level execution requirement (batch size $B = 1$), each epoch required significant processing time (~ 42 minutes per epoch, totaling 7 hours for 10 epochs).

2. **STDP Instability:** Empirical tracking revealed that the STDP model achieved its highest validation metrics during the first epoch. Subsequent epochs caused monotonic performance degradation, confirming that standard unsupervised Hebbian weight updates without supervised lateral inhibition or threshold stabilization cannot converge over multi-epoch temporal training.

Because models trained under 50:50 oversampling output predicted probabilities calibrated to a balanced prior, applying the default classification threshold (0.5) to un-resampled evaluation sets causes severe overprediction of the minority class (resulting in high false positives).

To restore optimal decision boundaries, threshold calibration was conducted on the validation set prior to test evaluation:

- Probability decision thresholds were swept from 0.05 to 0.95 in increments of 0.005.

- The optimal threshold T^* was selected by maximizing the Matthews Correlation Coefficient (MCC) on validation data. MCC was selected because it symmetrically evaluates all four quadrants of the confusion matrix and remains robust under class imbalance.

- The tuned threshold T^* was subsequently locked and applied to the test set for final reporting.

3 Experimental Setup and Results

This section describes our base architecture, based on which the three GNN-SNN architectures were developed, and the results from our experiments.

3.1 Base GNN-LSTM Architecture

As a baseline, we designed and implemented an end-to-end hybrid GNN-LSTM architecture using PyTorch and PyTorch Geometric (PyG). The model operates directly on raw graph-structured ECG windows without manual feature extraction and consists of three components:

- **Spatial GNN:** Composed of two graph convolution layers with ReLU activation functions to extract spatial inter-lead relationships at each timestep.
- **Temporal GNN:** Iterates sequentially through each timestep t in the ECG window, passing temporal slices through the Spatial GNN to form a sequence of 768-dimensional node-aggregated embeddings.
- **Hybrid Classifier Head:** Flattens the Temporal GNN output sequence and feeds it into a two-layer LSTM (hidden size 256), followed by a linear classification head for binary output.

After hyperparameter tuning, this baseline achieved a training accuracy of 99.88% and a testing accuracy of 99.75%.

3.2 GNN-SNN Architectures

To evaluate different spiking neural network paradigms, we redesigned the temporal integration head of the base GNN-LSTM into three distinct GNN-SNN architectures, maintaining end-to-end training across all variants. Each architecture adapts the presentation of spatial features to accommodate the specific learning mechanisms of its underlying SNN framework. All models preserved equivalent parameter capacities in their temporal heads for fair comparison.

3.2.1 BPTT with Surrogate Gradient

For gradient-based recurrent spiking, we implemented a GNN-SNN model using the snnTorch library. The spatial GNN feature extractor remains unchanged, while the temporal LSTM is replaced by a two-layer spiking recurrent network with Leaky Integrate-and-Fire (LIF) neurons (hidden size 256).

Information is presented sequentially across time steps. The temporal loop carries the continuous LIF membrane potentials forward across timesteps t, mirroring the hidden state persistence of an LSTM while matching its parameter count. To optimize the network via BPTT, fast sigmoid is utilized as a surrogate activation function during the backward pass. Rather than averaging discrete spike counts over short observation windows – which introduces high variance – classification is performed on the final LIF layer's continuous membrane potential. This final membrane voltage provides a smooth, differentiable continuous signal ideal for standard cross entropy loss optimization.

3.2.2 ANN-to-SNN Conversion

To leverage direct rate-coded conversion, a shadow ANN was implemented in SpikingJelly and converted to an SNN via threshold balancing. Because standard graph convolution operations are not natively supported in SpikingJelly's conversion pipeline, the Spatial GNN remains an ANN feature extractor, while the temporal head is converted to SNN format.

The temporal LSTM was replaced with a 1D CNN comprising three convolutional blocks ($768 \rightarrow 512 \rightarrow 256$ channels). The kernel sizes narrow progressively ($7 \rightarrow 5 \rightarrow 3$) to capture broad temporal context before fine-grained refinement. At the resampled sampling frequency ($f_s = 250$ Hz), a kernel size of 7 spans $\sim$ 27 ms, explicitly matching the duration of high-frequency ECG components such as the QRS complex. Three blocks provide sufficient depth without gradient saturation over short input windows ($T = f_s/2 \approx 125$ samples).

Information is presented as a 2D spatial-temporal feature map. Following convolution, a Global Average Pooling layer collapses the temporal axis into a 256-dimensional vector before passing to a linear classifier. Global average pooling naturally translates to mean firing rates in spiking space, making the representation optimal for the `ann2snn` conversion framework. The shadow network is trained using ReLU activations – ensuring mathematical equivalence to Integrate-and-Fire (IF) neuron firing rates (Pietrzak et al., 2023) – and is converted by calibrating maximum pre-activations across a calibration set (Rueckauer et al., 2017).

3.2.3 STDP

To evaluate local biological learning, we constructed a GNN-SNN model utilizing spike-timing-dependent plasticity (STDP) via the BindsNet library. The LSTM block is replaced with an unsupervised spiking layer consisting of 768 InputNodes connected to 256 LIFNodes via PostPre STDP connections, followed by a supervised `Linear(256, 2)` classification head.

Because BindsNet operates at the sample level, batch sizes are reduced to 1. At each timestep t, the continuous GNN feature embedding is converted into a probabilistic Poisson spike train and fed sequentially into the network.

To prevent activity state leakage between distinct ECG samples, a fresh BindsNet network instance is initialized per sample. However, to preserve unsupervised learning across the dataset without LSTM hidden states, we introduce a persistent weights buffer (768×256). This weight matrix is updated after each beat sample via an Exponential Moving Average, simulating slow Hebbian consolidation. STDP weight updates occur automatically and unsupervised during the forward pass. Supervised learning is isolated strictly to the final linear classifier, which maps the output LIF layer's final membrane voltages to class logits via

Adam optimization.

4 Results

Each paradigm was evaluated across three paired experimental runs utilizing distinct random data partitions and weight initializations governed by split seeds (42, 101, and 2023). All paradigms share identical data partitions to enable paired contrasts, thereby isolating paradigm behavior from dataset partition difficulty. Table 1 summarizes test-set performance, calibration, and training duration across all paradigms.

4.1 Ranking Discrimination vs. Calibration Dynamics

Under threshold-independent ranking metrics (PR-AUC and ROC-AUC), the two gradient-based paradigms perform comparably. Although shadow training achieves a higher marginal mean PR-AUC (0.779 vs. 0.697), the paired per-seed PR-AUC differences are $+0.008$, $+0.245$, and -0.007. The entire marginal gap is driven by a single partition (seed 101); excluding this outlier, shadow training averages 0.701 PR-AUC against 0.697 for BPTT with surrogate gradient. Given an observed within-paradigm spread of ± 0.028 for BPTT, performance differences below $\sim$ 0.06 PR-AUC fall within variance noise. Consequently, ranking quality between conversion and surrogate gradient optimization is statistically indistinguishable in this setup.

However, a marked performance divergence emerges in thresholded decision boundaries and output calibration:

- **Minority Class Detection:** Shadow training achieves a `V`-class recall of 0.933 ± 0.038 compared to 0.610 ± 0.027 for BPTT. The run ranges do not overlap.
- **Calibration Error:** Expected Calibration Error (ECE) for shadow training (0.007 ± 0.004) is an order of magnitude lower than BPTT (0.073 ± 0.016).

Because both paradigms order beats similarly well, this divergence indicates that their output probabilities occupy fundamentally different scales. Validation-tuned decision thresholds (T^*) transfer effectively to test sets under shadow training but underperform under BPTT. For BPTT, the selected operating point ($T^* = 0.825$) forfeited approximately one-third of available minority beat recall.

4.2 Partition Stability and Supervised Pre-Training

The paired experimental design reveals a distinct difference in stability across data partitions. Across the three seeds, BPTT exhibits consistent average precision (0.674, 0.688, 0.728; $\sigma = 0.028$). Conversely, shadow training displays high dispersion (0.682, 0.933, 0.721; $\sigma = 0.135$) – a variance nearly five times larger on identical splits.

This instability originates during the supervised source ANN pre-training phase rather than during conversion. The source ANNs achieved validation average precision scores of 0.575, 0.974, and 0.861 across partitions. Thus, while shadow training provides superior mean calibration and reduced computational overhead, individual runs exhibit lower predictability compared to end-to-end BPTT.

4.3 Unsupervised STDP Failure Modes

The unsupervised STDP paradigm failed to establish a functional decision boundary, yielding a PR-AUC of 0.133 ± 0.017 (near the no-skill baseline of 0.0885) and ROC-AUC of 0.603 ± 0.011. Because local Hebbian updates operate without class label supervision under severe imbalance ($10.3 : 1$), synaptic weight updates are dominated by majority-class waveform dynamics. Multi-epoch training exacerbated this failure: validation performance peaked during Epoch 1 and deteriorated monotonically thereafter, confirming that unsupervised local plasticity without lateral inhibition or supervised error feedback cannot converge on imbalanced sequence classification tasks.

5 Sensitivity Analysis and Readout Ablation

To determine whether the superior calibration of ANN-to-SNN conversion stems from its architecture or its output readout mechanism, we conducted an ablation study. BPTT outputs classification logits directly from unbounded, unnormalized LIF membrane potentials. Conversely, the converted SNN classifies from rate-normalized spike counts accumulated over time.

The conversion pipeline provides an internal control: the exact same trained parameters and extracted GNN features were evaluated under two conditions: (1) as the source ANN with a continuous logit readout, and (2) as the converted SNN with a spike-rate readout. Table 2 details the paired results.

The ablation separates ranking quality from calibration:

1. **Ranking Quality:** Rate coding slightly reduces ROC-AUC across all partitions, reflecting discrete quantization loss during conversion. PR-AUC paired differences alternate in sign ($+0.098$, -0.033, -0.150), indicating no systematic ranking advantage for rate readouts.
2. **Calibration Improvement:** Calibration improves consistently across every partition. ECE

Table 1. Test-Set Performance and Efficiency Across Three Paired Runs (Split Seeds 42 / 101 / 2023)

Paradigm	Parameters	PR-AUC	ROC-AUC	V-Recall	V-F1 Score	Accuracy	ECE	Train Time (h)
Shadow Training (Conversion)	1,907,650	**0.779 ± 0.135**	**0.966 ± 0.016**	**0.933 ± 0.038**	**0.860 ± 0.087**	**0.973 ± 0.015**	**0.007 ± 0.004**	**2.52 ± 0.37**
BPTT w/ Surrogate Gradient	267,460	0.697 ± 0.028	0.903 ± 0.023	0.610 ± 0.027	0.657 ± 0.015	0.940 ± 0.000	0.073 ± 0.016	5.03 ± 0.04
STDP (Unsupervised)	4,802 + 196,608*	0.133 ± 0.017	0.603 ± 0.011	–	–	–	–	7.91 ± 0.20

* Indicates 4,802 trainable parameters (GNN + classification head) and 196,608 persistent unsupervised STDP weight buffer parameters.

Table 2. Readout Ablation: Source ANN (Logit Readout) vs. Converted SNN (Rate Readout) across Three Paired Partitions using Identical Weights and Features.

Metric	ANN (Logit)	SNN (Rate)
PR-AUC	0.807 ± 0.199	0.779 ± 0.135
ROC-AUC	0.982 ± 0.011	0.966 ± 0.016
Sensitivity	0.923 ± 0.042	0.933 ± 0.036
ECE	0.032 ± 0.013	**0.007 ± 0.004**

drops by factors of 6.5, 4.7, and 3.9 under rate readout, with non-overlapping ranges (0.018–0.043 for logit vs. 0.0038–0.0112 for rate).

This demonstrates that accumulating discrete spikes over time naturally normalizes output variance, yielding well-calibrated probabilities that transfer robustly across decision thresholds.

6 Threats to Validity

We identify several threats to validity:

- **Sample Size and Significance:** Evaluation across three paired seeds provides descriptive spread and paired ordering but lacks statistical power for formal significance testing.
- **BPTT Optimization Constraints:** The BPTT model reached the 30-epoch ceiling without triggering early stopping. Reported BPTT metrics represent a lower bound on performance; extended training budgets could narrow the observed calibration gap, though at increased computational cost.
- **Beat-Wise Data Partitioning:** Splitting data at the beat level rather than the patient level allows subject-specific morphological features to appear in both training and test sets. While valid for comparative benchmarking of SNN learning paradigms under identical feature distributions, absolute metrics should not be interpreted as clinical cross-subject generalization performance.

7 Code Availability

To promote scientific transparency and reproducibility, all source code is publicly available at: https://github.com/xcambald/snn-gnn-paradigms.

8 Discussion

The experimental results demonstrate that temporal learning rules dominate overall performance and operational reliability in hybrid GNN-SNN architectures. While graph convolutions effectively capture spatial inter-lead relationships, temporal integration under severe class imbalance remains the primary bottleneck.

ANN-to-SNN conversion offers an advantageous efficiency-performance trade-off. Conventional optimization avoids surrogate gradient approximations, while rate-coded spike accumulation provides intrinsic regularizing effects that yield well-calibrated output probabilities. However, initialization sensitivity during source ANN training requires multiple restarts to ensure stability.

For surrogate-gradient BPTT, optimization instability and poor probability scaling highlight the need for specialized loss formulations (e.g., focal loss or class-weighted cross-entropy adapted for membrane dynamics) and dynamic neuron parameters (such as learnable firing thresholds or adaptive decay rates). Unsupervised STDP in its standard formulation is unsuited for imbalanced temporal classification, necessitating supervised feedback mechanisms such as reward-modulated STDP.

9 Conclusion

This study presented a controlled comparative analysis of three SNN learning paradigms integrated with spatial GNN encoders for ECG arrhythmia classification. Under realistic class imbalance (10.3 : 1), the ANN-to-SNN conversion paradigm achieves superior classification performance, minority beat recall, calibration, and training efficiency. End-to-end BPTT achieves competitive discrimination ranking but suffers from output miscalibration. Unsupervised STDP fails to converge to a usable decision boundary. Future work should focus on neuromorphic graph operations, hardware deployment metrics, and advanced gradient formulations tailored to event-driven processing.

Acknowledgments

This work was supported by the Slovak University of Technology in Bratislava (STU) under the internal grant scheme “Mladý výskumník” (Project No. 1364) awarded to Dominik Čambál.

References

Ayasi, B., Carmona, C. J., Saleh, M., & García-Vico, A. M. (2025). A Practical Tutorial on Spiking Neural Networks: Comprehensive Review, Models, Experiments, Software Tools, and Implementation Guidelines. *Eng*, *6*(11), 304. https://doi.org/10.3390/eng6110304

Bulíková, T. (2014). *EKG pre záchranárov nekardiológov* [ECG for Non-Cardiologist EMS Workers]. Grada Slovakia; Grada.

Chu, H., Yan, Y., Gan, L., Jia, H., Qian, L., Huan, Y., Zheng, L., & Zou, Z. (2022). A Neuromorphic Processing System With Spike-Driven SNN Processor for Wearable ECG Classification. *IEEE Transactions on Biomedical Circuits and Systems*, *16*(4), 511–523. https://doi.org/10.1109/tbcas.2022.3189364

Eshraghian, J. K., Ward, M., Neftci, E. O., Wang, X., Lenz, G., Dwivedi, G., Bennamoun, M., Jeong, D. S., & Lu, W. D. (2023). Training Spiking Neural Networks Using Lessons From Deep Learning. *Proceedings of the IEEE*, *111*(9), 1016–1054. https://doi.org/10.1109/jproc.2023.3308088

Hannun, A. Y., Rajpurkar, P., Haghpanahi, M., Tison, G. H., Bourn, C., Turakhia, M. P., & Ng, A. Y. (2019). Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine*, *25*(1), 65–69. https://doi.org/10.1038/s41591-018-0268-3

Hebb, D. O. (1949). *The Organization of Behavior: A Neuropsychological Theory*. John Wiley & Sons.

Kligfield, P., Gettes, L. S., Bailey, J. J., Childers, R., Deal, B. J., Hancock, E. W., van Herpen, G., Kors, J. A., Macfarlane, P., Mirvis, D. M., Pahlm, O., Rautaharju, P., & Wagner, G. S. (2007). Recommendations for the Standardization and Interpretation of the Electrocardiogram. *Journal of the American College of Cardiology*, *49*(10), 1109–1127. https://doi.org/10.1016/j.jacc.2007.01.024

Ma, H., & Xia, L. (2023). Atrial Fibrillation Detection Algorithm Based on Graph Convolution Network. *IEEE Access*, *11*, 67191–67200. https://doi.org/10.1109/ACCESS.2023.3291352

Malcolm, K., & Casco-Rodriguez, J. (2023). A Comprehensive Review of Spiking Neural Networks: Interpretation, Optimization, Efficiency, and Best Practices. https://doi.org/10.48550/ARXIV.2303.10780

Pietrzak, P., Szczęsny, S., Huderek, D., & Przyborowski, Ł. (2023). Overview of Spiking Neural Network Learning Approaches and Their Computational Complexities. *Sensors*, *23*(6), 3037. https://doi.org/10.3390/s23063037

Rueckauer, B., Lungu, I.-A., Hu, Y., Pfeiffer, M., & Liu, S.-C. (2017). Conversion of Continuous-Valued Deep Networks to Efficient Event-Driven Networks for Image Classification. *Frontiers in Neuroscience*, *11*. https://doi.org/10.3389/fnins.2017.00682

Tihonenko, V., Khaustov, A., Ivanov, S., & Rivin, A. (2007). St.-Petersburg Institute of Cardiological Technics 12-lead Arrhythmia Database. https://doi.org/10.13026/C2V88N

Wang, B., & Jiang, B. (2022). Spiking gats: Learning graph attentions via spiking neural network. https://arxiv.org/abs/2209.13539

Yi, Z., Lian, J., Liu, Q., Zhu, H., Liang, D., & Liu, J. (2023). Learning rules in spiking neural networks: A survey. *Neurocomputing*, *531*, 163–179. https://doi.org/10.1016/j.neucom.2023.02.026

Zenke, F., & Vogels, T. P. (2023). The Remarkable Robustness of Surrogate Gradient Learning for Spiking Neural Networks. *Neural Computation*, *35*(4), 899–925. https://doi.org/10.1162/neco_a_01566

Zhang, A., Lipton, Z. C., Li, M., & Smola, A. J. (2023, December). *Dive into Deep Learning*. Cambridge University Press.

Zhu, Z., Peng, J., Li, J., Chen, L., Yu, Q., & Luo, S. (2022). Spiking graph convolutional networks. https://arxiv.org/abs/2205.02767

From News Text to Causal Infographics via Semantic Triplets: Making an Opaque Image Pipeline Inspectable

Georgi Trajkov, Marko Grobelnik, Dunja Mladenić
Jožef Stefan Institute
Department for Artificial Intelligence
Jamova cesta 39, 1000 Ljubljana, Slovenia
geotrajkov0@gmail.com, {marko.grobelnik, dunja.mladenic}@ijs.si

Abstract. *Generative text-to-image models can turn news articles into a causal-network infographic, but the graph lives only inside the model – it cannot be inspected or corrected. We insert a semantic-triplet stage between text and image and compare direct and triplet-mediated renders on ten Slovene news events, under a rubric that separates triplet correctness from image faithfulness. We hypothesized that decomposition would improve both faithfulness and controllability; in our preliminary single-rater evaluation only controllability holds – the direct image is at least as faithful in every event. The triplet stage's value is auditability: extraction is roughly 70 percent clean, while the renderer mis-draws roughly 19 percent of an otherwise-correct graph's edges – an error the black box hides.*

Keywords. generative AI, text-to-image generation, semantic triplets, knowledge graph construction, causal graphs, compositional fidelity, infographic generation, RDF, large language models, explainability

1 Introduction

A news event of any complexity is covered by many articles across many publishers, and its causal structure – who did what, what followed from what – is spread over dozens of paragraphs. An infographic condenses that structure into a single image a reader can grasp at a glance, which is why editorial node – link graphics are a staple of data-driven outlets; generative image models now make them producible automatically, at the scale of whole event streams. Aggregator platforms such as Event Registry (Event Registry, 2025; Leban et al., 2014) and Ground News (Ground News, 2025) cluster news around events and expose bias and sentiment indicators. Building on Event Registry, our production system BetweenTheLines (MedVrsticami, https://medvrsticami.ijs.si) (Trajkov et al., 2025) analyzes how individual claims and entities are portrayed across publishers; its key lesson is that decomposing a hard large language model (LLM) task into staged prompts – extract, then classify – sharply improves accuracy. We carry that lesson to a different output: a *causal-network infographic*, a single editorial-quality image that lays out the actors and events of a story as a node – link diagram. The infographic module we built into BetweenTheLines generates such an image for each analyzed event, directly from the article text by a text-to-image model in a single step – and that direct text→image generation, our own production design, is the baseline we audit here.

In this one-step setup the article text goes in and a picture comes out, but the graph the picture depicts – which entities appear as nodes, which causal relations appear as edges, and in which direction – is never materialized; it exists only implicitly, inside the image model. This makes the artifact impossible to inspect, query, correct, or evaluate except by eye, and any error (a reversed arrow, an invented link, a dropped relation) is indistinguishable from a deliberate editorial choice.

This paper asks what is gained, and what is lost, when an explicit and inspectable knowledge-graph stage is inserted between the text and the image. We add a semantic-triplet extraction step – subject – predicate – object relations with stable entity identifiers, mirroring the Resource Description Framework (RDF) (W3C, 2014) – and feed the resulting graph to the renderer instead of the raw text. In other words, we test whether decomposing the image task the way BetweenTheLines (Trajkov et al., 2025) decomposed claim classification – extract first, then render – yields a similar gain.

We entered the study with a directional hypothesis: a triplet-mediated pipeline would improve both the *faithfulness* of the rendered infographic and the *controllability* of the pipeline, because the relations can be checked and edited before they are drawn. Our results support only the second half. On image quality the explicit graph does not help – the direct baseline is at least as faithful in every event we examined – but the decomposition delivers exactly the controllability and inspectability the hypothesis anticipated, and that is where its value lies.

Our contributions are: (1) a triplet schema and extraction prompt that produce connected, source-grounded causal graphs with confidence and source-

language constraints; (2) a two-condition comparison (direct vs. triplet-mediated image, same image model, same style prompt) under a two-part rubric that separates triplet correctness from rendered-image faithfulness on the same events; and (3) the finding that, in this preliminary single-rater evaluation, the decomposition's value is transparency and fault-localization rather than image quality – it separates the two error sources the black box conflates, localizing the largest avoidable failure to the generative rendering step.

2 Related Work

Cross-source news analysis. Ground News (Ground News, 2025) classifies publishers on a left – right spectrum and aggregates coverage, but does not track individual claims or produce structured graphs. Event Registry (Event Registry, 2025; Leban et al., 2014) clusters articles into events and provides article-level sentiment (Hutto & Gilbert, 2014). BetweenTheLines (Trajkov et al., 2025) builds on Event Registry with multi-prompt LLM analysis of entity and claim portrayal, and showed that decomposing the task into staged prompts – extraction then classification – markedly improved accuracy; the present study tests whether the same decomposition helps for infographic generation.

Relations and graphs from text. Extracting structured relations from natural language is the classic domain of Open Information Extraction (Banko et al., 2007; Fader et al., 2011), which produces subject – relation – object tuples without a fixed schema, and is conventionally evaluated by the precision and recall of the extracted tuples against a reference. More recently, large language models have largely taken over knowledge-graph and triplet construction (Bian, 2025), and have been applied specifically to extracting causal graphs from text in a zero-shot fashion (Antonucci et al., 2024), where LLM-derived priors are found especially useful for orienting edges (Darvariu et al., 2024) – a step our results show the image renderer often gets wrong. Our triplets are an LLM-produced, schema-light variant, but our aim differs: we extract the causal graph in order to render it as an infographic – and to measure how faithfully a downstream image model redraws it – rather than to populate a database. The entities map naturally to RDF subjects/objects and the predicates to RDF properties (W3C, 2014), which is what makes the intermediate queryable and standardizable in principle.

Text-to-image fidelity for diagrams. Text-to-image models render legible editorial graphics but give no guarantee of structural fidelity to a specified graph. General compositional benchmarks show that these models place plausible objects yet fail to bind the *relations* between them: VISOR (Gokhale et al., 2022) measures how often a stated spatial relation is actually realized, T2I-CompBench (Huang et al., 2023) covers attribute and relation binding, and the weakness persists on current frontier models (Huang et al., 2025; Wang et al., 2026). Diagram- and infographic-specific work confirms it directly: IGenBench (Tang et al., 2026) decomposes infographic reliability into atomic yes/no checks and finds even frontier models rarely satisfy every element of a target infographic, and DiagrammerGPT (Zala et al., 2024) shows these models fail on densely connected arrows and labels – which is why it plans an explicit layout and renders structure-critical content deterministically rather than trusting the generative model. Structural fidelity to a supplied graph is precisely what we measure here.

Graph-drawing readability. How readable a node – link diagram is has long been formalized by the graph-drawing community: Purchase (2002) defines seven measurable aesthetic criteria (edge crossings – the dominant factor – bends, symmetry, angular resolution, and so on). Our single relation-clarity rating corresponds to what this literature measures; a richer evaluation could score the criteria separately and, on a deterministic render, compute them objectively.

3 The Current Infographic Pipeline

For an event, our production step in BetweenTheLines concatenates the title and body of every article in the event cluster (capped at the ten most relevant articles) into one text block and sends it to a text-to-image model with a single prompt that asks the model to both *extract* a causal knowledge graph and *render* it as a Bloomberg/Economist-style node – link infographic (Listing 1). The two responsibilities – deciding the graph and drawing it – are fused in one opaque call. The output is a raster image; no node list, edge list, or evidence is emitted. There is therefore no way to ask which relations the picture claims, whether they are supported by the articles, or whether the drawing even matches the model's own intended graph.

Listing 1. Excerpt of the production prompt: a single call both extracts and renders the graph from the concatenated articles.

```
### Content Processing Rules
Extract a Knowledge Graph: Nodes = Entities +
Claims/Events; Links = Causal Relations (Cause
    -> Effect).
### Input Text
{articles}   # title + body of every article in
    the event
```

4 Proposed Method

We insert an explicit triplet stage and render from it (Fig. 1).

Triplet schema. Extraction returns a JSON object with an `entities` list (stable `id`, human-readable `label`, one `central` flag) and a `triplets` list (`subject_id`, `predicate`, `object_id`, `evidence`, `confidence`), where `evidence` is a verbatim support-

ing text span from the articles. Because subjects and objects reference entity `ids`, reuse across triplets is structural, not lexical – the graph is connected by construction. The `entities` map to RDF subjects/objects and the `predicate` to a (currently free-text) property (W3C, 2014).

Extraction prompt. It targets 10 – 15 triplets over 6 – 10 entities – a deliberate cap to keep the rendered graph legible – and enforces: one connected component, no singletons; specific predicates (no generic "affects/relates to"); a *source-language lock* (labels, predicates and evidence stay in the article language); *event reification* (a death or ruling becomes its own node, not collapsed onto a person); and *attribution-in-predicate* for contested claims (e.g. *naj bi*, "allegedly"), so a one-sided allegation is not stated as fact. A validator checks connectivity and id consistency, retrying once. The few-shot examples use a fictional scenario, so none of the ten evaluation events appears in the prompt (no train-on-test leakage); the provenance of every prompt revision is retained.

Confidence filtering. Each triplet carries a low/medium/high confidence tag, and low-confidence triplets are dropped before rendering. In our ten events this filter is barely exercised: 95 of 115 triplets were tagged high, 19 medium, and only 1 low, so a single triplet was dropped (leaving the 114 edges sent to the renderer) – too few low-confidence cases to judge how reliably the flag catches errors. We keep the mechanism because exposing a per-triplet confidence the pipeline can act on is itself a controllability lever the black box lacks.

Two rendering paths (evaluated). (1) *Direct*: the unchanged baseline, text → image. (2) *Triplet → image*: the same image model and the same visual-style prompt, but fed the explicit entity and edge lists instead of raw text and told to draw only those (Listing 2). Both use the same image model, so the only difference is the input.

Listing 2. Excerpt of the triplet→image prompt: the graph is supplied and the model is told to render *only* it – no extraction, no invention (contrast Listing 1).

```
You MUST use ONLY the entities and relations
     supplied below.
Do NOT introduce any node not in the entities
     list, invent
any edge not in the relations list, or translate
      a label.
### Entities (nodes)     {entities}   # id,
     label, central
### Relations (edges)    {relations}  # subject
     -> pred -> object
```

A third, programmatic path. Because the triplets are now explicit, they could also be rendered programmatically (e.g. a deterministic `graphviz` layout (Gansner & North, 2000)). Such a renderer reproduces the supplied graph with perfect structural fidelity by construction – there is no generative binding step to get wrong – so it removes the render-binding error entirely and collapses the pipeline's error onto the extraction stage. Diagram-generation systems already adopt this discipline (Zala et al., 2024). As a proof of the path we did produce a basic deterministic render (a `graphviz`-layouted SVG) of every evaluated event; they are included in the supplementary viewer, and they reproduce their graphs exactly. Producing a result as readable and information-dense as the generative image, however, requires a typed node ontology and an icon/visual-encoding library (flags, currency, institutional symbols) – substantial design work that is out of scope here. We therefore score only the two generative conditions and treat styled programmatic rendering as a direction enabled by, but not delivered in, this study.

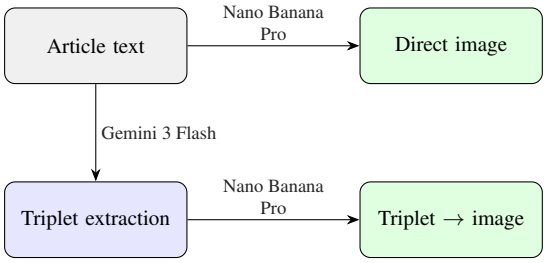


Figure 1. The two evaluated conditions. The direct baseline fuses extraction and rendering in one opaque call; the triplet condition exposes the graph, then renders it with the same image model. Extraction uses Gemini 3 Flash; both image renders use Gemini 3 Pro Image (*Nano Banana Pro*).

5 Evaluation Design

Data. Ten Event Registry events, each with 5 – 10 Slovene-language articles (mean 9.0 articles per event, mean article body ≈1,571 characters); per-event statistics are in Table 1. For triplet extraction, each article body is capped at 3,000 characters – a constraint inherited from BetweenTheLines; every article is always included, only the tails of individual long articles are cut. The cap touches 16 of the 90 articles (11.4% of corpus characters, concentrated in three events). The direct condition is the unchanged production baseline and receives the full text; Section 7 examines this asymmetry and shows the direct image's advantage is not explained by the extra input.

Models. Extraction uses `gemini-3-flash-preview` in JSON mode at temperature 0.2, with article bodies capped at 3,000 characters as above. Both image conditions use the same model, `gemini-3-pro-image-preview` ("Nano Banana Pro"), at 2K resolution and 16:9 with identical settings, so the rendering step itself is held constant. The comparison is deliberately pipeline-level rather than input-controlled: the triplet pipeline caps entities and relations while the direct baseline is free-form, and the extraction stage caps article length while the baseline does not. These asymmetries are part of what is being compared – each

Table 1. Per-event input statistics. Articles are Slovene; chars is the mean article-body length.

Event	Articles	Mean chars
E1	5	1,278
E2	7	920
E3	8	1,151
E4	10	1,088
E5	10	956
E6	10	967
E7	10	1,210
E8	10	1,825
E9	10	3,069
E10	10	3,247
Mean	9.0	1,571

condition runs its production configuration – and their consequences are analyzed in Section 7.

Rubric (two parts). A single rater – one of the authors – scores each event through a purpose-built side-by-side interface that shows the source articles, the extracted triplets, and both renders; scores are exported as per-rater JSON, so independent raters can later be added without changing the tooling. The rubric has two parts because the two pipelines expose two different things to inspect.

The first part is *the final image*, which both pipelines produce and which is judged against the articles: faithfulness (1 – 5), relation clarity (1 – 5), and a node count for density. Faithfulness scores how well the image is grounded in the source articles and how much of the story it manages to convey: 5 means the depicted content is fully supported by the articles and all important details of the story are communicated; 1 means hallucinated or wrong information and missing crucial points. An image that is fully grounded but omits crucial details therefore scores lower. Relation clarity scores how readable and digestible the drawn structure is – whether it is clear what happened, what led to what, and who is connected to whom; it corresponds to what the graph-drawing literature formalizes as readability (Purchase, 2002), compressed into a single rating: 5 means the flow of events and relations is immediately clear and the nodes look as they should; 1 means the image is effectively unreadable – overlapping arrows, confusing wording, an unclear flow of events, or icons and visual elements that confuse rather than clarify. The node count is taken by eye from the rendered image in both conditions. For the direct image this single number is all its hidden graph exposes; for the triplet image the accounting is finer-grained – the per-edge render verdicts of the second part record which supplied elements were drawn, omitted, or added – and the supplied graph size (8 in every event) cross-checks the count.

The second part is *the explicit graph*, which exists only in the triplet pipeline; because it is inspectable it can be scored partly objectively. Each triplet receives a per-triplet verdict on a severity scale (correct / *minor*: misleading, awkward, imprecise / *error*: wrong direction, wrong predicate, wrong endpoints, unsupported, redundant), judged against the articles – this measures extraction quality much as precision does for Open Information Extraction (Banko et al., 2007). Event-level coverage (1 – 5) captures the recall side – how much of the story the graph covers: 5 means all main strands are present, 3 means the core is covered but side strands are missing, 1 means major aspects are absent. Finally, each supplied edge gets a per-edge render verdict (drawn correctly, wrong direction, wrong connection, omitted) recording how the image redrew it, plus a count of unsupported additions; this per-element check mirrors the atomic verification used by infographic-reliability benchmarks (Tang et al., 2026). The direct pipeline has no such intermediate, so this second part has nothing to score there – which is exactly the asymmetry the study is about.

6 Results

Ten events, a single rater; 115 triplets (mean 11.5 per event) and 114 rendered edges scored.[1] The evaluation is preliminary – one rater, one render per condition – so we report means, sign consistency, and per-element counts rather than significance tests.

6.1 The triplet image does not win

The direct image is more faithful (4.10 vs. 3.10, mean difference 1.0) and clearer (3.60 vs. 2.80) (Table 2). The gap is not large, but it is consistent: the triplet image ties the direct image on faithfulness in 5 of the 10 events and never exceeds it – direct $\geq$ triplet in 10/10 events. With only ten paired events and five ties, we do not claim statistical significance; the robust, defensible statement is the sign consistency. The one decisive direct advantage is density: 18.4 vs. 8.1 nodes counted by the same by-eye procedure, a $2.3\times$ gap. Every one of the ten supplied graphs contained exactly 8 entities; in nine events the render drew exactly those 8 nodes, and in one it drew 9, duplicating a supplied node into two boxes. The gap is structural – the direct image freely promotes quantities, dates and outcomes to nodes, whereas the triplet schema admits only typed entities.

6.2 Extraction is mostly sound

Of the 115 rated triplets, 69.6% are correct and 25.2% have only minor issues (chiefly *imprecise* or *misleading* phrasing); 5.2% are hard errors, dominated by *unsupported* relations (3.5%) (Table 3). Event-level coverage averages $3.7/5$: the 10 – 15-triplet cap that keeps

[1] All extracted triplets, generated graphs, prompts, source articles, and rubric scores – with an interactive read-only viewer and a full data download – are available at https://causal-infograph.netlify.app.

Table 2. Final-image quality and density, direct vs. triplet→image (10 events, single rater). Faithfulness/clarity 1 – 5 (higher is better); mean ± SD. Nodes drawn is counted by eye from the rendered image in both conditions; each triplet condition's supplied graph had exactly 8 entities.

Metric	Direct	Triplet→image
Faithfulness (1 – 5)	4.10 ± 0.74	3.10 ± 0.74
Relation clarity (1 – 5)	3.60 ± 0.97	2.80 ± 1.03
Nodes drawn	18.4 ± 8.0	8.1 ± 0.3

Table 3. Per-triplet verdicts by severity tier (115 triplets). This scores each *extracted* triplet against the articles – whether the relation is correct – independent of how the image later drew it (Table 4).

Verdict	Count	Share
Correct	80	69.6%
Minor	29	25.2%
imprecise	15	13.0%
misleading	10	8.7%
awkward wording	4	3.5%
Error	6	5.2%
unsupported	4	3.5%
wrong endpoints	1	0.9%
redundant	1	0.9%

graphs legible also forces real omissions, the recall cost of the same budget that drives the density gap.

6.3 The renderer is the largest avoidable error

Given the explicit graph, the image model draws 80.7% of edges correctly, but 14.9% are *wrong connections* – an edge attached to the wrong node, the single most frequent failure across the whole pipeline – with a further 2.6% direction-flipped, 1.8% omitted, and 2 hallucinated edges plus 1 duplicated node added across the ten events (Table 4).

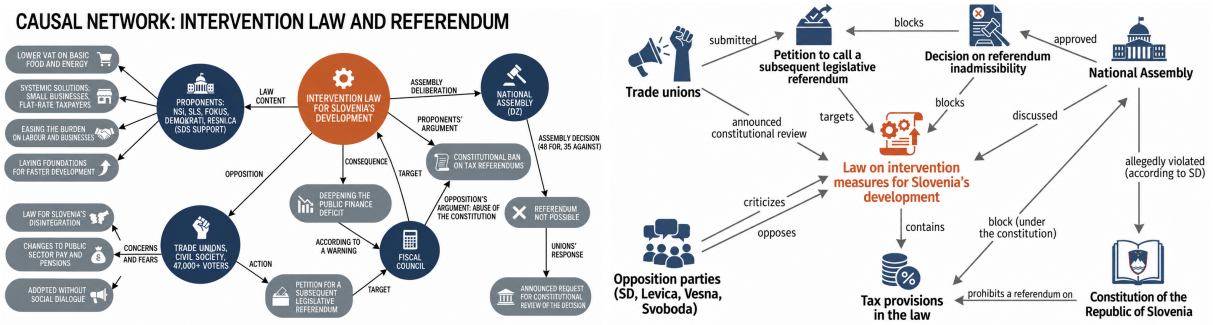
Fig. 2 is the illustrative case. Its triplets were rated "very good" and its triplet→image tied the direct image on *faithfulness* (4/5 each); what was poor was its *render fidelity* – only 8 of 12 supplied edges were drawn to the right nodes. Concretely, the four mis-connected edges were: Tax provisions → block (per constitution) → Referendum petition, whose arrow instead terminated at the National Assembly; two edges that should both end on the Inadmissibility-decision node – Opposition → opposes and Unions → announce constitutional review – which were attached to the Intervention-law node (the graph's central, most prominent node) instead; and a fourth, running *from* the decision to the Constitution, which was re-rooted at the Assembly. The pattern is that the model attaches an arrow to the wrong – typically nearer or more prominent – node, drawing a structurally different graph from the one supplied. This is the relational/spatial-binding weakness that text-to-image models are independently documented to exhibit on general compositional scenes (Gokhale et al., 2022; Huang et al., 2023), that persists on current frontier models (Huang et al., 2025; Wang et al., 2026), and that diagram-specific work confirms for node – link rendering in particular (Tang et al., 2026; Zala et al., 2024). A programmatic render would not suffer this class of error, the practical motivation for treating it as future work.

Table 4. Triplet→image render fidelity: how the model drew each of 114 supplied edges, plus 2 hallucinated additions.

Rendered as	Count	Share
correctly drawn	92	80.7%
wrong connection	17	14.9%
wrong direction	3	2.6%
omitted	2	1.8%

6.4 Drawing style: uniform triplet renders vs. tailored direct renders

Both conditions receive the identical visual-style prompt, so any difference in appearance comes from the input, not the specification. Alongside the uniform quality gap (direct ≥ triplet on faithfulness in 10/10 events and on clarity in 9/10), the two paths differ in visual character in a way that mirrors the density gap. The triplet renders are uniform and skeletal: every event hands the model the same kind of input – a fixed list of 6 – 10 entities and 10 – 15 edges – so it simply lays out the supplied skeleton, and the ten renders come out structurally alike. The direct render instead takes liberty and fits a richer graph to each event. Fig. 3 is typical: fed the raw articles, the direct image promotes event-specific detail to nodes – passenger counts, a timestamp, a casualty breakdown, separate emergency-service and media clusters – in a bespoke multi-flow layout, whereas the triplet image of the same event is a plain eight-node chain. This is the same input freedom that produces the density gap of Section 6.1, now visible as composition: fed raw articles the model has both the most to work with and the most latitude, and the result is a looser, less rigid layout, whereas the triplet render is locked to the supplied skeleton. Narrowing the gap is therefore mainly a matter of giving the triplet renderer comparable latitude – a broader node ontology, or licence to add supporting detail.



(a) Direct (text→image) **(b)** Triplet→image

Figure 2. Hero event (the intervention-law referendum). The triplets are sound and (b) tied (a) on faithfulness, but the model mis-connected four of its twelve edges – "good triplets, wrong renders." The direct image (a) is denser and more polished. Labels are translated into English for publication; the depicted connections, including rendering errors, are retained.

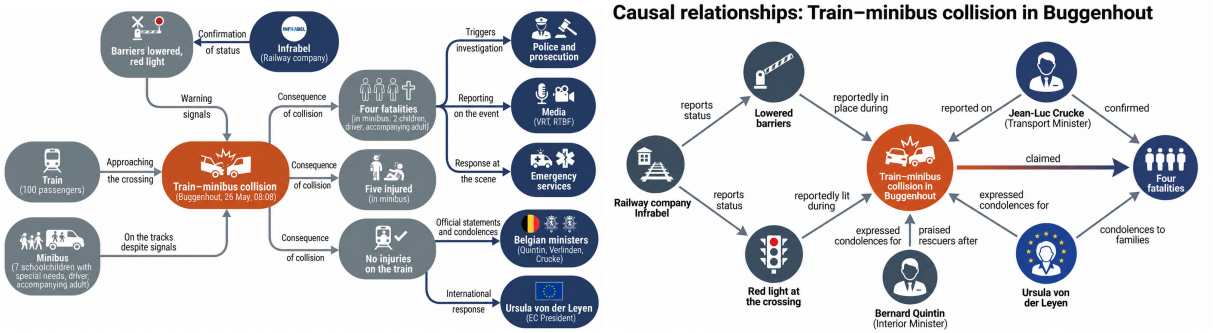
7 Discussion

The result reframes the contribution. Decomposition did not produce a better infographic – on a black-box basis the direct baseline wins – but it turns a single unverifiable image into an *auditable chain*: extraction is $\sim$70% clean ($\sim$5% hard errors), while the rendering stage mis-draws $\sim$19% of edges, predominantly wrong connections. The black box conflates these two error sources; the explicit graph pulls them apart, localizing the largest avoidable failure to generative rendering, not extraction. This is the Semantic-Web payoff in miniature: a standardizable intermediate enables verification that an end-to-end model forecloses.

The density gap follows from the triplet budget – but closing it is not free. The direct prompt asks for "key Entities *and* Key Claims/Events" and caps neither node nor edge count, so the model freely promotes quantities, dates and outcomes to nodes (mean 18.4); the triplet pipeline caps both axes (6 – 10 typed entities, 10 – 15 triplets) and every one of the ten events converged on exactly 8 entities. The same cap that limits density also limits coverage (the $3.7/5$ above). Is the lower node count also the reason for the triplet image's lower relation clarity? The per-event scores suggest not: within the triplet condition every graph has the same 8 nodes, yet clarity varies from 2 to 5 and tracks render fidelity – the correlation between clarity and the share of correctly drawn edges is $r = 0.75$ – while in the direct condition node count and clarity are essentially unrelated ($r = -0.18$; the densest render, 40 nodes, scores only 3). Low clarity comes from mis-drawn edges, not from sparsity. Could the extractor simply be encouraged to produce more nodes? Technically yes – the cap is a prompt parameter – but our render results argue against treating density as a dial one can simply turn up: the model already mis-binds $\sim$19% of edges on these sparse eight-node graphs, and adding nodes means more – and longer – edges, which would most likely make the binding worse, not better. Raising density dependably would therefore require more than a larger target count: extraction rules that yield graphs whose connections render reliably (e.g. locally clustered, layout-friendly structure), or a constrained/programmatic layout that removes the binding error outright. Density and render fidelity trade off, so the rendering failure must be addressed before added density would help rather than hurt.

Combining the two approaches. The pipelines need not be alternatives. The explicit graph can serve the better-looking direct image rather than replace it: (i) *triplets as a QA layer* – render the direct image for density and polish, and use the extracted graph to audit it for missing, unsupported or reversed relations; (ii) *hybrid input* – feed the model both the text (for style and density) and the triplets (as a checklist of relations that must appear, drawn in the stated direction); (iii) *structure-locked styling* – fix the layout with a deterministic graph renderer and let the image model only style and iconify it, removing the binding error while keeping the editorial look; and (iv) *wider triplet latitude* – broaden the node ontology or let the extractor add supporting detail, so the triplet render can tailor a richer graph per event the way the direct one does (Section 6.4). Options (i) and (iii) are the most promising: the first keeps today's better image and adds inspectability for free, the second attacks the dominant error directly.

An asymmetry in the evaluation itself. For the triplet image the rater knows the intended graph, so every mis-binding is findable; for the direct image there is no ground truth, so its structural errors are inherently harder to spot. The direct image's faithfulness (4.10) is therefore plausibly *over*-estimated relative to the triplet image, and the measured gap is in part an artifact of exactly the asymmetric inspectability this paper argues for. Far from undermining the result, this sharpens it: the black box's errors are invisible even to a careful human evaluator, which is the strongest possible case for materializing the graph.



(a) Direct – rich, event-tailored

(b) Triplet→image – uniform, skeletal

Figure 3. Same style prompt, different composition (a fatal train – minibus collision). Fed the raw articles, the direct render (a) tailors a rich, event-specific graph (passenger counts, timestamp, casualty breakdown, emergency-service and media clusters); fed the fixed triplet list, the triplet render (b) is a uniform, skeletal layout. Bound to the supplied skeleton, the triplet render is also the more rigid of the two. Labels are translated into English for publication; the depicted connections are retained.

An input asymmetry. The extraction stage capped each article body at 3,000 characters while the direct baseline received the full text (Section 5). Only three events were affected at all, and the asymmetry runs the wrong way to explain the headline result: the direct image's faithfulness advantage is larger on the seven untruncated events (mean gap 1.29) than on the three truncated ones (0.33), and direct $\geq$ triplet holds in both subsets. The extra input therefore cannot account for the direct image's win; the asymmetry's plausible cost falls on the triplet pipeline's coverage, which we already report as a limitation.

What the explicit intermediate unlocks. Materializing the graph enables what the black box cannot: *selective rendering* (over-extract, then draw only a chosen subgraph rather than let the model silently omit); *cross-event bridging* via shared entities, surfacing links invisible across separate images; *sentiment linking*, joining nodes to the system's per-publisher sentiment; and *interactive images*, where OCR'd nodes let a reader click through to a node's cross-source coverage.

8 Conclusion, Limitations and Future Work

Making the hidden knowledge graph of a news infographic explicit does not, by itself, make the picture better – but it makes the pipeline honest. In our ten-event comparison the direct render was at least as faithful in every event (mean faithfulness 4.10 vs. 3.10) and substantially denser (18.4 vs. 8.1 nodes), yet the explicit stage turned a single unverifiable image into an auditable chain: extraction was roughly 70% clean with only ∼5% hard errors, while the generative renderer mis-drew ∼19% of the supplied edges, predominantly by attaching them to the wrong node – the pipeline's largest avoidable failure, and one the black box hides. We can see where the pipeline succeeds (extraction) and where it fails (generative rendering), and can substitute a faithful renderer where structure matters. For a Semantic-Web view of generative pipelines, that inspectability is the point.

These conclusions carry the limitations of a preliminary study. All ratings come from a single rater who is also an author, so the subjective scores are a self-assessment; with ten paired events and five ties we report sign consistency (10/10) rather than a significance test; and each condition was sampled once per event although image generation is stochastic, so per-event sampling effects cannot be ruled out. Only a handful of prompt iterations preceded freezing, so the results characterize this pipeline instance rather than an upper bound; the narrow node ontology explains part of the density and coverage gaps; and the dominant render error – edge mis-binding – is a documented weakness of current text-to-image models (Gokhale et al., 2022), whose mitigation through a constrained or programmatic layout is a prerequisite to scaling density.

An extended version will accordingly add independent second and third raters with Cohen's κ agreement on the categorical verdicts (the scoring tool already exports per-rater scores), multiple renders per condition with reported variance, and gold-standard triplets; further directions are RDF-star reification (Hartig et al., 2021) for attribution, controlled-vocabulary predicates with ontology alignment, e.g. Wikidata (Chepurova et al., 2025), and a richer deterministic renderer building on the `graphviz` proof-of-path included in the supplementary material.

Acknowledgments

This research was financially supported by the Slovenian Research and Innovation Agency (ARIS) via: the Research Program Knowledge Technologies (grant P2-

0103); the Gravity project AI for Science (grant GC-0001); and EC via the projects TWON (101095095) and DataPact (grant 101189771).

References

Antonucci, A., Piqué, G., & Zaffalon, M. (2024). Zero-shot causal graph extrapolation from text via LLMs [arXiv:2312.14670]. *XAI4Sci Workshop, AAAI Conference on Artificial Intelligence*. https://arxiv.org/abs/2312.14670

Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., & Etzioni, O. (2007). Open information extraction from the web. *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI)*, 2670–2676. https://dl.acm.org/doi/10.5555/1625275.1625705

Bian, H. (2025). LLM-empowered knowledge graph construction: A survey. *arXiv preprint arXiv:2510.20345*. https://arxiv.org/abs/2510.20345

Chepurova, A., Bulatov, A., Burtsev, M., & Kuratov, Y. (2025). Wikontic: Constructing wikidata-aligned, ontology-aware knowledge graphs with large language models. *arXiv preprint arXiv:2512.00590*. https://arxiv.org/abs/2512.00590

Darvariu, V.-A., Hailes, S., & Musolesi, M. (2024). Large language models are effective priors for causal graph discovery [arXiv:2405.13551]. https://arxiv.org/abs/2405.13551

Event Registry. (2025). *Event registry api documentation*. Retrieved September 7, 2025, from https://eventregistry.org/documentation

Fader, A., Soderland, S., & Etzioni, O. (2011). Identifying relations for open information extraction. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1535–1545. https://aclanthology.org/D11-1142

Gansner, E. R., & North, S. C. (2000). An open graph visualization system and its applications to software engineering. *Software: Practice and Experience*, *30*(11), 1203–1233.

Gokhale, T., Palangi, H., Nushi, B., Vineet, V., Horvitz, E., Kamar, E., Baral, C., & Yang, Y. (2022). Benchmarking spatial relationships in text-to-image generation [arXiv:2212.10015]. https://arxiv.org/abs/2212.10015

Ground News. (2025). *Ground news — breaking news headlines and media bias*. Retrieved September 7, 2025, from https://ground.news/

Hartig, O., Champin, P.-A., Kellogg, G., & Seaborne, A. (2021). *RDF-star and SPARQL-star (w3c community group final report)*. Retrieved June 1, 2026, from https://www.w3.org/2021/12/rdf-star.html

Huang, K., Duan, C., Sun, K., Xie, E., Li, Z., & Liu, X. (2025). T2I-CompBench++: An enhanced and comprehensive benchmark for compositional text-to-image generation [arXiv:2307.06350]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. https://doi.org/10.1109/TPAMI.2025.3531907

Huang, K., Sun, K., Xie, E., Li, Z., & Liu, X. (2023). T2I-CompBench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*. https://proceedings.neurips.cc/paper_files/paper/2023/hash/f8ad010cdd9143dbb0e9308c093aff24-Abstract-Datasets_and_Benchmarks.html

Hutto, C., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, 216–225. https://doi.org/10.1609/icwsm.v8i1.14550

Leban, G., Fortuna, B., Brank, J., & Grobelnik, M. (2014). Event registry: Learning about world events from news. *Proceedings of the 23rd International Conference on World Wide Web*, 107–110. https://doi.org/10.1145/2567948.2577024

Purchase, H. C. (2002). Metrics for graph drawing aesthetics. *Journal of Visual Languages & Computing*, *13*(5), 501–516. https://doi.org/10.1006/jvlc.2002.0232

Tang, Y., Liu, X., Zhang, B., Lan, T., Xie, Y., Lao, J., Wang, Y., Li, H., Gao, T., Pan, B., Weng, L., Huang, X., Zhu, M., Feng, Y., Luo, Y., & Chen, W. (2026). IGenBench: Benchmarking the reliability of text-to-infographic generation. *arXiv preprint arXiv:2601.04498*. https://arxiv.org/abs/2601.04498

Trajkov, G., Grobelnik, M., & Mladenić Grobelnik, A. (2025). BetweenTheLines - cross source news analysis. *Proceedings of the Slovenian KDD Conference 2025*. https://doi.org/10.70314/is.2025.sikdd.26

W3C. (2014). *RDF 1.1 concepts and abstract syntax*. World Wide Web Consortium. Retrieved June 1, 2026, from https://www.w3.org/TR/rdf11-concepts/

Wang, Z., Hu, X., Wang, Y., Xiong, F., Zhang, M., & Chu, X. (2026). Everything in its place: Benchmarking spatial intelligence of text-to-image models. *arXiv preprint arXiv:2601.20354*. https://arxiv.org/abs/2601.20354

Zala, A., Lin, H., Cho, J., & Bansal, M. (2024). DiagrammerGPT: Generating open-domain, open-platform diagrams via LLM planning [arXiv:2310.12128]. *Conference on Language Modeling (COLM)*. https://arxiv.org/abs/2310.12128

Evaluating AI-Generated Code: A Taxonomy and Comparative Catalog of Vibe-Coding Benchmarks

Marko Horvat, Sven Leo Kupres
University of Zagreb
Faculty of Electrical Engineering and Computing
Department of Applied Computing
Unska 3, HR-10000 Zagreb, Croatia
{Marko.Horvat3, Sven-Leo.Kupres}@fer.unizg.hr

Abstract. *As Large Language Models and AI coding agents are increasingly integrated into software development processes through the novel vibe-coding paradigm, selecting appropriate benchmarks is becoming essential for assessing their correctness, security, efficiency, and practical utility. This paper systematizes the large and expanding landscape of benchmarks suitable for vibe-coding by providing a structured inventory of benchmark families used to evaluate AI-generated code. It introduces a compact six-dimensional taxonomy that classifies benchmarks according to intended use, evaluation scope, programming context, task type, assessment method, and practical applicability. The paper also presents methodological recommendations for selecting the most appropriate benchmark according to the employed software architectures and programming technologies, emphasizing that reliable assessment of modern AI coding systems requires an aggregated score from multiple complementary benchmarks rather than a single evaluation framework.*

Keywords. Artificial intelligence; Large Language Models; Software engineering; Computer programming; Benchmark testing; AI-assisted programming; Vibe-coding

1 Introduction

Generative AI and Large Language Models (LLMs) are rapidly disrupting software engineering by enabling code generation, repository editing, automated repair, test generation, and interactive debugging through natural language conversational and agentic interfaces. In practitioner discourse, in a conversational language of software developers, this workflow is increasingly described as *vibe-coding*: a style of software development in which the human developer specifies intent at a relatively high level, and the model autonomously generates computer code, proposes architectural solutions, and executes code artifacts or tools supporting software development (Horvat, 2025; Meske et al., 2025; Sarkar & Drosos, 2025). Although the term is informal rather than academic, the underlying evaluation problem is scientifically important: *how should one measure the quality of the AI-generated code?* Is it correct, secure, maintainable, context-aware, and useful for real-world software engineering work?

The rapid adoption of new AI technologies by the developer community raises serious concerns about reliability, architectural soundness, code explainability, long-term maintainability and, ultimately, the security of generated computer code (Elgendy et al., 2026). The models' ability to comprehend specific technical systems must be tested and objectively evaluated before they can be used reliably. While it has been demonstrated that LLMs can generate functional boilerplate code (Sapkota et al., 2025), particularly in domains such as web front-end (Horvat et al., 2025), and even backend development (Horvat et al., 2026), their ability to consistently make optimal decisions in software development remains the subject of active academic and professional research.

Early evaluations of code models relied primarily on compact programming problems, where a model receives a docstring or short natural language description and generates a function that passes hidden tests (Austin et al., 2021; M. Chen et al., 2021). These benchmarks remain useful because they are accessible, deterministic, and simple to replicate. However, their scope is limited. They rarely assess project context, multi-file dependencies, build systems, database access, front-end rendering, security, deployment assumptions, or long-term tool utilization. Later benchmark families expanded the evaluation target from function-level synthesis to pragmatic and repository-level software work, including LiveCodeBench, BigCodeBench, DevEval, SWE-bench, frontend benchmarks, backend benchmarks, secure code-generation benchmarks, and human-calibrated agentic tasks (Jain et al., 2025; Jimenez et al., 2024; J. Li et al., 2024; Rein et al., 2025; Shen et al., 2026; Vero et al., 2025; H. Zhu et al., 2025; Zhuo et al., 2025).

This paper examines the evaluation of AI-generated code using a taxonomy and comparative catalog of vibe-coding benchmarks. The resulting catalog of vibe-coding benchmarks includes only benchmark families

for which a traceable public source can be found, such as scientific publication, official repository, benchmark website, or technical documentation. This criterion is important because the benchmark landscape is rapidly expanding and contains numerous emerging frameworks whose scope, public availability, versioning, and evaluation criteria are not necessarily consistent.

The paper makes three distinct contributions. First, it provides a systematic inventory and structured list of the main benchmark families currently used to evaluate AI-generated code. Second, it provides concise methodological recommendations for selecting benchmark portfolios in scientific and industrial evaluations. Third, it proposes a compact taxonomy for the structured classification of vibe-coding benchmarks along six orthogonal dimensions: intended use, evaluation scope, programming context, task type, assessment method, and practical applicability.

2 Vibe-coding and AI-assisted coding

Vibe-coding is a novel software development paradigm in which a developer specifies the desired software functionality through natural language prompts addressed to a LLM, rather than implementing the software artifact through precise line-by-line programming (Horvat, 2025; Meske et al., 2025; Sarkar & Drosos, 2025). In this process, the developer acts primarily as a high-level coordinator who guides the AI coding agent through an iterative dialogue of code generation, testing, correction, and refinement (Horvat, 2026). As a result, the key challenge is no longer limited to mastering programming language syntax, understanding software libraries, or low-level implementation details but is transformed to efficiently articulating intent, decomposing requirements, evaluating generated artifacts, and detecting functional, architectural, security, or maintainability defects in the produced code. Vibe coding should therefore be distinguished from conventional AI-assisted programming, where the developer remains the principal author of the code and uses AI tools only to augment the manual coding process. In vibe coding, by contrast, the generated code may partly function as a black-box artifact whose quality depends on the interaction between natural language specifications, model capabilities, evaluation feedback, and corrective prompting. This makes reliable benchmarking especially important, since the assessment of vibe-coded software must extend beyond simple functional correctness and include repository-level behavior, security, efficiency, maintainability, and practical applicability.

The main difference between AI-assisted programming, or AI-assisted coding, and vibe coding is in the developer's role and level of responsibility for the generated code. In AI-assisted programming, the developer remains the primary author of the software artifact while AI tools are used to accelerate or improve certain elements of the manual coding process, such as code completion, refactoring, debugging, or test generation.

The generated code still needs to be understood, reviewed, and integrated by the human developer. In contrast, in vibe coding, the developer only expresses high-level intent in natural language and guides the LLM or AI agent through an iterative dialogue, with the model generating the majority or all of the software implementation.

In summary, the vibe-coding paradigm allows the human developer to operate without comprehending or even reading the programming code, whereas in AI-assisted programming, the developer is entirely aware of the code and can manually address various issues. Both approaches, vibe-coding and AI-assisted coding, produce AI-generated code and can benefit from the catalog presented in the paper.

3 AI-generated Code Benchmarks

Benchmarks for assessment of AI-generated code differ from standard natural language benchmarks used for text comparison (for example, to detect AI-generated text) because their outputs are not merely texts assessed for plausibility. They are executable artifacts that may compile, fail, pass tests, expose vulnerabilities, render a user interface, corrupt data, or interact with a live environment. For example, instead of merely checking how closely AI's text matches a human solution, OpenAI's HumanEval and HumanEval+ verify functional correctness by automatically executing the generated code against hidden unit tests. Consequently, thorough coding evaluation normally requires test oracles (i.e., providers of information that describe correct output): unit tests, integration tests, build logs, browser assertions, database state checks, proof-of-concept exploit tests, or carefully calibrated human and machine judging. Reviews of code-generation evaluation consistently point out that lexical similarity metrics such as exact-match, Levenshtein distance, BLEU, or ROUGE are insufficient when functional equivalence and software quality matter (Garcia & Chakravarty, 2024; Lu et al., 2021).

Typically, AI-code generation benchmarks use pass@k metric (Brown et al., 2024). This estimates the probability that at least one of the k code attempts generated by an AI model passes all the unit tests. Some other common metrics used in this context are test pass rate and executability rate (Paul et al., 2024). Code-centric structural and semantic metrics are used as well (for example, CodeBLEU, CodeBERTScore and CodeScore). However, even if code passes a unit test, it might be poorly written or inefficiently. Therefore, performance constraint metrics are also used, such as runtime speed (latency), memory footprint and cyclomatic complexity (Paul et al., 2024).

For the purposes of this overview, a vibe-coding benchmark is defined as a benchmark whose primary target is production, repair, completion, translation, or execution-grounded assessment of code or code-centered software development workflows.

This includes single-function synthesis, competitive programming, data science coding, repository editing, front-end UI generation, backend/API implementation, SQL generation, program repair, secure code generation, terminal-based agent tasks, and human-calibrated software work. This definition excludes purely conceptual programming questions, general natural language reasoning tasks, and code-understanding tasks that do not require the model to produce an executable or directly software-relevant artifact.

The literature review also shows that AI-coding benchmarks are not uniformly available across the entire software development life cycle. Development-phase tasks are more strongly represented, whereas requirements engineering, architecture design, release engineering, maintenance economics, and long-term quality assurance remain less systematically evaluated (K. Wang et al., 2026). This imbalance explains why high scores on popular coding benchmarks, such as HumanEval and MBPP, should not be interpreted as evidence of end-to-end software engineering competence.

4 Methodology

The comparative catalog of vibe-coding benchmarks was constructed in five steps. First, a comprehensive literature search was conducted using Google Scholar search engine with the following two sets of keywords: (*AI-assisted programming, AI-assisted coding, vibe-coding*) and (*assessment, evaluation, benchmark, evaluation benchmarks*). Each keyword from the first set was combined with each keyword from the second set, resulting in 12 two-term search queries. These searches generated a total of 349,290 results. For each query, the first 5 pages of Google Scholar results were examined, with each page containing 10 literature items. Papers that were accessible and relevant to the topic were then manually downloaded for further analysis and selection. The initial search results included HumanEval, MBPP, SWE-bench, CoderEval, LiveCodeBench, APPS, TACO, Android and mobile-related benchmarks, frontend benchmarks, backend benchmarks, and SQL/database benchmarks. Tag cloud generated from the retrieved literature is shown in Fig. 1.

In the second step, the retrieved literature was analyzed to establish benchmark frameworks for various technology stacks, including Spring Boot, Node.js, TypeScript, and PostgreSQL. The third step involved mapping each candidate benchmark to a primary public artifact, such as an archival paper, preprint, official repository, benchmark website, or dataset release. Fourth, benchmark families were identified for closely related variants that have the same core task formulation but differ in language coverage, test type, modality,

Figure 1: Tag cloud of key concepts and terms related to vibe-coding benchmarks and the evaluation of AI-generated code from the retrieved literature.

or validation method. In the fifth and final step, each family was described using a common schema based on their intended use, task type, input modality and granularity, programming language or domain, evaluation oracle and metric, availability and miscellaneous constraints, if they exist.

5 Taxonomy of Vibe-Coding Benchmarks

We propose a taxonomy in which the AI-assisted coding and vibe-coding benchmark landscape can be organized along six orthogonal dimensions. The first dimension in this taxonomy or classification is *task formulation* and includes the following software attributes: generation, completion, translation, repair, refinement, security hardening, test generation, or agentic workflow execution. The second dimension is *input modality* and applies to: natural language prompts, docstrings, partial code, issue reports, repository context, screenshots, diagrams, flowcharts, tests-as-specification, database schemas, or terminal states. The third dimension is *granularity* and covers: function, method, class, file, repository, application, or full workflow. The fourth dimension is *software domain*: algorithmic programming, data science, frontend, backend, mobile, SQL/database, spreadsheets, high-performance computing, security, or research replication. The fifth dimension is *oracle*: unit tests, compilation, integration tests, end-to-end browser execution, database checks, exploit validation, visual similarity, or human/LLM judging. The final sixth dimension is *evaluation metric* and is represented by AI-software evaluation metrics such as pass@k, execution accuracy, exact match, CodeBLEU, secure-pass variants, visual similarity, task success rate, human-time calibration, or economic value.

This classification was used as the methodological

structure for assembling the catalog of vibe-coding benchmarks provided below. Each benchmark identified in the literature review was analyzed and characterized according to the proposed dimensions, providing a common structure for describing otherwise heterogeneous benchmark families. The proposed taxonomy, which classifies vibe-coding benchmarks according to six orthogonal dimensions, is shown in Fig. 2.

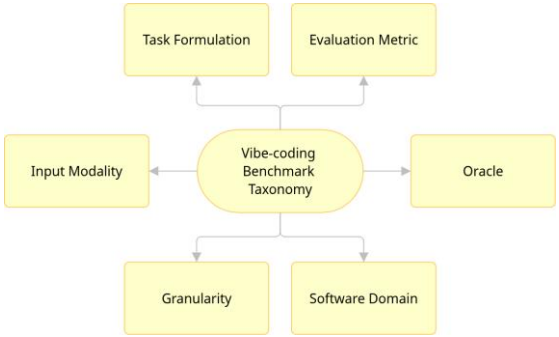


Figure 2: The proposed taxonomy of vibe-coding benchmarks defined by six orthogonal dimensions used in the construction of the benchmark catalog.

These dimensions reflect an important methodological trade-off. Function-level benchmarks may be simple to reproduce, scale, and execute repeatedly, but they do not completely reveal how software is developed, tested, and used in real-world settings. In contrast, repository-level, UI, mobile, backend, security, and terminal-based benchmarks more closely approximate realistic software engineering work, but they are also more expensive to execute, more sensitive to infrastructure differences, and more difficult to com-pare across evaluation environments. For this reason, comprehensive and reliable assessment of AI-generated code should rely on a portfolio of complementary benchmark families rather than a single benchmark score.

6 Comprehensive Benchmark Catalog

The complete catalog of AI-assisted programming and vibe-coding benchmarks is presented in three separate tables:

1. Table 1: Canonical code synthesis, competitive programming, multilingual, and data science benchmarks.
2. Table 2: Repository, UI, mobile, backend, database, and specialized-code benchmarks.
3. Table 3: Repair, agentic workflows, security, and human calibration benchmarks.

Table 1, Table 2, and Table 3 form one logical catalog split across three consecutive pages at the end of the paper to improve readability. Fig. 3 shows the timeline of initial release dates for selected representative benchmarks for AI-assisted coding listed in the catalog. The publication of benchmarks for assessing AI-generated program code has been steadily increasing following the introduction of LLMs, and this trend has been particularly evident since mid-2023. It should be noted that some benchmarks were developed long before the vibe coding paradigm has been introduced. Nevertheless, they remain relevant for evaluating code generated with this paradigm.

7 Recommendations for Benchmark Use

Based on the survey results, a comprehensive evaluation of AI-generated code should follow these five concise recommendations:

1. Use a set of benchmarks that combines at least one canonical synthesis benchmark, one contamination-aware or recent benchmark, and one domain-specific benchmark aligned with the intended software task.
2. When possible, use execution-grounded oracles. Additionally, report compilation, unit tests, integration tests, browser tests, database state checks, or exploit validation instead of relying on text similarity.
3. Report the harness (i.e., infrastructure that surrounds a software component to execute it) precisely. Repository-level and agentic benchmarks usually require specific versions of datasets, Docker images, library dependencies, different parameters, and tool access to be used successfully.
4. Separate correctness from quality. Functional success, maintainability, architectural soundness, readability, performance, and security are related but distinct dimensions.
5. Account for contamination and benchmark diffusion. Prefer fresh, time-split, private-test, or dynamically updated benchmarks when evaluating frontier models that may have seen public tasks during training or evaluation-time optimization.

For Spring Boot/PostgreSQL or Node.js/TypeScript/PostgreSQL backend evaluation, the implication is direct. Existing public benchmarks provide useful components: JavaBench for Java/OOP structure, Spider-style tasks for database query competence, BaxBench and ABC-Bench for backend testing, and SWE-bench-style tasks for repository repair. Additionally, a stack-specific evaluation should include a custom integration harness. Such a harness should compile generated code, start an ephemeral database, run framework-specific tests, validate serialization boundaries, and check software behavior relevant for security.

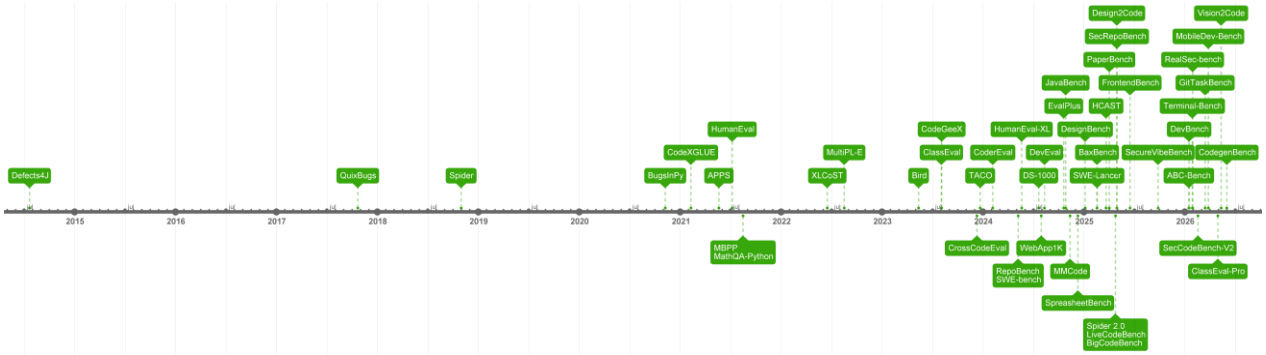


Figure 3: Timeline of first release dates for selected representative benchmarks for AI-assisted coding from the comparative catalog listed in Table 1, Table 2, and Table 3. The timeline clearly shows how the adoption of LLMs in code generation substantially accelerated the development of new benchmarks.

8 Discussion

The catalog shows a clear transition from narrow function synthesis to broader software engineering evaluation. Three observations are especially important for scientific benchmarking.

The most important observation is that *functional correctness is necessary but insufficient for a complete evaluation of AI-generated program code*. HumanEval, MBPP, EvalPlus, APPS, LiveCodeBench, BigCodeBench, and many repair benchmarks rely on unit tests as the primary oracle. This is methodologically stronger than string/text similarity because code may be lexically and syntactically different yet behaviorally equivalent. However, a test suite is only as strong as its coverage. EvalPlus was developed precisely because original HumanEval and MBPP tests could miss edge cases (J. Liu et al., 2023). In repository and back-end settings, test adequacy becomes even more difficult because models may pass local tests while violating different architectural, serialization, security, or performance requirements.

The second observation is that *context has become the most important factor*. Repository-level benchmarks show that coding ability depends on non-local information such as imports, APIs, naming conventions, build systems, framework idioms, and domain models. Consequently, CoderEval, RepoBench, CrossCodeEval, DevEval, JavaBench, and SWE-bench, for instance, are more congruent with professional development than just function synthesis (Cao et al., 2024; Ding et al., 2023; Jimenez et al., 2024; J. Li et al., 2024; T. Liu et al., 2024; H. Yu et al., 2024). This shift also introduces new methodological challenges. A model may fail due to a lack of reasoning ability, retrieval selecting incorrect files, dependencies not being installed, or the test harness differing from the original repository environment.

The third observation shows that *risk-sensitive evaluation is no longer an option*. Benchmarks such as BaxBench, SecRepoBench, RealSec-bench, SecureAgentBench, and SecCodeBench-V2 show that code can be correct under ordinary tests and still be insecure (J. Chen et al., 2025; L. Chen et al., 2026; Shen et al., 2026; Vero et al., 2025; Y. Wang et al., 2026). For backend, API, database, and mobile applications, a benchmark that ignores security can overestimate practical deployability. Software attributes such as security, privacy, authorization, injection resistance, dependencies, and usability should all be treated as equally important in applied environments as features evaluated by functional benchmarks.

9 Conclusion

Benchmarks for evaluating AI-generated code have evolved considerably beyond the first generation of short Python functions and have become highly complex. Given the speed with which large volumes of program code can be automatically generated using AI, benchmarks have become a necessity in the evaluation of vibe-coded software. Choosing the optimal benchmark has become very important but, at the same time, a difficult task.

HumanEval and MBPP benchmark families still remain valuable baselines, but modern vibe-coding evaluation must also address many additional modalities, such as repository context, multi-file dependencies, UI rendering, backend and database behavior, security, mobile ecosystems, terminal workflows, and human-calibrated software work. Therefore, *the correct methodological approach in the evaluation of AI-generated software is not a single benchmark score but an aggregated score based on a set of selected benchmarks aligned with the software features being tested*. To improve dissemination transparency and quality, such a joint score can be represented as a vector, with each individual score displayed separately, or, in a simpler way, as a scalar (i.e., a single real number) derived as a weighted sum of individual scores.

The most robust evaluation strategy should combine selected benchmarks with functional testing, testing strategies resistant to contamination, repository scaling tasks, and domain test oracles. For applied software engineering, additional dimensions such as security,

maintainability, performance, and integration correctness should be measured. The field of AI-generated code evaluation is converging towards more realistic and risk-sensitive assessment, but reproducibility, infrastructure cost, programming language imbalance, and benchmark contamination remain open challenges.

Acknowledgments

This paper is partly based on the research developed through the project "Maritime mAInd – Advanced platform for comprehensive design of marine vessels supported by machine learning algorithms" supported by the European Union within the Operational Programme Competitiveness and Cohesion in the financial period 2021–2027 (PK.1.1.12.0334).

References

Austin, J., Odena, A., Nye, M., Bosma, M., Michalewski, H., Dohan, D., Jiang, E., Cai, C., Terry, M., Le, Q., et al. (2021). Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*.

Brown, B., Juravsky, J., Ehrlich, R., Clark, R., Le, Q. V., Ré, C., & Mirhoseini, A. (2024). Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*.

Cao, J., Chen, Z., Wu, J., Cheung, S.-C., & Xu, C. (2024). Javabench: A benchmark of object-oriented code generation for evaluating large language models. *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering*, 870–882.

Cassano, F., Gouwar, J., Nguyen, D., Nguyen, S., Phipps-Costin, L., Pinckney, D., Yee, M.-H., Zi, Y., Anderson, C. J., Feldman, M. Q., et al. (2022). Multipl-e: A scalable and extensible approach to benchmarking neural code generation. *arXiv preprint arXiv:2208.08227*.

Chen, J., Huang, H., Lyu, Y., An, J., Shi, J., Yang, C., Zhang, T., Tian, H., Li, Y., Li, Z., et al. (2025). Secureagentbench: Benchmarking secure code generation under realistic vulnerability scenarios. *arXiv preprint arXiv:2509.22097*.

Chen, L., Zhao, J., Cui, L., Su, T., Pan, X., Li, Z., Wu, Y., Cao, Q., Cai, Q., Zhang, J., et al. (2026). *Seccodebench-v2 technical report* (tech. rep.). Chen, M., Tworek, J., Jun, H., Yuan, Q., de Oliveira Pinto, H. P., Kaplan, J., Edwards, H., Burda,

Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., … Zaremba, W. (2021). Evaluating large language models trained on code. https://arxiv.org/abs/2107. 03374

Chen, Y., Xie, C., Shi, Y., Zeng, W., Wang, Y., Zhang, H., & Gu, X. (2026). Classeval-pro: A cross-domain benchmark for class-level code generation. https://arxiv.org/abs/2604.26923

Cui, Y. (2024). Webapp1k: A practical code-generation benchmark for web app development. https://arxiv.org/abs/2408.00019

Ding, Y., Wang, Z., Ahmad, W., Ding, H., Tan, M., Jain, N., Ramanathan, M. K., Nallapati, R., Bhatia, P., Roth, D., et al. (2023). Crosscodeeval: A diverse and multilingual benchmark for cross-file code completion. *Advances in Neural Information Processing Systems*, *36*, 46701–46723.

Elgendy, I. A., Dwivedi, Y. K., Al-Sharafi, M. A., Hosny, M., Helal, M. Y. I., Crick, T., Hughes, L., Alwahaishi, S., Mahmud, M., Dutot, V., & Al-Busaidi, A. S. (2026). Responsible vibe coding: Architecture, opportunities, and re-search agenda. *Journal of Computer Information Systems*, *0*(0), 1–19. https://doi.org/10. 1080/08874417.2026.2621186

Fakorede, M. A., Upadhyay, K., Siddique, A. B., & Farooq, U. (2026). Mobiledev-bench: A comprehensive benchmark for evaluating language models on mobile application development. https://arxiv.org/abs/2603.24946v1

FudanSELab. (2023). ClassEval: A manually-crafted benchmark for evaluating LLMs on class-level code generation. https://github.com/ FudanSELab/ClassEval

Garcia, M., & Chakravarty, M. M. T. (2024). A re-view of evaluation methods for code generation. https://arxiv.org/abs/2408.16498

Gottumukkala, A. (2025). Kotlin-bench. https://github. com/firebenders/Kotlin-bench

Hendrycks, D., Basart, S., Kadavath, S., Mazeika, M., Arora, A., Guo, E., Burns, C., Puranik, S., He, H., Song, D., et al. (2021). Measuring coding challenge competence with apps. *arXiv preprint arXiv:2105.09938*.

Horvat, M. (2025). What is vibe coding and when should you use it (or not)? *Authorea Preprints.*

Horvat, M. (2026). Which prompting technique is better? Improving vibe coding code quality with efficient prompt design for web front-end development. *interaction*, *9*, 10.

Horvat, M., Kralj, B., & Gledec, G. (2025). A comparative study of vibe coding with chatgpt and gemini in front-end web development. *Proceedings of the 36th International Scientific Conference: Central European Conference on Information and Intelligent Systems: CECIIS 2025*, 787–796.

Horvat, M., Ursić, I., & Krmpotić, K. (2026). Comparative evaluation of gemini and deepseek for llm-generated code quality and architectural robustness in backend software engineering. *Electronics*, *15*(13). https://doi.org/10.3390/ electronics15132805

Jain, N., Gu, A., Li, W.-D., Yan, F., Zhang, T., Wang, S., Solar-Lezama, A., Sen, K., & Stoica, I. (2025). Livecodebench: Holistic and contamination free evaluation of large language models for code. *International Conference on Learning Representations, 2025*, 58791–58831.

Jimenez, C. E., Yang, J., Wettig, A., Yao, S., Pei, K., Press,

O., & Narasimhan, K. (2024). Swe-bench: Can language models resolve real-world github issues? *International Conference on Learning Representations, 2024*, 54107–54157.

Just, R., Jalali, D., & Ernst, M. D. (2014). Defects4j: A database of existing faults to enable controlled testing studies for java programs. *Proceedings of the 2014 international symposium on software testing and analysis*, 437–440.

Kumarappan, A., Golnari, P. A., Wen, W., Liu, X., Ryan, G., Sun, Y., Fu, S., & Nallipogu, E. (2026). De-vbench: A realistic, developer-informed benchmark for code generation models. https://arxiv. org/abs/2601.11895

Lai, Y., Li, C., Wang, Y., Zhang, T., Zhong, R., Zettlemoyer, L., Yih, S. W.-t., Fried, D., Wang, S., & Yu, T. (2022). Ds-1000: A natural and reliable benchmark for data science code generation. https://arxiv.org/abs/2211.11501

Lei, F., Chen, J., Ye, Y., Cao, R., Shin, D., Su, H., Suo, Z., Gao, H., Hu, W., Yin, P., et al. (2025). Spider 2.0: Evaluating language models on real-world enterprise text-to-sql workflows. *International Conference on Learning Representations, 2025*, 28691–28735.

Li, J., Li, G., Zhao, Y., Li, Y., Liu, H., Zhu, H., Wang, L., Liu, K., Fang, Z., Wang, L., et al. (2024). Deveval: A manually-annotated code generation benchmark aligned with real-world code repositories. *Findings of the Association for Computational Linguistics: ACL 2024*, 3603–3614.

Li, J., Wu, W., Hu, J., Zheng, Q., Wu, B., Zheng, J., Lu, Y., & Fu, H. (2026). Codegenbench: Can llms write efficient code across architectures? https://arxiv.org/abs/2606.04023

Li, J., Hui, B., Qu, G., Yang, J., Li, B., Li, B., Wang, B., Qin, B., Cao, R., Geng, R., Huo, N., Zhou, X., Ma, C., Li, G., Chang, K. C. C., Huang, F., Cheng, R., & Li, Y. (2023). Can llm already serve as a database interface? a big bench for large-scale database grounded text-to-sqls. https://arxiv.org/abs/2305.03111

Li, K., Tian, Y., Hu, Q., Luo, Z., Huang, Z., & Ma, J. (2024). Mmcode: Benchmarking multimodal large language models for code generation with visually rich programming problems. *Findings of the Association for Computational Linguistics: EMNLP 2024*, 736–783.

Li, R., Fu, J., Zhang, B.-W., Huang, T., Sun, Z., Lyu, C., Liu, G., Jin, Z., & Li, G. (2023). Taco: Topics in algorithmic code generation dataset. *arXiv preprint arXiv:2312.14852*.

Lin, D., Koppel, J., Chen, A., & Solar-Lezama, A. (2017). Quixbugs: A multi-lingual program repair benchmark set based on the quixey challenge. *Proceedings Companion of the 2017 ACM SIGPLAN international conference on systems, programming, languages, and applications: software for humanity*, 55–56.

Liu, J., et al. (2023). EvalPlus: A rigorous evaluation framework for LLM4Code. https://github.com/evalplus/evalplus

Liu, T., Xu, C., & McAuley, J. (2024). Repobench: Benchmarking repository-level code auto-completion systems. *International Conference on Learning Representations, 2024*, 47832–47850.

Lopez-Latorre, Z. (2026). Android bench. https://github. com/android-bench/android-bench

Lu, S., Guo, D., Ren, S., Huang, J., Svyatkovskiy, A., Blanco, A., Clement, C., Drain, D., Jiang, D., Tang, D., et al. (2021). Codexglue: A machine learning benchmark dataset for code understanding and generation. *arXiv preprint arXiv:2102.04664*.

Ma, Z., Zhang, B., Zhang, J., Yu, J., Zhang, X., Zhang, X., Luo, S., Wang, X., & Tang, J. (2024). Spreadsheetbench: Towards challenging real world spreadsheet manipulation. *Advances in Neural Information Processing Systems, 37*, 94871–94908.

Maninger, D., Chemnitz, L., Sharifloo, A. M., Brugger, J., & Mezini, M. (2025). Benchmarking web api integration code generation. *2025 2nd IEEE/ACM International Conference on AI-powered Software (AIware)*, 240–248.

Merrill, M. A., Shaw, A. G., Carlini, N., Li, B., Raj, H., Bercovich, I., Shi, L., Shin, J. Y., Walshe, T., Buchanan, E. K., Shen, J., Ye, G., Lin, H., Poulos, J., Wang, M., Nezhurina, M., Jitsev, J., Lu, D., Mastromichalakis, O. M., . . . Schmidt, L. (2026). Terminal-bench: Benchmarking agents on hard, realistic tasks in command line interfaces. https://arxiv.org/abs/2601.11868

Meske, C., Hermanns, T., Von der Weiden, E., Loser, K.U., & Berger, T. (2025). Vibe coding as a reconfiguration of intent mediation in software development: Definition, implications, and research agenda. *IEEE Access, 13*, 213242–213259.

Miserendino, S., Wang, M., Patwardhan, T., & Hei-decke, J. (2025). Swe-lancer: Can frontier llms earn $1 million from real-world freelance software engineering? https://arxiv.org/abs/2502.12115

Nam, J., et al. (2025). Autobe - ai backend builder for prototype to production. https://github.com/ wrtnlabs/autobe

Ni, Z., Wang, H., Zhang, S., Lu, S., He, Z., Tang, Z., Hu, S., Li, B., Hu, C., Jiao, B., et al. (2026). Git-taskbench: A benchmark for code agents solving real-world tasks through code repository leveraging. *Proceedings of the AAAI Conference on Artificial Intelligence, 40*(38), 32564–32572.

Paul, D. G., Zhu, H., & Bayley, I. (2024). Benchmarks and metrics for evaluations of code generation: A critical review. *2024 IEEE International Conference on Artificial Intelligence Testing (AITest)*, 87–94.

Peng, Q., Chai, Y., & Li, X. (2024). Humaneval-xl: A multilingual code generation benchmark for cross-lingual natural language generalization. *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024)*, 8383–8394.

Periasami, A. V., Wang, J., & Dhingra, B. (2026). Vision2code: A multi-domain benchmark for

evaluating image-to-code generation. https://arxiv.org/abs/2605.11307

Princeton NLP. (2024). SWE-bench official repository. https://github.com/princeton-nlp/SWE-bench

Rein, D., Becker, J., Deng, A., Nix, S., Canal, C., O'Connel, D., Arnott, P., Bloom, R., Broadley, T., Garcia, K., et al. (2025). Hcast: Human-calibrated autonomy software tasks. *arXiv preprint arXiv:2503.17354*.

Sapkota, R., Roumeliotis, K. I., & Karkee, M. (2025). Vibe coding vs. agentic coding: Fundamentals and practical implications of agentic ai. *arXiv preprint arXiv:2505.19443*.

Sarkar, A., & Drosos, I. (2025). Vibe coding: Programming through conversation with artificial intelligence. *arXiv preprint arXiv:2506.23253*.

Shen, C., Dilgren, C., Chiniya, P., Griffith, L., Ding, Y., & Chen, Y. (2026). Secrepobench: Benchmarking code agents for secure code completion in real-world repositories. https://arxiv.org/abs/ 2504.21205

Si, C., Zhang, Y., Li, R., Yang, Z., Liu, R., & Yang, D. (2025). Design2code: Benchmarking multimodal code generation for automated front-end engineering. *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 3956–3974.

Starace, G., Jaffe, O., Sherburn, D., Aung, J., Chan, J. S., Maksin, L., Dias, R., Mays, E., Kinsella, B., Thompson, W., et al. (2025). Paperbench: Evaluating ai's ability to replicate ai research. *arXiv preprint arXiv:2504.01848*.

Vero, M., Mündler, N., Chibotaru, V., Raychev, V., Baader, M., Jovanović, N., He, J., & Vechev, M. (2025). Baxbench: Can llms generate correct and secure backends? *arXiv preprint arXiv:2502.11844*.

Wang, K., Li, T., Zhang, X., Wang, C., Sun, W., Liu, Y., Liu, A., Liu, X., Shen, C., & Shi, B. (2026). Software development life cycle perspective: A survey of benchmarks for code large language models and agents. https://arxiv.org/abs/2505.05283

Wang, Y., Zhang, Z., Wang, C., Xu, X., Liu, M., Wang, Y., Chen, J., & Zheng, Z. (2026). Realsec-bench: A benchmark for evaluating secure code generation in real-world repositories. *arXiv preprint arXiv:2601.22706*.

Widyasari, R., Sim, S. Q., Lok, C., Qi, H., Phan, J., Tay, Q., Tan, C., Wee, F., Tan, J. E., Yieh, Y., et al. (2020). Bugsinpy: A database of existing bugs in python programs to enable controlled testing and debugging studies. *Proceedings of the 28th ACM joint meeting on european software engineering conference and symposium on the foundations of software engineering*, 1556–1560.

Xiao, J., Wang, M., Lam, M. H., Wan, Y., Liu, J., Huo, Y., & Lyu, M. R. (2025). Designbench: A comprehensive benchmark for mllm-based front-end code generation. *arXiv preprint arXiv:2506.06251*.

Yang, J., Guo, H., Ji, L., Zhou, J., Zheng, R., Lei, Z., Zhang, S., Xi, Z., Liu, S., Wang, Y., Wang, B., Zheng, Y., Gui, T., & Qiu, X. (2026). Abcbench: Benchmarking agentic backend coding in real-world development. https://arxiv.org/ abs/2601.11077

Yu, H., Shen, B., Ran, D., Zhang, J., Zhang, Q., Ma, Y., Liang, G., Li, Y., Wang, Q., & Xie, T. (2024). Codereval: A benchmark of pragmatic code generation with generative pre-trained models. *Proceedings of the 46th IEEE/ACM International Conference on Software Engineering*, 1–12.

Yu, T., Zhang, R., Yang, K., Yasunaga, M., Wang, D., Li, Z., Ma, J., Li, I., Yao, Q., Roman, S., et al. (2018). Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task. *Proceedings of the 2018 conference on empirical methods in natural language processing*, 3911–3921.

Zhang, F., Wu, L., Bai, H., Lin, G., Li, X., Yu, X., Wang, Y., Chen, B., & Keung, J. (2025). Humaneval-v: Benchmarking high-level visual reasoning with complex diagrams in coding tasks. https://arxiv.org/abs/2410.12381

Zheng, Q., Xia, X., Zou, X., Dong, Y., Wang, S., Xue, Y., Shen, L., Wang, Z., Wang, A., Li, Y., et al. (2023). Codegeex: A pre-trained model for code generation with multilingual benchmarking on humaneval-x, 5673–5684.

Zhu, H., Zhang, Y., Zhao, B., Ding, J., Liu, S., Liu, T., Wang, D., Liu, Y., & Li, Z. (2025). Frontend-bench: A benchmark for evaluating llms on front-end development via automatic evaluation. *arXiv preprint arXiv:2506.13832*.

Zhu, M., Jain, A., Suresh, K., Ravindran, R., Tipirneni, S., & Reddy, C. K. (2022). Xlcost: A benchmark dataset for cross-lingual code intelligence. https://arxiv.org/abs/2206.08474

Zhuo, T. Y., Vu, M. C., Chim, J., Hu, H., Yu, W., Widyasari, R., Yusuf, I. N. B., Zhan, H., He, J., Paul, I., et al. (2025). Bigcodebench: Benchmarking code generation with diverse function calls and complex instructions. *International Conference on Learning Representations, 2025*, 66602–66656.

Table 1: Comprehensive catalog of AI-generated code benchmarks, part I: canonical code synthesis, competitive programming, multilingual, and data-science benchmarks.

Benchmark family	Intended use	Input and granularity	Domain/languages	Oracle and metrics	Key properties and limitations
HumanEval family: HumanEval, HumanEval+, HumanEval-X, HumanEval-XL, HumanEval-V	Canonical code synthesis; strengthened, multilingual, and visual variants	Docstring or natural-language prompt to function; HumanEval-V adds visual input	Python baseline; variants cover additional programming and natural languages	Hidden unit tests; pass@k; stronger tests in EvalPlus	Reproducible and historically central, but small, Python-centered in the original form, and vulnerable to contamination; variants improve test strength, language coverage, and visual reasoning (M. Chen et al., 2021; J. Liu et al., 2023; Peng et al., 2024; Zhang et al., 2025; Zheng et al., 2023).
MBPP family: MBPP, MBPP+	Entry-level standalone program synthesis	Short natural-language problem descriptions to small programs	Python, beginner-to-intermediate algorithmic tasks	Unit tests; pass@k; strengthened tests in MBPP+	Useful for quick baseline evaluation and novice-style tasks; less realistic than project-level benchmarks and limited in language/ecosystem coverage (Austin et al., 2021; J. Liu et al., 2023).
AixBench	Method-level code generation with automatic and manual assessment	Natural-language or specification-like prompt to method implementation	General programming tasks	Automatic correctness metric and manual evaluation	Bridges pure unit-test evaluation and qualitative assessment; smaller and less widely adopted than HumanEval or MBPP.
APPS, TACO, LiveCodeBench, CodeElo	Competitive-programming and algorithmic reasoning evaluation	Long problem statement to complete solution; some variants include repair or execution prediction	Algorithmic programming, mainly contest-style settings	Hidden tests, contest submission, pass@k, Elo-style scoring	More difficult and contamination-aware than static function suites, especially LiveCodeBench; still measures contest programming more than industrial software engineering (Hendrycks et al., 2021; Jain et al., 2025; R. Li et al., 2023).
BigCodeBench, BigCodeBench-Hard, CodeMixBench	Practical function generation requiring library and API use	Instruction or docstring to function using many libraries and complex constraints	Python-centered practical coding across library domains	Unit tests with stronger branch coverage; pass@k	Better reflects tool/library use than HumanEval; still mostly function-level and sensitive to execution environment availability (Zhuo et al., 2025).
DS-1000, DSBench	Data-science code generation and agentic data-analysis tasks	StackOverflow-style or data-analysis prompt plus data context	Python data science, scientific and analytics libraries	Unit tests, output comparison, analy-sis/modeling task completion	Strong domain realism for data science; narrower transferability to general software engineering (Lai et al., 2022).
CodeXGLUE, XLCoST	Broad code-intelligence suites, including generation, translation, summarization, search, and refinement	Natural-language-to-code, code-to-code, code-to-text, and related formats	Multiple languages and code-intelligence tasks	Mixed metrics: BLEU, CodeBLEU, exact match, retrieval and task-specific scores	Important for broad benchmarking but not all subtasks are execution-grounded; scores should not be equated with functional correctness (Lu et al., 2021; M. Zhu et al., 2022).
AVATAR, MultiPL-E, McEval, MCoNaLa	Code translation and multilingual natural-language-to-code evaluation	Parallel code, translated prompts, multilingual natural-language inputs	Java/Python and many additional programming/natural languages	Unit tests where available; translation and code-similarity metrics in some settings	Addresses English/Python monoculture; automatic translation can introduce artifacts and may not capture idiomatic language-specific programming (Cassano et al., 2022).

Table 2: Comprehensive catalog of AI-generated code benchmarks, part II: repository, UI, mobile, backend, database, and specialized-code benchmarks.

Benchmark family	Intended use	Input and granularity	Domain/languages	Oracle and metrics	Key properties and limitations
MMCode, Code-Vision, Vision2Code	Multimodal code generation from visual problem statements, flowcharts, diagrams, or other visual sources	Image plus optional text to executable code or rendered artifact	Multimodal programming and visual reasoning	Unit tests, visual similarity, VLM or human judging depending on task	Captures visual specification ability absent from text-only benchmarks; evaluation may depend on rendering and judge quality (K. Li et al., 2024; Periasami et al., 2026).
CoderEval, RepoBench, CrossCodeEval, DevEval, DevBench	Pragmatic generation and completion using project context, repository dependencies, or telemetry-derived developer tasks	Partial code, repository context, local APIs, or development traces; method/file/repo level	Python, Java, TypeScript, C#, and mixed repositories	Unit tests, exact match, retrieval utility, task success, and judge-assisted usefulness	Measures non-local context use and dependency awareness; results can be confounded by retrieval quality, environment setup, and repository availability (Ding et al., 2023; Kumarappan et al., 2026; J. Li et al., 2024; T. Liu et al., 2024; H. Yu et al., 2024).
ClassEval, ClassEval-Pro, JavaBench	Class-level and project-level structured code generation	Class skeletons, method descriptions, signatures, and project context	Python classes; Java OOP/project settings	Unit tests, class-wise and method-wise pass@k	Bridges single functions and full repositories; useful for OOP and enterprise-like structure, but less comprehensive than repository repair benchmarks (Cao et al., 2024; Y. Chen et al., 2026; FudanSELab, 2023).
Design2Code, DesignBench, FrontendBench, WebSight/WebSight-Test	Frontend, UI, screenshot-to-code, and interactive web-component evaluation	Screenshot, design image, prompt, or component specification to HTML/CSS/JS/framework code	Web UI; React, Vue, Angular, vanilla HTML/CSS in some variants	Rendering similarity, browser tests, end-to-end assertions, human/VLM agreement	Essential for UI coding; deterministic testing is harder because visual fidelity, layout similarity, and interactivity must be evaluated jointly (Si et al., 2025; Xiao et al., 2025; H. Zhu et al., 2025).
WebApp1K, BaxBench, ABC-Bench, Web API Integration benchmark	Web application, backend, and API-integration code generation	Application requirements, tests-as-specification, API descriptions, or backend tasks to multi-file code	Web applications, backend services, API clients	Unit/integration tests, API checks, exploit/security validation in BaxBench	Stronger ecological validity for production-style development; setup is framework-sensitive and more expensive than function-level evaluation (Cui, 2024; Maninger et al., 2025; Vero et al., 2025; Yang et al., 2026).
Spider, BIRD, Spider 2.0	Text-to-SQL and database-grounded code generation	Natural-language question plus schema to SQL or SQL-agent workflow	Relational databases and enterprise-like data access	Execution accuracy, exact match, SQL-specific metrics	Crucial for database code, but SQL-specific results should not be generalized uncritically to full backend software engineering (Lei et al., 2025; J. Li et al., 2023; T. Yu et al., 2018).
SpreadsheetBench, CodegenBench	Spreadsheet automation and performance-oriented code generation	Natural-language spreadsheet task or kernel-generation specification	Spreadsheets; BLAS/HPC and architecture-specific code	File-level tests, execution correctness, performance metrics	Expands evaluation beyond general-purpose scripts; highly domain-specific and infrastructure-dependent (J. Li et al., 2026; Ma et al., 2024).
Android Bench, Kotlin-Bench, AutoBe	Mobile and backend evaluation	Mobile/backend-oriented benchmarks	Android, Kotlin, backend development		(Gottumukkala, 2025; Lopez-Latorre, 2026; Nam et al., 2025)

Table 3: Comprehensive catalog of AI-generated code benchmarks, part III: repair, agentic workflows, security, and human calibration.

Benchmark family	Intended use	Input and granularity	Domain/languages	Oracle and metrics	Key properties and limitations
Defects4J, Defects4J-NL2Fix, BugsInPy, QuixBugs	Automated program repair and bug-fix generation	Buggy code, tests, and sometimes natural-language issue text to patch	Java, Python, and small algorithmic programs	Unit tests, patch correctness, top-*k* repair accuracy	Foundational for repair; test-suite overfitting and incomplete specifications remain classic threats (Just et al., 2014; Lin et al., 2017; Widyasari et al., 2020).
SWE-bench family: SWE-bench, Lite, Verified, Multimodal, Multi-SWE-bench, SWE-Bench Pro, MobileDev-Bench	Real-world issue resolution over repositories; increasingly multimodal, multilingual, long-horizon, and mobile variants	Issue report plus repository, build/test environment, sometimes screenshots or mobile app context	Python core benchmark; variants cover JavaScript, multiple languages, and mobile ecosystems	Build-and-test harnesses, issue-resolution rate, pass@1, patch success	Reference family for coding agents; high ecological validity but expensive, environment-sensitive, and increasingly exposed to benchmark diffusion (Fakorede et al., 2026; Jimenez et al., 2024; Princeton NLP, 2024).
GitTaskBench, Terminal-Bench, GitGoodBench	End-to-end repository and terminal workflows for code agents	Natural-language task plus repository or terminal environment	GitHub tasks, CLI tasks, software maintenance workflows	Task success, com-mand/environment checks, cost-aware or alpha-style metrics	Strong measure of tool-using autonomy; difficult to reproduce and compare because task execution depends on environment state (Merrill et al., 2026; Ni et al., 2026).
SecRepoBench, SecureAgentBench, RealSec-bench, SecCodeBench-V2	Correct-and-secure code generation, secure repair, and vulnerability-aware repository tasks	Security-focused prompts, vulnerable repositories, project scaffolds, or exploit scenarios	C/C++, Java, and multi-language industrial settings	Unit tests plus static analysis, exploit proof-of-concept, secure-pass metrics	Shows that functional correctness and security diverge; security must be evaluated explicitly rather than inferred from ordinary tests (J. Chen et al., 2025; L. Chen et al., 2026; Shen et al., 2026; Y. Wang et al., 2026).
SWE-Lancer, HCAST, PaperBench	Human-calibrated, economically grounded, or research-replication software work	Freelance software tasks, calibrated human baselines, or paper-replication subtasks	Software engineering, ML engineering, cybersecurity, scientific code	Task success, monetary value, human-time calibration, rubric grading	Closer to real value and effort than pass@k; costly, complex, and sometimes judge-dependent (Miserendino et al., 2025; Rein et al., 2025; Starace et al., 2025).

Using Mind Maps to Investigate the Typology of Students' Use of Large Language Models: A Preliminary Study

Goran Bubaš, Andreja Kovačić
University of Zagreb Faculty of Organization and Informatics
Pavlinska 2, 42000 Varaždin, Croatia
{gbubas, akovacic}@foi.unizg.hr

Abstract. *This preliminary qualitative study examines how hand-drawn mind maps can investigate higher education students' educational and private uses of large language models (LLMs). Of 74 participants in a convenience sample of Croatian undergraduate students enrolled in IT- and economics-related study programs, 59 readable, hierarchically structured mind maps were selected, transcribed, translated from Croatian into English, and analyzed using a hybrid approach combining LLM-based analysis with expert interpretation. Specific educational and private use categories were identified and supported. LLM-generated typologies of students' generative AI use converged around integration into everyday life, STEM/technical problem-solving, and reflective or skeptical evaluation of LLMs.*

Keywords. large language models, generative AI, student behavior, higher education, mind maps, qualitative analysis, typology, educational use, private use

1 Introduction

LLMs such as ChatGPT and Gemini are used for a multitude of purposes, and numerous empirical studies, such as those discussed below, have revealed that most of their use falls into two broad categories: (1) *private* and (2) *professional and/or educational* purposes. For instance, a study of messages exchanged with ChatGPT in July 2025 (Chatterji et al. 2025) revealed that the *predominant* conversation topics were related to practical guidance (28.8%), seeking information (24.4%), writing (23.9%; e.g., text editing or correction, translation, personal communication, etc.), multimedia generation (7.3%), self-expression (5.3%), and technical help (5.1%). Another large consumer-oriented study (Carloan et al., 2025), conducted among *U.S. adults*, uncovered that typical everyday usage is related to writing emails (18%), researching topics of interest (18%), managing to-do lists (18%), writing support (18%), meal planning (16%), managing expenses (15%), taking and organizing notes (15%), creating images (15%), researching purchases (15%), and researching health questions (14%). The *Deloitte 2025 Connected Survey* (Fineberg et al., 2025) showed that users have mostly conversed with generative AI tools (chatbots) regarding work assistance (67%, only employed respondents), product recommendations (53%), personal advice (44%), physical health or medical assistance (38%), travel planning (36%), nutritional assistance (35%), financial assistance (35%), and fitness coaching (33%). A survey presented on the *Statista* portal (Statista Research Department, 2025) found that in 2025, *adults in the U.S.* predominantly used AI for activities like tasks at work (28%), to learn something new (26%), and for entertainment (22%).

The most recent *Pew Research Centre* survey (Gotfried et al., 2026), conducted in February 2026, reported the following most frequent uses of AI chatbots by *adults in the U.S.*: search for information (42%), tasks at work - among employed adults (38%), fun or entertainment (25%), creation or editing of images or videos (24%), getting medical advice (20%), for diet and fitness information (20%), obtaining news (13%), emotional support or advice (10%). A similar survey performed in March 2026 among *adults* (18+) in the U.S. by *Epoch AI* and *Ipsos* (2026) showed that the most frequent uses for AI services (by LLMs like ChatGPT, Gemini, etc.) were looking up information or getting recommendations (80%), writing or editing text - emails, documents, summaries (59%), advice or learning - including "how-to" guidance, tutoring (55%), brainstorming or generating ideas (53%), creating or editing images (44%), summarizing or analyzing a long document or webpage (41%), technical work - math, data analysis, programming, etc. (37%), and personal reflection or chatting for fun (26%). Interestingly, the *AP-NORC* poll (AP-NORC Center for Public Affairs Research, 2025), conducted in July 2025, indicated that *younger adults* (aged 18-29) report much higher frequencies of AI use, in comparison to *all adults* in their sample, for various tasks like search for information (74%), coming up with ideas (62%), work tasks (52%), helping write emails (50%), creating or editing images (47%), entertainment (44%), shopping (40%), and companionship (25%).

When the research focus was on the *student population* and educational use of ChatGPT and other chatbots, one of the earliest global surveys (Ravšelj et al., 2025), conducted from November 2023 to February 2024 with 23,218 *higher education students* from 109 different countries and territories, revealed the following most frequent uses: brainstorming, summarizing long texts, finding articles for research, and writing texts. A similar large-scale study (Chung et al., 2026), conducted on *higher education students* in Australia in the August–October 2024 period, uncovered the following most

frequent uses of generative AI: editing or improving writing (67%), generating ideas and examples (64%), summarizing notes, readings or other materials (54%), teaching me how to do something (51%), getting feedback from generative AI (50%), finding information or conducting research (47%), generating or error-checking code, formula, mathematical calculations or data analysis (36%). A more recent survey (Chatzopoulos et al., 2025), conducted at *one university in Greece* from April to November 2025, showed that the most frequent uses of generative AI tools were: information search (85%), explaining and understanding difficult concepts in simple terms (59%), help with academic work - writing papers, generating ideas and inspiration, gathering study material (48%), exam preparation (43%), programming - code assistance, algorithm explanation, code debugging (36%), text writing (30%), creating or editing images (27%), and language assistance - conversation practice, translation, grammar (24%).

One of the most recent surveys (Stephenson and Armstrong, 2026), conducted by the professional organization *Savanta* in the United Kingdom in December 2025, detected an almost universal use of various types of AI tools (up to 95%) by *higher education students*, with the following activities being most frequently reported: generating text, e.g. ChatGPT (56%), summarizing, note-taking or quizzing university textbooks, e.g. Kortext (38%), enhancing and editing personal writing, e.g. Grammarly (37%), translation or language support, e.g. Google Translate (31%), generating images, videos or audio, e.g. DALL-E (22%), speech-to-text transcription, e.g. automatic captions on YouTube (18%), writing computer code, e.g. GitHub Copilot (17%), data analysis and presentation, e.g. Julius AI (17%), for friendship, company, advice or tackling loneliness (15%), another use related to my studies, e.g. medical software (13%), for counselling or therapy provided solely by AI (8%). The authors of this study also noted that AI tools are increasingly being used by students to prepare assessed work, for instance to explain concepts, summarize relevant articles, suggest ideas for research, structure their "thoughts", search the internet, generate text that is later edited by the student before it is, more or less, directly incorporated in the assessed work.

Finally, in a *review study* of 26 related academic papers by Naznin et al. (2025), the following categories/domains of types of AI uses by higher education students were identified: (1) *personalized learning* (real time feedback, adaptive learning, collaborative learning, student engagement, and self-regulatory learning); (2) *academic writing* (academic writing process, content generation, writing tools, over-reliance, and critical thinking); (3) *coding* (code generation, code debugging, and code explanation).

However, it must be noted that a *taxonomy or categorization* of students' use of large language models (LLMs) and generative AI (GenAI) based on analyses of data collected from the students themselves was rarely studied. Most research articles list the students' frequent activities with LLMs and GenAI. Still, the study by Zhang et al. (2026) investigated how the Latent Class Analysis (LCA) would identify *subgroups of students* based on their preference for the use of LLMs/GenAI for general purposes labelled "Fun," "Homework," "Research," "Study," across several tool types ("Code," "Data," "LLM," "Write," and "Other"). This analysis indicated five subgroups of higher education students: (1) study/assignment–focused users; (2) research-oriented power users; (3) routine mainstream/regular learners; (4) light users, whose use is more selective, bounded to fewer contexts, and narrower in tool coverage; (5) all-purpose, multi-tool, cross-context users who move across learning, assignments, research, and leisure and draw on a wider set of tools.

In the study that is presented in this paper, the *educational use* of LLMs and GenAI in itself is not the main focus, but instead the *main objective* is to investigate a potential *typology of students' LLM/GenAI-related behaviors* in their daily activities, using our *small convenience sample* as an *exploratory starting point*. Therefore, in this study, we will observe the students in our study as a *non-representative subsample* of LLM and GenAI student-users of different age groups and other sociodemographic characteristics. The *second objective* is to test an approach to qualitative analysis that is based on using a less conventional method in which students create mind maps of their LLM/GenAI use. Also, in this *hybrid approach*, following the coding of the students' mind maps into hierarchical verbal descriptions, several LLMs will be employed for *initial qualitative analyses* to examine the frequency of various activities and identify potential typologies of LLM/GenAI use.

2 Mind Maps and Mental Models

Mind maps can be used in qualitative research for various purposes like planning a project, data collection, analysis of visual and narrative data, as well as presentation of findings (Wheeldon and Åhlberg, 2019). In data collection, participants may only need a blank sheet of paper and a pen(cil) to begin drawing from a central main topic or 'core,' thereby creating a visual representation of how they perceive and experience the phenomenon under investigation. Both mind maps and concept maps can be used by visually oriented social science researchers since they are a form of a participant-centric means of associating user-collected data with theory (Wheeldon and Faubert, 2009). One means to apply mind maps as a data collection method is to start by demonstrating to the subjects in research at least one visual model or example of what a mind or conceptual map is, with some additional explanation of what is required by the participants. In fact, mind maps can be used not only to effectively collect personalized visualization data from research subjects who draw them, but also to organize complex transcripts or recordings of data collected with interviews (Fearnley, 2022). In addition, mind maps can be effectively used in STEM *(Science, Technology, Engineering, and Mathematics)* education since they have the potential to reduce cognitive load and enhance learning outcomes, including improvement of theoretical comprehension and critical thinking (Kefalis et al., 2025). In fact, one meta-analysis confirmed the effectiveness of their use on students' achievements in science and technology from primary to higher education (Wang et al., 2025, June).

Contrary to the relatively simple, basic, mostly hierarchical structure of mind maps, *mental models* are "internal representations that individuals construct to explain and predict the behavior of complex systems, including their expectations of the system's capabilities and limitations" that can be effectively used, for instance, to investigate *human–AI interaction* (Bodamer et al., 2026).

While there are many definitions of mental models, basically, in human-computer interaction (HCI) research, they are founded on *eliciting users' cognitive and behavioral model(s) of a specific computing system and its use*, i.e., they can help in "explaining how and why people use, rely on, trust, or reject these systems", including AI (Sanchez et al., 2026). For instance, Wang et al. (2025, March) used transcripts of oral interviews and drawing data to investigate, in simulated chatbot ecosystem settings, mental models of a small sample of users regarding their use of chatbots for hotel booking and they classified these models into four categories: (1) chatbot as a key player; (2) chatbot as a medium; (3) chatbot as a representation; (4) chatbot as an agent. Mehmood et al. (2025) used qualitative research design to identify the following three primary types of users' mental models about roles and implications of LLMs: (1) *enthusiastic view* of LLMs as helpful assistants and/or personalized support systems; (2) viewing LLMs as *conditional partners*, and (3) *apathetic views* of LLMs that involve seeing them as potentially dangerous and/or suspicious actors.

In our study, the focus will be on the qualitative research method of drawing mind maps of students' use of LLMs for educational and private purposes to (a) investigate the *variety of LLM use* for such purposes and (b) develop broad *initial categorizations or typologies* based on the students' drawings. The goal of this study is not so much to inspect the educational uses of LLMs by a small convenience sample of higher education students, but rather the *initial exploration of potential typologies* of their LLM/GenAI-related user behavior. Also, less conventional *qualitative* methods will be applied and tested, consisting of (a) using an LLM to assist in creating written transcriptions of students' mind maps and (b) using multiple LLMs for frequency analysis and creation of potential categorizations or typologies of students' educational and private uses of LLMs. In addition, the resulting typologies of LLM use may provide *partial insights into users' mental models of LLMs* and could perhaps be considered indirect external representations of some of their segments.

3 Study Goals, Problems, and Research Questions

The main goal of the *qualitative pilot study* that is presented in this paper was to investigate the use of mind maps as a method to collect data for the creation of an *initial broad typology* of students' educational and private uses of large language models. Accordingly, the following research problems (objectives) are defined for this pilot study:

1) To identify the potential variety of students' educational and private uses of large language models that are represented in their mind maps.
2) To investigate specific categories of students' educational and private uses of large language models based on data collected through mind maps.
3) To examine whether large language models can generate meaningful and interpretable typologies of students' LLM use from written transcripts of their mind maps.

From the previously cited studies (i.e., Chatterji et al., 2025; Carloan et al., 2025; Fineberg et al., 2025; Statista Research Department, 2025; Gotfried et al., 2026; Epoch AI and Ipsos, 2026; AP-NORC Center for Public Affairs Research, 2025) among the most common uses of LLMs in the general *adult population* were the following: searching for information and research; writing and editing text; work and educational assistance; advice and recommendations; brainstorming and idea generation; image and multimedia creation; technical assistance; summarization and information organization; health, nutrition, and fitness support; personal life support and entertainment. Specifically, regarding recently published research focused on the *students populations* (i.e., Ravšelj et al., 2025; Chung et al., 2026; Chatzopoulos et al., 2025; Stephenson and Armstrong, 2026), the following uses and activities were most present: information search and academic research; writing, editing, and improving academic texts; brainstorming and idea generation; summarizing texts, notes, and learning materials; explaining and understanding concepts; learning support and tutoring (learning how to do something); exam preparation and studying; programming, coding, and technical problem solving; translation and language support; image and multimedia generation; feedback on writing and academic work; data analysis and presentation.

Based on the first problem of this study and having in mind a possible comparison with the previous lists of LLMs uses, this first research question was formulated:

RQ1: *What educational and private uses of large language models are represented in students' mind maps?*

As was briefly discussed in the previous chapter 2. "Mind Maps and Mental Models," as a research method, mind maps can be used for collecting qualitative data that illustrate how the participants in a study visually represent their knowledge, experiences, and perceptions (Wheeldon & Åhlberg, 2019). In fact, use of mind maps and mental models are forms of participant-centered qualitative analysis (Wheeldon & Faubert, 2009) and may provide an alternative means of identifying representations of participants' thinking and behavior. Mind maps are a simplified form of mental models that are often used in human–AI interaction research (Sanchez et al., 2026). Some qualitative studies have successfully developed typologies of users' mental models from interviews and drawings, demonstrating that qualitative visual data can support the creation of meaningful user typologies (Wang et al., 2025, March; Mehmood et al., 2025). However, although typologies of users' mental models of AI exist, broad categorizations that are specifically related to *students' educational and private uses* of LLMs are still scarce, which provides a rationale for the following second research question:

RQ2: *What specific categories of students' educational and private uses of large language models can be identified from mind maps created by students?*

Finally, since the third problem of our pilot study was to investigate if and how LLMs can generate typologies of students' LLM use from written transcripts of their mind maps, a third research question is defined:

RQ3: *What is the potential of large language models to support qualitative analysis by generating meaningful typologies of students' educational and private uses of LLMs from mind map transcripts?*

For each of the previously listed research questions (RQ1–RQ3), a separate explanation of methodology and presentation of results of this preliminary study will be presented.

4 Methodology

This preliminary study uses *survey data* for collecting demographic information about the convenience sample of respondents and a *hybrid* qualitative approach using students' drawings of mind maps that depict their use of LLMs in educational and private contexts together with LLM-assisted thematic and content analyses of transcripts of the hand-drawn mind maps. Furthermore, selected LLMs were used to identify potential typologies (or "styles") of students' educational and private use of LLMs. This is a rather novel approach that uses mind maps and LLMs for qualitative analyses of their transcripts.

4.1 Subjects in Research and Data Collection

The initial convenience sample of participants in this qualitative study consisted of 74 students at a Croatian university who were enrolled in IT- and economics-related undergraduate study programs. These students volunteered for a set of learning activities on using ChatGPT in a course on *English for Special Purposes*. For participating in these learning activities, which lasted a total of 60 minutes and were performed in the English language in a computer laboratory with the use of ChatGPT as an LLM, the students were given small course credit. After completing each of the learning activities, the students used a survey to evaluate them. Then, after all of the brief learning activities were completed and evaluated, the students were asked to draw on an A4 piece of blank paper a mind map of their educational and private LLM use.

Out of 74 mind maps created by the students, 59 were selected for further analysis based on the following criteria: (1) the mind map had a clear hierarchical structure, (2) the hand-written text of the mind map was readable, (3) the students' mind map was not a copy or too similar to the *model mind map* they were provided with as an example. In other words, the final sample in our study consisted of 59 students who had drawn a *readable* and *original* hierarchical mind map on their *educational and private LLM use*. It must be noted that all students were anonymous in the survey for evaluation of learning activities and in drawing the mind maps.

The initial convenience sample of 74 students consisted of 66.2% male and 33.8% female participants. Also, most respondents (51, or 68.9%) were aged 18–19, while other respondents were aged 20–21 (21, or 28.4%), or 22 and above (2, or 2.7%). Almost all respondents (73 out of 74) were first-year students. Due to anonymity and the fact that the students were not writing their age and gender on the paper with their mind maps, the precise demographic composition is not known for the 59 students (Case #1 to Case #59) whose mind maps were chosen for further analyses.

Regarding the amount of use of tools such as ChatGPT prior to this study, more than two-thirds (67.6%) of the students in the initial convenience sample of 74 participants reported high or very high use, while only a small fraction (6.8%) reported no or very little prior use of ChatGPT. When asked specifically to assess the *average frequency* of their use of ChatGPT or similar tools in the previous 3–4 weeks, 16% of subjects reported using ChatGPT less than once in a week, 16% of respondents declared using it 1–2 times per week, 31% stated that they used it 3-4 times per week, and 37% indicated a frequency of use of 5 or more times per week. In addition, when the participants were asked to estimate the amount of time that they had spent using ChatGPT during the previous 3–4 weeks, 49% of them reported using it 1–2 hours per week, 26% of declared using it 3–4 hours per week, 17% used it 5–6 hours per week, and 4% for 7 or more hours per week. Also, three respondents, or 4%, did not report any ChatGPT use during the previous 3–4 weeks.

The respondents were also asked to list the AI tools they had previously used. The most frequently mentioned AI tools were ChatGPT (85%), Gemini (74%), and DeepSeek (28%). Other reported tools included Claude (18%), Perplexity AI (15%), Copilot (12%), Grok (11%), Astra AI (7%), and NotebookLM (3%). Among the tools rarely mentioned by individual respondents were Comet, ThetaWise, Medley AI, MathGPT, Sora, DALL·E, and Codex. With regard to the number of AI tools used by individual students, most respondents reported having previously used three AI tools (35%), followed by two tools (27%) and four tools (15%).

It should be emphasized that the students voluntarily participated in the study, that the data collection method ensured their full anonymity, and that the research was approved by the institutional Ethics Committee. Also, since the students were in their first year of study and from a single higher education institution, the findings of data collection and analyses cannot be generalized to other students in Croatia.

4.2 Qualitative Methods

The qualitative methods and analytical procedures that were used in this study were (1) drawing mind maps regarding respondents' personal and educational use of LLMs, (2) creation of written transcriptions of student-generated mind maps into a hierarchical textual structure, (3) thematic frequency analysis of written transcripts using LLMs, (4) identification of initial categorizations of general "styles" or "typologies" of students' use of LLMs in the educational and private contexts.

4.2.1 Drawing Mind Maps

For this rather simple qualitative data collection *(elicitation)* procedure the main intention was that the respondent-generated mind maps have a similar basic structure: (a) central point or root node consisting of one or more most used LLMs; (b) two main primary branches or first-level nodes that are related to *educational* and *private* use of LLMs; (c) secondary branches (other subbranches) that are freely identified by the respondents regarding common means of their LLM use. The students were instructed to create a mind map of their choice, but to indirectly guide their design, the instructor used a whiteboard marker to draw by hand on a whiteboard how a potential model mind map for a higher education teacher's use of LLMs may look. This "model" mind map consisted of the following: (1)

"ChatGPT" and "Gemini" written in the *primary/central node*; (2) "PRIVATE" and "BUSINESS" as *two main primary branches*; (3) subbranches "ARTICLES SEARCH," "THEORY EVALUATION," "IDEAS FOR RESEARCH," "TRANSLATION AND EDITING" in the "BUSINESS" part of the mind map; (4) subbranches "INFORMATION SEARCH," "NEWS CHECKING," "COOKING RECIPES," "HEALTH COMMENTS" in the "PRIVATE" part of the mind map.

The students were specially instructed not to make a copy of the potential ("model") mind map on higher-education teachers' use of LLMs, but to freely draw their own mind maps regarding subbranches and specific content that relate to their own uses of LLMs.

It must be emphasized that, for this study, the use of mind maps was chosen because (a) it is a semi-structured visual and textual method for eliciting students' data about their uses of LLMs, (b) they usually maintain a simple tree-structure branching radially, and (c) they are most often produced without cross-links. In other words, it was expected that most respondent-produced mind maps would be represented as hierarchically organized textual content that is not difficult for content analyses. Still, the use of a "model" mind map, despite the instruction to draw their own versions, may have influenced the creation of some students' mind maps.

4.2.2 Transcriptions of Drawings and Text in Mind Maps

If a hand-made drawing of a mind map has a simple hierarchical tree-structure that branches radially and has no cross-links then creating its textual transcript is a simple task of (a) denoting the central point or root node, (b) labelling the primary branches or first-level nodes, (c) listing the elements of the subbranches of the first level nodes or primary branches, (d) if there are third-level subbranches, or even lower level, then the hierarchical top-down procedure is repeated until the lowest level of subbranches has been documented.

For this pilot study (and to accelerate the manual process of making textual transcriptions of respondents' mind maps), several LLMs were tested for their ability to transcribe PDF scans of hand-drawn mind maps containing block and/or cursive handwriting. Since *Gemini 3.1 Pro* proved to be a convenient choice for this task, this LLM was requested to make an interpretation of each uploaded PDF scan of mind maps, one by one, by repeating the following basic prompt written in English: "Interpret the [N-th] mind map in the PDF (the text is in Croatian language, written by hand)." In the case of all drawings of mind maps, the interpretation or textual transcription of a given mind map created by Gemini was checked by the researcher, and when there was a misreading or misinterpretation, Gemini was prompted to correct its interpretation of such maps. In other words, textual transcriptions of all mind maps created by Gemini were checked by the researcher and corrected when needed. The use of the Gemini LLM considerably accelerated the process compared with *fully manual transcription* of all mind maps by a human researcher. After all the transcriptions of mind maps with a clean hierarchical structure were completed and checked, one final Word document was assembled with all acceptable transcriptions labelled as Case #1 to Case #59. It must be noted that Gemini not only made transcriptions in the Croatian language, but also translated all of the text in the Croatian to English. Following is an example of a translation to English of the basic structure of a mind map for "Case #21" (the exact transcript of the Croatian text is written in *italics*):

- Central Node: Main Context
  - *CHAT GPT*: The central hub of the map.
  - Top Branch: *EDUKACIJA* (Education)
    - *pomoć oko učenja*: Help with studying.
    - *pomoć pri rješavanju zadataka*: Assistance in solving tasks or assignments.
    - *pomoć za projekte, radove i sl.*: Help with projects, papers, and similar academic work.
  - Bottom Branch: *PRIVATNO* (Private/Personal Use)
    - *saznavanje više informacija o određenim stvarima koje me zanimaju*: Finding out more information about specific topics of personal interest.
    - *rješavanje određenih pitanja s kojima se susrećemo na dnevnoj bazi*: Resolving or answering specific questions that are encountered on a daily basis.

As can be concluded from the previous example, this preparatory phase for further qualitative analyses resulted not only in transcripts regarding the basic structure and Croatian text of the mind maps, but also in translations of the Croatian text to English.

4.2.3 Frequency Count of Mind Map Elements

Before the categorization of elements of the respondents' mind maps, the elements of those mind maps were counted from their transcripts for all 59 cases (Case #1 to Case #59). To avoid doing this process manually or using specialized content-analysis software, the Word document with the 59 transcripts was uploaded to four different LLMs *(Microsoft's Copilot, ChatGPT 4.6, Gemini 3.1 Pro, and Claude Sonnet 4.6)* and, for the *first round* (initially), they were prompted to perform a frequency count with the following additional remark in the prompt: "Conduct the analysis in the spirit of qualitative thematic analysis or content analysis, allowing yourself interpretive freedom in grouping, naming, and structuring themes."

Before the *second round* of counting, the labels and groupings for the frequency count generated by the aforementioned LLMs were examined by an *expert in evaluating LLM use*, and a unified list of categorization labels was created based on semantic analysis, synthesis, and correction of the initial labels/names and groupings/categorizations made by the three LLMs. Then, for the second round of frequency count, all four LLMs *(Microsoft's Copilot, ChatGPT 4.6, Gemini 3.1 Pro, and Claude Sonnet 4.6)* were employed to test and possibly confirm this, now final for this phase of research, unified categorization ("thematic analysis"). Finally, the results of the second round of frequency count by different LLMs were presented, compared, and commented on.

Due to the limited number of pages and the pivotal, "proof-of-concept" nature of the research reported in

this paper, we do not explain in greater detail how the previously mentioned categories were consolidated or what criteria were applied when determining whether different LLM-generated categories represented the same underlying concept.

4.2.4 Investigation of the Typologies of Students' LLM Use

After brief experimentation with prompts of various complexity and methodological detail, as well as having in mind the exposed diversity in the categorization and counting of denotations of "educational" and "private" uses of LLMs regarding the content of the subbranches of two primary nodes of respondents' mind maps (see subchapter 5.1 "Educational and Private Uses of LLMs"), it was decided to explore the potential of LLMs to identify potential typologies (or broad categories in the form of "styles") of LLM use by the most simple prompts. The exact prompt that was used with all four LLMs *(Copilot, ChatGPT 5.5, Gemini 3.1 Pro, and Claude Opus 4.8)*, after uploading the Word document with the transcriptions of respondents' mind maps, and their outputs is presented in Chapter 5.2 "Typologies of Students' LLM Use".

For this task, the selected latest and most advanced (best ranked) LLMs and their high "thinking level" were used at the time the analysis was conducted: ChatGPT 5.5 High, Gemini 3.1 Pro Extended, and Claude Opus 4.8 High. It is repeatedly emphasized that, since the LLMs approached the same complex categorization task quite differently in the probe trials, only the simplest and broadest prompt was used for this purpose.

4.2.5 Evaluation of Retrieved Initial Typologies of Students' LLM Use

In a *hybrid* approach, the initial typologies generated by the LLMs were evaluated and grouped by (1) a human expert in LLM use evaluation and (2) selected, most advanced LLMs available at the time the study was completed.

4.3 Instruments

Apart from the blank paper and pencil or pen that the students used for drawing mind maps, after completing the online learning activities in a computer laboratory for their *English for Special Purposes courses*, a survey was employed for the evaluation of each of these activities. At the beginning of this survey, a set of demographic questions was used to gather data on gender, age, and frequency of LLM use. This demographic data described the composition of the convenience sample.

For the investigation related to the problems of this study and research questions *(RQ1–RQ3)*, the following families of LLMs were used with a basic subscription plan: Microsoft's Copilot, OpenAI's ChatGPT, Google's Gemini, and Anthropic's Claude models. However, for the creation of written transcripts from the students' hand drawings of mind maps, only Google's Gemini was used.

4.4 Procedure

The procedure used for the study that is presented in this paper included a preliminary phase in February 2026 when the ability of an LLM to produce a transcript of a hand-drawn mind map was tested, as well as the potential to use LLMs in the analysis of textual transcript data derived from students' mind maps.

In February 2026, a group of 38 first-year student volunteers was recruited for the online learning activities, followed by a survey for the evaluation of each of these activities and their semi-directed free drawing of a mind map related to their use of LLMs for educational and private purposes. In June 2026, a different group of 36 new first-year student volunteers was recruited for the same purpose and activities.

The transcriptions of the mind maps and main data analyses with the use of LLMs were performed using the most advanced (latest) versions of LLMs also available to other researchers with a basic subscription at the time this study was completed. To avoid excessive token usage for large text processing tasks, the Claude Opus 4.8 model was not used in the first phases of analyses, but Claude Sonnet 4.6 was used instead.

5 Results and Discussion

In this section, the results of qualitative data analyses that were predominantly performed with the assistance of LLMs are presented, together with the basic prompts that were used for that purpose. The conducted analyses are broadly associated with (a) *thematic analysis* and (b) *content analysis*, which are among the most frequent qualitative methods (Chandrasekar et al., 2024). In addition, for coding the types of students' use of LLMs the *inductive method* was used, with the assistance of initial free categorizations of specific uses by several LLMs. For the inductive method, "codes and categories are not predetermined, but are instead identified and named as the researcher reads through the data," which may include "open or initial coding, and the constant comparative method of analysis" (Bingham, 2023). It is important to acknowledge that LLMs have already been applied to the analysis of qualitative data (see: Xiao et al., 2023; Zhang et al., 2025). However, the use of AI in qualitative data analyses may have insufficient transparency and generate divergent results upon repeated queries (Chen, 2025), as was the case in the study that is presented in this paper. Still, the application of a reliable research protocol for using AI tools like ChatGPT may considerably facilitate the qualitative data analysis process in the initial research stages (Goyanes et al., 2025).

5.1 Educational and Private Uses of LLMs

Having in mind the first problem of the study that is presented in this paper *("To identify the variety of students' educational and private uses of large language models that are represented in their mind maps")* and the related first research question (RQ1), the frequencies of each identified type of LLM use are presented in Table 1, Table 2, and Table 3. It must be emphasized that the row rubrics in Tables 1-3 are unified categorizations of LLM uses for which an expert combined the initial frequency count of four different LLMs, as was described in section 4.2.3 ("Frequency Count of Mind Map Elements") of this paper.

The exact prompt used with all four LLMs for generating the numerical data in Tables 1-3 was: "Using the

hierarchical transcript of students' mind maps on their use of LLMs such as ChatGPT and Gemini (provided in the Word document), perform a thematic analysis by producing separate concept counts for each of the following: A) central node categories (count each category separately); B) primary nodes (count each node separately); C) first-level branch category 1. Educational / Academic / School / College / Faculty (count each subcategory from 1a–1n separately); and D) first-level branch category 2. Private / Personal / Free Time / Everyday Life (count each subcategory separately). Use only the following fixed, predefined labels for the nodes and subbranches, and report the exact number of occurrences for each. (Do not count any concept more than once!)" *[Note: This first part of the prompt is followed by a list of predefined categories that are displayed in Tables 1-3.]*

The data presented in Table 1 indicates very high agreement among all four LLMs regarding the most frequently used LLMs mentioned in the central node of respondents' mind maps. ChatGPT was mentioned in 54–55 of the 59 processed mind maps, Gemini in 24 mind maps, and DeepSeek in 6 mind maps. A later, more detailed analysis revealed that in all 24 cases, Gemini was co-mentioned in the central node with ChatGPT. It should be recognized that the instruction was *not to include in the central node all the LLMs that the students use*, but only the one (or more) that they use most frequently.

Table 1. Frequency count of top three LLMs in the central node and top three categorizations of elements of mind maps in first-level nodes

LEVEL AND CONTENT	**LLM**†			
CENTRAL NODE *(TOP 3)*	**CO**	**CG**	**GM**	**CS**
ChatGPT	54	54	55	55
Gemini	24	24	24	24
Deep Seek	6	6	6	6
FIRST LEVEL NODES *(TOP 3)*	**CO**	**CG**	**GM**	**CS**
Educational / Academic / School / College / Faculty / As a student	50	**48***	58	49
Private / Personal / Free time / Everyday life	54	54	56	52
Business / Professional / Work	22	21	20	7

† CO – Copilot, CG – ChatGPT, GM – Gemini, CS – Claude Sonnet
* After prompting for a recount.

Regarding the first-level nodes related to broad categories of "education" (more specifically labelled for LLM prompting as "Educational / Academic / School / College / Faculty / As a student") and "private" (more extensively labelled for LLMs as "Educational / Academic / School / College / Faculty / As a student"), there was a notable disagreement (in the counting range from 48 to 58) for categorizations in the first-level node as "educational" and a much greater agreement (in the range between 52 and 56) among LLMs in categorizations in the first-level node "private." Some of the students also included the third type of primary node defined to the LLMs as "Business / Professional / Work". For brevity, the elements of this third primary node will not be analyzed, but this illustrates that the respondents chose the elements of the primary nodes rather freely, like the alternative wordings for the first two primary nodes that they have used.

The results of categorization of respondents' activities with LLMs, placed as *subbranches* in their mind maps, in the "education" category ("Educational / Academic / School / College / Faculty / As a student") in Table 2 demonstrate a rather high level of discrepancy or dissonance among the LLMs in counting the occurrences of concepts in the second-level branches (or further levels) of this category. These fixed categories of subbranches, which were used for prompting, were defined by the human expert as a unification of initial "free" categorizations performed by the four LLMs used in this study. The least diversity in categorizations was in relation to "Summarizing / shortening textual content" (18–20) and "Generating ideas for academic work / inspiration / project ideas" (13–15). In almost all cases (except one count by Claude Sonnet LLM regarding "Professional / client / business tasks linked to formal work"), the LLMs placed at least 4 and as many as 44 elements of subbranches in the predefined fixed categories, which can be interpreted as *support* ("confirmation") of those categories. Therefore, the data presented in Table 2 indirectly address the first two research problems and research questions (RQ1 and RQ2) in relation to the educational domain of students' activities with LLMs, *at least regarding the convenience sample that was used in this study*. It can also be stated that the revealed categories of educational activities in Table 2 correspond to the students' activities most frequently reported with LLMs in other recent surveys that were performed on much larger samples (for instance, see: Ravšelj et al., 2025; Chung et al., 2026; Chatzopoulos et al., 2026; Stephenson and Armstrong, 2026).

Regarding the previously presented findings, it is important to restate the possibility that some of the observed structures of students' mind maps at least partly reflected "task framing" instead of entirely spontaneous representations of students' mental models. In other words, some students may have relied on the model mind map of their professors' use of LLMs and the suggestion of two basic nodes (educational vs. private). This limitation of the presented study should be noted before interpreting its findings (see section 4.2.1 "Drawing Mind Maps").

Having in mind the discrepancies in counting the subbranches of the "education" primary node by different LLMs, it must be considered that within the same family of LLMs, for instance Claude, the versions Claude Sonnet 4.6, Claude Sonnet 5, and Claude Haiku 4.5 produced different counts for almost all fixed definitions of categories of subbranches. The same type of inconsistency was observed between Gemini 3.1 Pro and Gemini 3.5 Flash, as well as between Copilot with the "Think Deeper" and Copilot with the "Smart" option activated. In fact, the closest match, with mostly 0-2 different counts per type of subbranch, was between the ChatGPT 5.6 High and ChatGPT 5.5 High versions. However, in all the mentioned additional parallel/alternative analyses, the fixed categories for the "educational" subbranches were supported by only a few

individual counts as low as 2, while most counts were 5 or higher across all other categories.

Table 2. Frequency count of identified occurrences of *subbranches* that the LLMs placed in fixed categories for the **"education"** primary node

	LLM†			
EDUCATIONAL USES	**CO**	**CG**	**GM**	**CS**
Explanation, understanding and clarifying of study materials/concepts	33	44	33	40
Summarizing / shortening textual content	20	19	20	18
Solving tasks / problems / assignments / homework / answers	34	31	24	24
Creating study materials / study notes / scripts	28	12	11	10
Searching for sources / literature /papers / research	20	23	16	18
Exam preparation / quizzes / practice tests / knowledge checks / study review cards	18	15	12	13
Programming help	20	8	10	9
Writing support / grammar and spelling checking / proofreading / emails	22	20	20	14
Translation / foreign languages	10	8	6	6
Creating visual output, graphs, presentations, PPT / visual layout	20	13	13	13
Study planning / organization	10	7	6	5
Generating ideas for academic work / inspiration / project ideas	12	12	15	13
Analysis of data or theories	8	7	5	4
Professional/client/business tasks linked to formal work	15	10	7	1

† CO – Copilot, CG – ChatGPT, GM – Gemini, CS – Claude Sonnet

As can be concluded from the data displayed in Table 3, the respondents' most frequent activities in the category "private" (i.e., "Private / Personal / Free Time / Everyday Life") were related to information search using an LLM, cooking and nutrition, seeking advice / decision making, hobbies/skills and self-improvement, and entertainment. When the whole list of retrieved "private" uses of LLMs by the respondents in our study is taken into account, it very much corresponds to the non-work-related activities of adults that were reported in several recent studies by other authors (see, for instance: Carloan et al., 2025; Fineberg et al., 2025; Gotfried et al., 2026). In any case, regarding the first two problems of this study and research questions (RQ1 and RQ2), it can be concluded that very diverse private uses of LLMs were represented in students' mind maps and that the specific categories of students' private uses of LLMs have been successfully identified and supported. However, as was the case with the "educational" uses of LLMs, there was considerable diversity between the LLMs used, as well as, after additional checks, within versions of LLMs of the same family.

Table 3. Frequency count of identified occurrences of *subbranches* that the LLMs placed in fixed categories for the **"private"** primary node

	LLM†			
PRIVATE/PERSONAL USES	**CO**	**CG**	**GM**	**CS**
Searching for general information / trivia / general knowledge / curiosity	35	51	30	38
Cooking / recipes / nutrition / meal planning / calories	18	24	21	20
Travel and excursion planning / destinations search / restaurants and cafes	15	9	7	7
Health queries / symptoms / medical information	12	9	9	8
Decision-making / advice	20	30	14	18
Private image generation / visuals / multimedia creation or editing	22	9	10	9
Personal organization / private time management	12	9	6	6
Fitness / sports / training	14	12	10	10
Entertainment / media / games / movies / music	18	15	11	10
Language learning / translation – private context	10	7	7	6
Home repairs / practical household help / technical advice	8	9	6	7
Emotional support / stress reduction	6	4	3	2
Shopping / product comparison	8	3	4	3
Hobbies / skills / self-improvement / learning new things	20	22	5	19
News / current events	15	6	6	6
Writing private messages / emails	12	6	4	5

† CO – Copilot, CG – ChatGPT, GM – Gemini, CS – Claude Sonnet

The frequency counts by LLMs presented in Table 2 (subbranches of the "education" node) and Table 3 (subbranches of the "private" node) were not checked by a human expert, even though the counts differed across LLMs, because the intention was to find support for the categorizations and illustrate LLM inconsistencies, rather than to meticulously evaluate the diversity of outputs across various LLMs, which is a common phenomenon.

Another important issue is that the human experts were not used as coders of mind map elements (as a reference) to compare their findings with the LLM counts, as an evaluation of the LLMs' output. Even though this could be viewed as a limitation of the previously elaborated results of the study (in Section 5.1, "Educational and Private Uses of LLMs"), it was not the intention of the authors to perform such a comparison, since one of the main problems was to "examine whether large language models can generate meaningful and interpretable typologies of students' LLM use from written transcripts of their mind maps." Also, both RQ1 and RQ2 were directed only toward the potential usefulness of initial qualitative analyses performed by the LLMs and not the factually correct counting of individual students' mind map elements. Finally, because the convenience sample consists of first-year students from

IT- and economics-related study programs at a single Croatian university, the previous categorizations and consequent typologies cannot be generalized to the wider higher education student population in Croatia.

5.2 Typologies of Students' LLM Use

Bearing in mind the previously demonstrated diversity in the output of various LLMs and their versions regarding occurrence count of "educational" and "private" forms of LLMs use mentioned in the subbranches of the primary nodes of respondents' mind maps, it was decided to use only the simplest wording for the prompt for investigating the *typologies* of students' use of LLMs that could be deduced from the structure of the mind maps. Also, only the latest and most advanced (best ranked) LLMs at the time this analysis was conducted, with the highest optional "thinking" level, of the ChatGPT, Gemini, and Claude families were used for this task. After uploading the Word document with the transcriptions of 59 mind maps (Case #1 to Case #59), the following brief prompt was used in all three cases that will be briefly presented: "Read the Word document. Using an approach similar to cluster analysis, create a typology of LLM use (e.g., ChatGPT, Gemini) based on the textual representations (Cases #1 – #59) of students' mind maps visualizing their use of LLMs. The students were provided with an example containing two central nodes: (1) educational and (2) private. However, they were free to create a mind map of their own design. Create no more than 4–5 categories, with each category containing at least 5–8 student cases."

5.2.1 Typologies by ChatGPT 5.5 High

After the previously mentioned prompt, *ChatGPT 5.5 High* summarized its findings as "the dominant representation of LLMs is pragmatic and task-oriented. […] The original educational/private example clearly influenced many maps, but the actual clusters show a broader typology: LLM as tutor, producer, operator, life assistant, and tool ecosystem." Here are the descriptions of the main clusters proposed by ChatGPT 5.5 (italics are added by the authors):

1. ***"Study coach / concept scaffolder.*** […] Students in this group mainly represent ChatGPT/Gemini as a personal tutor: explaining difficult concepts, summarizing material, creating mini-tests, supporting exam preparation, and helping students understand required coursework. […] The tool is valued because it *explains, simplifies, tests, and repeats content in a personalized way.*"
2. ***"Academic producer / researcher / communicator.*** This type treats LLMs more as academic-work production infrastructure. Typical uses include finding sources, writing or editing emails, generating ideas for papers, creating presentations, checking grammar, summarizing academic texts, and producing documentation. These students use LLMs as productivity systems, not only to *help me learn*, but *help me produce, organize, polish, and communicate.*"
3. ***"STEM, code and task-solving operator.*** This group clusters around technical, computational, and problem-solving uses. The maps mention programming tasks, math tasks, code checking, bug fixing, IT support, C++, client design/communication, and task solutions. […] These students represent LLMs as operators or technical assistants: tools for getting through concrete tasks, debugging, coding, calculating, and solving."
4. ***"Whole-life personal assistant.*** This coherent type represents the LLM as a general companion for everyday life, not just a study or work tool. It includes personal interests, decisions, health, recipes, household issues, recommendations, motivation, communication advice, and practical life management. […] These students use LLMs as life-management assistants. The boundary between educational and private use becomes porous: the same tool helps with studying, health, food, organization, entertainment, and everyday decisions."
5. ***"Tool-ecology strategist / experimental or boundary-setting user.*** This type is defined not only by what students do with LLMs, but by how they differentiate, compare, or limit tools. Some cases explicitly separate tools by domain, compare prompts, use one tool as backup, or express skepticism. […] These students are more reflective about the AI ecosystem. They are not simply *using ChatGPT*; they are selecting, comparing, substituting, or resisting tools."

5.2.2 Typologies by Gemini 3.1 Pro Extended

The introductory comment by *Gemini 3.1 Pro Extended*, after the prompt to create users' typologies, was: "Based on a cluster-style qualitative analysis of the 59 student mind maps, the ways students use Large Language Models (LLMs) can be grouped into 5 distinct typologies. […] their self-designed configurations reveal unique behavioral archetypes based on their cognitive approach, technical depth, and level of life integration." The following are brief descriptions of the five typologies generated by Gemini 3.1 Pro:

1. ***"The Technical Developer & STEM Specialist.*** Core Approach: Algorithmic, functional, and highly specialized. These students primarily treat LLMs as an interactive IDE, a code-review partner, or a computational calculator. Their mind maps largely bypass standard essay writing or general brainstorming in favor of programming languages, logical debugging, and technical data manipulation. Key Use Cases: Debugging code syntax, learning C++, automating script creation, practicing mathematics, and handling server data encryption.
2. ***"The Academic Researcher & Strategic Writer.*** Core Approach: High-volume text synthesis and institutional compliance. These users deploy LLMs to navigate heavy academic workloads, process dense educational literature, and ensure strict alignment with rigid academic formatting criteria. Key Use Cases: Digesting long study scripts, generating targeted practice quizzes before exams, hunting down literature sources, and checking presentation layouts against strict styles like APA standards."
3. ***"The Holistic Lifestyle Designer & Personal Coach.*** Core Approach: Deep personal and somatic integration. For these individuals, the LLM functions as a proxy for specialized human experts—acting as a personal trainer, nutritional

scientist, therapist, or home economics consultant. Key Use Cases: Generating hyper-specific workout regimens (e.g., Ironman training), tracking daily calorie/carbohydrate counts, analyzing medical symptoms, managing burnout/stress, resolving communication dilemmas, and troubleshooting domestic malfunctions."

4. ***"The Pragmatic Generalist & Search Engine Substitute.*** Core Approach: Frictionless information retrieval and remedial academic assistance. This represents the largest group of students, who view the LLM essentially as an accelerated, conversational version of Google. Their interaction is highly reactive, broad, and conversational rather than heavily structural. Key Use Cases: Answering trivial or 'banal' daily questions, quick foreign language translations, basic homework assistance, looking up movie recommendations, and drafting initial outlines for assignments when they entirely lack personal motivation."
5. ***"The Strategic Tool-Evaluator & Skeptical Outlier.*** Core Approach: Meta-cognitive and highly evaluative. These users do not just map their tasks; they explicitly map the performance, constraints, and relationships of the AI tools themselves. They intentionally construct multi-model workflows or display strong, critical ethical boundaries regarding AI use. Key Use Cases: Setting up redundant backup tools (e.g., using DeepSeek strictly when ChatGPT fails), designing iterative prompt-chaining sequences to cross-compare results across different models, or entirely siloing the AI as a mandatory 'plagiarism machine' while rejecting it in personal life."

5.2.3 Typologies by Claude Opus 4.8 High

The response to the previously presented prompt by *Claude Opus 4.8 High* started with the following remark: "Below is a typology built by treating each mind map as a data point and clustering on the dimensions along which the maps actually vary, rather than the surface tool names (almost everyone uses ChatGPT, often plus Gemini, so the tool itself doesn't discriminate well). The clustering axes that turned out to be diagnostic: **Balance/tilt** between the academic and personal branches (one dominant vs. genuinely two-sided); **Depth of life-integration** on the private side (occasional info-lookup vs. AI woven through health, food, organization, well-being); **Tool ecosystem** (single ChatGPT vs. several tools vs. tools deliberately assigned to tasks); **Structural form** (did they follow the provided educational/private template or invent their own — function-based, role-based, workflow-based); **Relational stance** (AI as neutral utility vs. as advisor/companion vs. critically/sceptically)." The more detailed description by Claude Opus 4.8 High of these five types follows:

1. ***"Balanced Dual-Use Generalists.*** The modal pattern and largest cluster. These students adopted the provided template almost exactly: a two-sided split into educational and private, with a roughly even weighting and conventional, generic uses on each side — summarizing, explaining concepts, solving tasks and writing on the study side; information search, trivia, news, recipes and translation on the private side. One or two tools, no structural innovation, no strong tilt in any direction. They treat the LLM as a general-purpose helper for whatever comes up."
2. ***"Academic-Dominant Study Users.*** The map is heavily weighted toward studying; the private branch is thin, absent, or clearly secondary, and the framing is often about academic survival, efficiency, or shortcuts. Several explicitly foreground exam prep, summarizing course material step-by-step, or — at the sharper end — drafting work they're unmotivated to do or cheating. When a private branch exists it's minimal (one or two info-lookup items)."
3. ***"Everyday-Life Integrators.*** Here the personal side is rich and central rather than an afterthought. AI is genuinely embedded in daily life: health and symptoms, nutrition and cooking, fitness and training plans, travel, daily routines, organization (to-do lists, scheduling), home repairs, recommendations, and — at the deepest end — mental well-being. The academic branch is present and normal, but the distinctive signature is the breadth and specificity of life-management on the private side."
4. ***"Strategic Multi-Tool Power Users.*** These students relate to AI as a toolkit to be managed. The signature is one or more of: deliberately running several tools (often including specialist ones — MathGPT, DeepSeek, Grok, Copilot, NotebookLM, Comet, Sora), assigning tools to tasks or using one as a backup for another, building prompting/comparison workflows, technical/production work (coding, bug-fixing, client deliverables, dev & design), or abandoning the educational/private template entirely for a function-, process-, or role-based structure."
5. ***"Relational & Reflective Users.*** The smallest but most conceptually distinct cluster, defined by an unusual stance toward the AI. Two sub-flavors sit together here. Some personify or relationally lean on it — as an advisor, thinking-partner, source of an "objective opinion," or even explicitly a "psychologist," "musician," and stand-in "doctor." Others adopt a reflective or critical posture — using it to weigh perspectives and evaluate quality, or, at the extreme, rejecting it. What unites them is that the AI is treated as a counterpart for judgment or scrutiny rather than a neutral lookup utility."

5.2.4 Comparison and Integration of Retrieved Initial Typologies of Students' LLM Use

While, at first glance, the typologies of students' use of LLMs that were created by ChatGPT 5.5, Gemini 3.1 Pro, and Claude Opus 4.8 may seem rather diverse, a closer analysis of their content by an *expert in LLM use evaluation* revealed the following:

- Similarity between types labelled *Academic producer / researcher / communicator* (#2 ChatGPT 5.5), *The Pragmatic Generalist & Search Engine Substitute* (#4 Gemini 3.1 Pro), and *Academic-Dominant Study Users* (#2 Claude Opus 4.8).
- Strong resemblance between *Whole-life personal assistant* (#4 ChatGPT), *The Holistic Lifestyle Designer & Personal Coach* (#3 Gemini 3.1 Pro), and *Everyday-Life Integrators* (#3 Claude Opus 4.8).

- Moderate correspondence of *STEM, code and task-solving operator* (#3 ChatGPT 5.5), *The Technical Developer & STEM Specialist* (#1 Gemini 3.1 Pro), and *Strategic Multi-Tool Power Users* (#4 Claude Opus 4.8).
- Visible parallels between *Tool-ecology strategist / experimental or boundary-setting user* (#5 ChatGPT 5.5), *The Strategic Tool-Evaluator & Skeptical Outlier* (#5 Gemini 3.1 Pro), and *Relational & Reflective Users* (#5 Claude Opus 4.8).
- Some possible likeness between *Study coach / concept scaffolder* (#1 ChatGPT 5.5) and *The Academic Researcher & Strategic Writer* (#2 Gemini 3.1 Pro).
- Very little or no similarity with other categorizations for *Balanced Dual-Use Generalists* (#1 Claude Opus 4.8).

To test the correspondence between the previously listed associations performed by a human expert with similar semantic analyses of the typologies performed by some of the most advanced and best rated LLMs at the time this study was completed (ChatGPT 5.5, Gemini 3.1 Pro, Claude Opus 4.8, and Claude Fable 5), the following prompt was delivered to each of the LLMs: "Read the Word document. Having in mind the content of the Word document, answer which categories of students' use of LLMs in the three typologies proposed by A. ChatGPT, B. Gemini, and C. Claude are most similar? For example, there may be semantic overlap among the following categories: Academic Producer / Researcher / Communicator (#2, ChatGPT 5.5); The Pragmatic Generalist & Search Engine Substitute (#4, Gemini 3.1 Pro); Academic-Dominant Study Users (#2, Claude Opus 4.8). Analyze the semantic similarities among the five categories developed by each LLM, even when they are labeled differently. Identify and connect the categories across the three typologies that appear to represent the same or closely related patterns of student LLM use. Sometimes the cases that fall into specific categories are mentioned. For each set of similar categories, provide a brief explanation of the shared characteristics that justify the comparison."

Concerning the previous prompt, the LLMs were in consensus with the human expert and among themselves only with respect to (a) the *lifestyle set of categories* (i.e., Whole-life personal assistant, The Holistic Lifestyle Designer & Personal Coach, and Everyday-Life Integrators), (b) *STEM, code & power users categories* (i.e., STEM, code and task-solving operator, The Technical Developer & STEM Specialist, and Strategic Multi-Tool Power Users), and (c) *reflective tool evaluator and sceptic categories* (Tool-ecology strategist / experimental or boundary-setting user, The Strategic Tool-Evaluator & Skeptical Outlier, and Relational & Reflective Users). The other categories were distributed across various academic-use spectrums of interpretations. These may eventually be placed on a *single dimension* from (1) a simpler, task-oriented user with low specialization, and mostly reactive general-purpose use of LLMs, on the one end, to (2) the more sophisticated, "high-end" user with a higher level of specialization and structured, academic-production use of LLMs, on the other end.

Considering the third problem and research question (RQ3) of this pilot study, it can be concluded that the three most consistent and stable typologies of students' use of LLMs were identified as exhibiting (1) the *lifestyle set of attributes*, (2) *STEM and coding set of attributes*, and (3) *reflective tool evaluator and sceptic attributes.*

Finally, even though these categorizations are predominantly action-oriented rather than direct cognitive valuations, the students' types of LLM use may be broadly associated with their underlying *mental models*. The typologies indicate distinct perceptions of utility and modes of interaction, which are inherently connected to how students mentally construct and understand LLMs.

5.2.5 Important Remark on the Retrieved Initial Typologies of Students' LLM Use

The typologies of students' LLM use that were investigated and identified in the previous sections (5.2.1–5.2.4) demonstrate that similar assistance from LLMs could be useful and informative during the early phases of exploratory qualitative analysis, perhaps even shortening the initial probing phase of a study. Nevertheless, such LLM-generated groupings should be regarded as exploratory patterns or initial examples that may require further validation through (a) more scientifically established methodologies and (b) empirical data collection with greater precision than mind maps can provide, especially in the case of mental models research.

6 Conclusion

This pilot study shows that, with a sufficiently large sample, the use of student-created mind maps in combination with LLM-assisted data analysis can reveal meaningful categories of LLM use in both educational and private contexts. For ***RQ2***, students' *educational use* of LLMs was predominantly organized around learning support, academic production, problem solving, programming/STEM tasks, writing, summarization, and study organization, while *private use* centered on information search, everyday-life assistance, advice, entertainment, health, food, travel, hobbies, and personal communication. For ***RQ3***, LLMs showed clear potential to *support qualitative analysis* by generating interpretable initial ("probe") *typologies*, especially around (1) lifestyle integration, (2) STEM/coding use, and (3) reflective or skeptical tool evaluation. However, differences in categorization outputs among models for both RQ2 and RQ3 indicated that LLM-based typologies, both in this and other research areas, should be treated as exploratory and validated with expert interpretation of collected data.

References

AP-NORC Center for Public Affairs Research. (2025, July). *Young adults leading the way in AI adoption*. https://apnorc.org/projects/young-adults-leading-the-way-in-ai-adoption/

Bingham, A. J. (2023). From data management to actionable findings: A five-phase process of qualitative data analysis. *International Journal of Qualitative Methods, 22*, 1–11. https://doi.org/10.1177/16094069231183620

Bodamer, W. G., Schoenmakers, S., Nuñez Castellar, E., & IJsselsteijn, W. A. (2026). Users' mental models within human–AI interaction: A systematic scoping review. *AI & Society*. Advance online publication. https://doi.org/10.1007/s00146-026-03038-1

Carolan, S., Wu Martin, A., Gong, C. C., & Borja, S. (2025, June 26). *2025: The state of consumer AI*. Menlo Ventures. https://menlovc.com/perspective/2025-the-state-of-consumer-ai/

Chandrasekar, A., Clark, S. E., Martin, S., Vanderslott, S., Flores, E. C., Aceituno, D., Barnett, P., Vindrola-Padros, C., & Vera San Juan, N. (2024). Making the most of big qualitative datasets: A living systematic review of analysis methods. *Frontiers in Big Data, 7*, Article 1455399. https://doi.org/10.3389/fdata.2024.1455399

Chatterji, A., Cunningham, T., Deming, D. J., Hitzig, Z., Ong, C., Shan, C. Y., & Wadman, K. (2025, September). *How people use ChatGPT* (NBER Working Paper No. 34255). National Bureau of Economic Research. https://doi.org/10.3386/w34255

Chatzopoulos, A., Zacharia, P., & Kantaros, A. (2026). Generative AI in Higher Education: A Large-Scale Study of Student Usage Patterns, Applications and Motivations. *Applied Sciences*, 16(12), 5972. https://doi.org/10.3390/app16125972

Chen, W. (2025). Exploring the application of AI to qualitative data analysis: A comparative study in the field of industrial design education. In M. Kurosu (Ed.), *Artificial intelligence in HCI: 6th International Conference, AI-HCI 2025, held as part of the 27th HCI International Conference, HCII 2025, Gothenburg, Sweden, June 22–27, 2025, proceedings, part II* (pp. 131–142). Springer. https://doi.org/10.1007/978-3-031-93415-5_8

Chung, J., Henderson, M., Slade, C., Liang, Y., Pepperell, N., Corbin, T., Walton, J., Yu, A. S., Bearman, M., Buckingham Shum, S., Fawns, T., McCluskey, T., McLean, J., Oberg, G., Seligmann, A., Shibani, A., Bakharia, A., Lim, L.-A., & Matthews, K. E. (2026). The use and usefulness of GenAI in higher education: Student experience and perspectives. *Computers and Education Open, 10*, Article 100347. https://doi.org/10.1016/j.caeo.2026.100347

Epoch AI, & Ipsos. (2026, April). *Topline & methodology: Epoch AI / Ipsos National AI Usage Survey*. Ipsos. https://www.ipsos.com/sites/default/files/ct/news/documents/2026-06/Topline_Epoch%20AI%20-%20Ipsos%20National%20AI%20Usage%20Survey.pdf

Fearnley, C. J. (2022). Mind mapping in qualitative data analysis: Managing interview data in interdisciplinary and multi-sited research projects. *Geo: Geography and Environment, 9*(1), e00109. https://doi.org/10.1002/geo2.109

Fineberg, S., Hupfer, S., Loucks, J., Steinhart, M., & Mazumder, S. (2025, September 25). *In the gen AI economy, consumers want innovation they can trust: Deloitte's 2025 Connected Consumer Survey reveals that, by pairing strong data responsibility with bold innovation, tech companies can win greater trust, loyalty, and spending*. Deloitte Insights. https://www.deloitte.com/us/en/insights/industry/telecommunications/connectivity-mobile-trends-survey.html

Gottfried, J., Bishop, W., Anderson, M., Faverio, M., Park, E., & McClain, C. (2026, June 17). *Americans and AI 2026: Chatbots, smart devices and views on impact*. Pew Research Center. https://www.pewresearch.org/internet/2026/06/17/americans-and-ai-2026-chatbots-smart-devices-and-views-on-impact/

Goyanes, M., Lopezosa, C., & Jordá, B. (2025). Thematic analysis of interview data with ChatGPT: Designing and testing a reliable research protocol for qualitative research. *Quality & Quantity, 59*, 5491–5510. https://doi.org/10.1007/s11135-025-02199-3

Kefalis, C., Skordoulis, C., & Drigas, A. (2025). A systematic review of mind maps, STEM education, algorithmic and procedural learning. *Computers, 14*(6), 204. https://doi.org/10.3390/computers14060204

Mehmood, K., Verleye, K., De Keyser, A., & Larivière, B. (2025). From promises to practice: Unraveling users' mental models about the role and implications of large language models (LLMs) for individuals and society. In *AIRSI 2025 Proceedings Book* (pp. 153–157). AIRSI.

Naznin, K., Al Mahmud, A., Nguyen, M. T., & Chua, C. (2025). ChatGPT integration in higher education for personalized learning, academic writing, and coding tasks: A systematic review. *Computers, 14*(2), 53. https://doi.org/10.3390/computers14020053

Ravšelj D, Keržič D, Tomaževič N, Umek L, Brezovar N, A. Iahad N, et al. (2025) Higher education students' perceptions of ChatGPT: A global study of early reactions. *PLoS ONE 20*(2): e0315011. https://doi.org/10.1371/journal.pone.0315011

Sanchez, T., Vereschak, O., & Deroy, O. (2026). Mental models in human-AI interaction: Systematic review of empirical methodologies and guidelines. In *Proceedings of the 31st International Conference on Intelligent User Interfaces (IUI '26)* (pp. 663–682). Association for Computing Machinery. https://doi.org/10.1145/3742413.3789223

Statista Research Department. (2025, November 28). *U.S. adults using ChatGPT 2025, by activity* [Graph]. In *Statista*. https://www.statista.com/statistics/1461638/us-adults-chatgpt-usage-by-activity/

Stephenson, R., & Armstrong, C. (2026). *Student generative AI survey 2026* (HEPI Report No. 199). Higher Education Policy Institute. https://www.hepi.ac.uk/wp-content/uploads/2026/03/HEPI-Report-199-Gen-AI-Survey-2026.pdf

Wang, X., Wang, X., Park, S., & Yao, Y. (2025, March). Users' mental models of generative AI chatbot ecosystems. In *Proceedings of the 30th International Conference on Intelligent User Interfaces (IUI '25)* (pp. 1016–1031). Association for Computing Machinery. https://doi.org/10.1145/3708359.3712125

Wang, X.-M., Wang, J.-L., Xu, S.-Y., & Xu, S.-J. (2025, June). Concept mapping in STEM education: A meta-analysis of its impact on students' achievement (2004–2023). *International Journal of STEM Education, 12*, Article 30. https://doi.org/10.1186/s40594-025-00554-2

Wheeldon, J., & Åhlberg, M. (2019). Mind maps in qualitative research. In P. Liamputtong (Ed.), *Handbook of research methods in health social sciences* (pp. 1113–1129). Springer. https://doi.org/10.1007/978-981-10-5251-4_7

Wheeldon, J., & Faubert, J. (2009). Framing experience: Concept maps, mind maps, and data collection in qualitative research. *International Journal of Qualitative Methods, 8*(3), 68–83. https://doi.org/10.1177/160940690900800307

Xiao, Z., Yuan, X., Liao, Q. V., Abdelghani, R., & Oudeyer, P.-Y. (2023). Supporting qualitative analysis with large language models: Combining codebook with GPT-3 for deductive coding. In *Companion proceedings of the 28th International Conference on Intelligent User Interfaces* (pp. 75–78). Association for Computing Machinery. https://doi.org/10.1145/3581754.3584136

Zhang, J., Wang, C., & Liu, M. (2026). Generative AI use profiles and academic outcomes in higher education: The role of institutional support. *Interactive Learning Environments*. Advance online publication. https://doi.org/10.1080/10494820.2026.2668610

Zhang, H., Wu, C., Xie, J., Lyu, Y., Cai, J., & Carroll, J. M. (2025). Harnessing the power of AI in qualitative research: Exploring, using and redesigning ChatGPT. *Computers in Human Behavior: Artificial Humans, 4*, Article 100144. https://doi.org/10.1016/j.chbah.2025.100144

A Hybrid Document Reading, Few-Shot Classification, and LLM Extraction Pipeline for Accounting Document Processing

David Slavik
Artificial Intelligence Laboratory
University of Zagreb Faculty of Organization and Informatics
Pavlinska 2, 42000 Varaždin, Croatia
dslavik@foi.unizg.hr

Tim Juić
Junior Developer
Softcon
Jakopovec, Marijana Lončara 20A, Croatia
tjuic@softcon.hr

Abstract. *Croatian accounting and insurance-brokerage companies handle large volumes of Croatian-language paperwork, including invoices, ledgers, contracts, commission lists, and many kinds of insurance policies, that need to be read, classified, and have key fields entered into a bookkeeping system. We present AppAI, a pipeline for processing documents that combines a vision-language model for grounded text extraction, a few-shot SetFit classifier trained on Croatian bookkeeping and insurance document classes labeled by project partners, and a self-hosted large language model for structured attribute extraction per document type. We position this design against recent work on OCR-free document understanding, few-shot text classification, and layout-aware extraction. On a single-partner held-out external validation set, classification reached 94.4% accuracy across 14 document types; over scored structured fields, extraction reached 87.80% field accuracy.*

Keywords. document classification, vision-language models, few-shot learning, large language models, information extraction

1 Introduction

Croatian accounting companies that also work as insurance brokers handle a steady stream of paperwork in Croatian: output invoice ledgers (cro. *izlazni računi*, *IRA*) and input invoice ledgers (cro. *ulazni računi*, *URA*), general invoices (cro. *fakture*), contracts (cro. *ugovori*), commission lists (cro. *provizijske liste*), and many kinds of insurance policies: compulsory motor vehicle liability (cro. *obvezno auto osiguranje*), CASCO (Casualty and Collision) insurance (cro. *kasko osiguranje*), property (cro. *osiguranje imovine*), liability (cro. *osiguranje odgovornosti*), travel (cro. *putno osiguranje*), life (cro. *životno osiguranje*), voluntary and supplementary health (cro. *dobrovoljno* and *dopunsko zdravstveno osiguranje*), employee collective accident (cro. *kolektivna nezgoda*), crop (cro. *osiguranje usjeva*), animal (cro. *osiguranje životinja*), and construction-risk (cro. *objekti u izgradnji*) insurance, plus vehicle registration certificates (cro. *prometne dozvole*). The workflow requires staff to identify the document type, then transcribe fields such as policyholder, OIB (cro. *Osobni identifikacijski broj*, the Croatian personal identification number), policy number, premium, covered risks, and dates into a bookkeeping or claims system. In the partner workflow motivating this work, much of this is still done by hand: slow, error-prone, and harder as volume grows.

AppAI is designed as a service for this workflow: given a document, it reads content in the Croatian language from digital PDFs or scans, classifies the document within a bookkeeping and insurance taxonomy, and extracts the attributes needed by downstream business systems. We built this as a hybrid pipeline, not one end-to-end model, because no single method is best at all three steps. Vision-language models are good at OCR-free, layout-aware reading. Few-shot classifiers work well when labels are scarce and the taxonomy keeps changing. Instruction-tuned LLMs can extract a different schema per document type without retraining. These documents also carry personal and financial data subject to Croatian and EU privacy requirements, so every model in the pipeline is self-hosted rather than called through a third-party API; Section 3 describes the anonymization measures applied before any document enters a training or evaluation corpus.

Two simpler alternatives were set aside: a single end-to-end vision-language model, which would need retraining or re-prompting whenever the taxonomy or a per-type schema changes; and per-class LLM prompting for classification (our earlier design, Section 4), which cost about five seconds per document. Its main contribution is the applied integration of these existing techniques into a working, self-hosted pipeline for Croatian accounting and insurance-brokerage document processing, evaluated on genuinely held-out, single-partner data.

The rest of the paper is organised as follows. Sec-

tion 2 reviews the literature behind each stage; Section 3 describes the architecture as currently implemented; Section 4 covers the classification model's training and evaluation; Section 5 evaluates the full pipeline on a held-out validation set; Section 6 discusses the results and directions for future work; Section 7 concludes.

2 Related Work

2.1 OCR-free and vision-language document reading

Classical document pipelines typically separate text recognition from downstream understanding: an OCR engine such as Tesseract (Smith, 2007) detects and recognizes text, and subsequent models consume the OCR output, often together with layout information. Transformer-based recognizers such as TrOCR (Li et al., 2023) improved recognition accuracy, but still function primarily as text-recognition components rather than end-to-end document-understanding systems.

Donut (Kim et al., 2022) showed a different approach: an OCR-free transformer that processes document images directly and reports competitive performance on visual-document-understanding benchmarks, avoiding a separate OCR stage and its error propagation. More general instruction-following multimodal systems such as LLaVA (Liu et al., 2023) and GPT-4V (OpenAI, 2023) established prompted vision-language interaction as a general capability.

By early 2025, VLMs with document understanding capabilities, such as Qwen2.5-VL (Bai, Chen, et al., 2025), were reporting OCR, document parsing, layout, and grounding capabilities closer to the requirements of AppAI's reading stage. Alongside these generative approaches, conventional document-conversion tools such as Docling (Livathinos et al., 2025) can parse programmatic text from digital PDFs and use OCR/layout/table models where needed to build a structured representation, a non-generative alternative to VLM-based conversion.

In the current implementation, AppAI follows the OCR-free direction, but uses a general-purpose vision-language model rather than one trained specifically for documents: a Qwen3-VL-8B-Instruct instance (Bai, Cai, et al., 2025), served through vLLM (Kwon et al., 2023), is prompted to return recognized text and bounding boxes for detected text regions in one pass. Docling is also available as a configurable alternative for use cases where a lightweight, non-generative conversion pipeline is preferred. We did not adopt layout-fused encoders like the LayoutLM family or Donut's own classifier head, mainly because AppAI's reading stage is decoupled from classification and has to support whatever prompt the extraction stage needs later, not a fixed label set.

2.2 Few-shot document-type classification

Once text is available, a small taxonomy that keeps changing is often better served by methods that need only a handful of labeled examples than by training a large classifier from scratch for every new type. SetFit (Tunstall et al., 2022) first fine-tunes a small sentence-embedding model on contrastive pairs drawn from a small labeled set, such as eight examples per class, then trains a lightweight classifier on top of the resulting embeddings. In the reported benchmark settings, a 110M-parameter SetFit model scores close to T-Few 3B, a prompt-based few-shot method roughly 27 times larger, while needing far less compute: SetFit is estimated to train and run about 19 times faster, and in practice took about 30 seconds to train on a single V100 GPU ($0.025 per run) versus roughly 700 seconds on an A100 for T-Few 3B ($0.70 per run), 28 times cheaper.

Multilingual embedding models such as BGE-M3 (Chen et al., 2024) provide multilingual retrieval within a single backbone, building on the broader sentence-embedding tradition established by Sentence-BERT (Reimers & Gurevych, 2019), while Croatian-specific encoders also exist, notably BERTić (Ljubešić & Lauc, 2021), an ELECTRA-style model trained on Bosnian, Croatian, Montenegrin, and Serbian web text. Contrastive training methods such as SimCSE (Gao et al., 2021) support the same representation-learning intuition behind SetFit: training on positive and negative pairs helps embeddings place similar texts close together and dissimilar texts farther apart in the embedding space.

A 2024 study re-evaluating document image classification with large language models (Scius-Bertrand et al., 2024) reports zero-shot GPT-4 at 61.8% accuracy on OCR text and GPT-4V at 69.9% on page images, on a low-label RVL-CDIP subset, with OCR-text embedding k-NN reaching up to 67.8% as labeled examples grow from 160 to 1,600. This comparison is useful for AppAI because it evaluates prompted classification against embedding-based baselines under low-label conditions, where repeated API calls, variance, and operational cost also matter.

This data efficiency will matter for AppAI: its Croatian bookkeeping and insurance taxonomy is expected to grow as partner companies and document types are added, and collecting thousands of labeled Croatian examples for every new class would be impractical, so a SetFit classifier was the natural fit. We fine-tune the multilingual BGE-M3 embedding model rather than BERTić, mainly because BGE-M3 was already available and tested on a much broader set of benchmarks; Whether a regional encoder such as BERTić would do better on this taxonomy is still untested. The corpus later grew to 2,427 documents (Section 4) as more classes were added; SetFit's few-shot efficiency is per class, independent of total corpus size.

2.3 Generative extraction with large language models

Classification identifies the document type; downstream processing also requires the document contents. DocLLM (Wang et al., 2024) shows layout-aware generative extraction can work without a full vision encoder, by feeding OCR text and spatial layout straight into the model. More generally, instruction-following models such as GPT-4 (OpenAI et al., 2023) helped popularize prompt-based workflows that return structured output, although reliable production use still requires validation around the model output. Dedicated Croatian generative models are also emerging, including HR-LLM, a completed pilot effort that built an early monolingual Croatian LLM still in beta, presented at NeTTT 2024 (Štefanec et al., 2024).

We assign extraction to a separate instruction-tuned LLM rather than extend the classifier, since the output schema differs by document type and changes more often than the type taxonomy does. We follow the simpler pattern used by GPT-4 rather than DocLLM's layout-aware approach: a self-hosted LLM (Qwen3-14B (Qwen Team, 2025), served through vLLM) receives a document-type prompt template plus the OCR text, and returns the requested attributes as JSON for downstream schema validation. We use a general multilingual model rather than a monolingual Croatian one because efforts such as HR-LLM have so far produced research pilots and resources still in beta rather than production alternatives for this pipeline.

2.4 Benchmark scope and evaluation caution

RVL-CDIP (Larson et al., 2023) remains one of the most widely used benchmarks in document classification, but it is known to contain label noise, ambiguous classes, and train/test overlap. A 2026 follow-up by the same group quantifies these issues further (12% label error, roughly 35% test-train duplication) and releases corrected RVL-CDIP variants, finding that fixing label errors improves out-of-distribution generalization by an average of 8.1% (Larson et al., 2026). The *Beyond Document Page Classification* study (Van Landeghem et al., 2024) goes further, arguing that real document classification is usually multi-page and spans far more varied domains and class distributions than a single-page, generic-class benchmark captures.

AppAI is not directly benchmarkable against RVL-CDIP, since its taxonomy is specific to Croatian accounting and insurance documents. These critiques mainly matter as a reason to evaluate against a held-out dataset labeled by project partners instead, rather than use a public leaderboard number to claim deployment readiness; this is the approach taken in Section 5.

3 System Architecture

AppAI runs as a set of independently deployable services behind a central backend that orchestrates the pipeline, so that each stage can be scaled, updated, or swapped for an alternative without redeploying the others. The document flow has three stages: reading, classification, and extraction, illustrated in Fig. 1.

Reading. When a document is ingested, the system routes it to Qwen3-VL-8B-Instruct, deployed in FP8 precision through vLLM, as the primary VLM-based reading engine. The model is prompted to return recognized text and bounding boxes for detected text regions. Docling is available as a configurable alternative for cases where a more lightweight, non-generative conversion pipeline is preferred. This stage outputs recognized text plus layout metadata, cached for later stages. The bounding boxes are currently kept mainly for caching and auditability; classification and extraction, below, use only the recognized text. Section 6 returns to this and discusses using the layout signal more directly for document types where text alone is not enough.

Classification. The recognized text goes to a separate classification component, which embeds it with a BGE-M3-based SentenceTransformer model (Chen et al., 2024) fine-tuned with SetFit, then runs a logistic regression head trained on a labeled corpus provided by project partners, covering Croatian bookkeeping classes (output and input invoice ledgers, general invoices, contracts, commission lists), vehicle registration certificates, and insurance policy classes (compulsory motor vehicle liability and CASCO insurance, property, liability, travel, health, life, employee collective accident, crop, animal, and construction-risk insurance). It returns a predicted document-type label and an uncalibrated classification score. Section 4 covers how this model was trained and evaluated.

Extraction. Once a document type is known, the system selects a prompt template listing the attributes to extract, such as policyholder name, OIB, policy number, covered risks, premium, and dates for an insurance policy, or invoice number, sender, and total amount for an invoice, and sends the recognized text plus the template to a self-hosted LLM (Qwen3-14B-AWQ (Qwen Team, 2025), served through vLLM, temperature 0.1, thinking mode off, 16,384-token context with 2,048 reserved for output). The model returns the attributes as structured output. Schema validation checks required fields and data types before the response goes out; semantic correctness is evaluated separately in Section 5. Each template states the required and optional fields, a few document-specific hints, and an instruction to return only structured output, parsed straight out of the model's response.

Keeping the three stages loosely coupled has practical value: the classification taxonomy and extraction schemas can change without retraining or redeploying

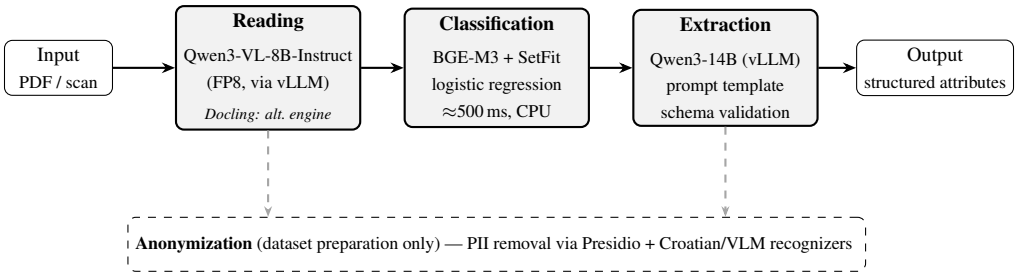


Figure 1. AppAI system architecture. The production pipeline (top) processes each document through three stages; the anonymization component (dashed) operates during dataset preparation for training and evaluation and is not part of the live inference path.

the reading stage, and the reading stage itself can be swapped (VLM versus Docling) without modifying the rest of the pipeline. A separate anonymization component applies Presidio-based PII detection (Data Privacy Stack, 2024), custom Croatian recognizers, and a VLM-based visual-region detector to remove identifying text and visual elements (e.g., signatures, stamps, handwriting) before documents enter training or evaluation corpora; it is not a stage of the live inference pipeline above.

4 Classification Model: Training and Evaluation

Classification is the best-tested stage so far: it has gone through two trained iterations as the labeled corpus grew. Classification in AppAI was not always done this way; an earlier version sent one LLM prompt per candidate class, 16 calls per document, taking around five seconds, with each prompt hand-curated around concrete cues (e.g. CASCO's theft, fire, vandalism, insured vehicle value, and deductible amount). The system was migrated to a SetFit classifier instead: after OCR/text extraction, one embedding pass plus a logistic regression lookup takes about 500 ms on a multi-core CPU server (an AMD EPYC 7V13, 64 cores).

Training and evaluation for both SetFit iterations were done by our project team, using GPU infrastructure provided by project partners. Both iterations fine-tune BAAI/bge-m3 (568M parameters) with SetFit, using a CoSENT contrastive loss for the embedding stage and a scikit-learn logistic regression head for classification, on an 80/20 train/test split. These held-out splits, labeled by project partners, support comparing model iterations within the same corpus, under the current taxonomy and document collection. The split was randomized at the document level rather than stratified by issuer, template, time period, or near-duplicate document family, so internal-split accuracy can partly reflect template recognition; the external validation set in Section 5, evaluated on genuinely unseen documents, is the more decisive measure of generalization. The 114-document split used to pick the embedding backbone (Table 2) is likewise a model-selection set, not an independent test set.

Table 1 reports two engineering iterations. Because the OCR reader, corpus size, label taxonomy, and training settings changed together, the comparison reflects overall system progress rather than the isolated effect of one variable.

Before settling on this approach, three options were shortlisted. Classifying directly on frozen sentence embeddings needs only a small labeled set and no training step, but was expected to struggle distinguishing similar classes, for example two insurance types with overlapping wording. Fine-tuning a BERT-style classifier typically needs a larger labeled set and a full training run, with potentially higher accuracy at a much higher labeling cost. SetFit sits in between: it needs only a small labeled set, uses a fast contrastive fine-tuning step, and, according to the literature, adapts better than frozen embeddings while needing far less data than full classifier fine-tuning. SetFit was selected as the most practical trade-off for the available labeled data.

First iteration. Within SetFit, three embedding backbones were fine-tuned and compared on the same corpus of 612 documents across 17 classes: a small multilingual MiniLM, multilingual-E5-base, and BGE-M3. Table 2 summarizes the comparison. The MiniLM result also reflects its shorter input window, so the comparison is operational rather than a controlled comparison of backbone choice. A fourth candidate, multilingual-E5-base used only as a frozen embedding model, represented the embedding-only alternative above and reached 91% test accuracy on the same held-out split; it was dropped before the check on the full corpus, consistent with the expectation that frozen embeddings adapt less well to closely related classes within the same domain.

Four classes were underrepresented in this corpus: 1 document for the input invoice ledger, 6 for crop insurance, 12 for life insurance, and 14 for animal insurance, with roughly 50 documents per class otherwise. BGE-M3 had the most parameters and the largest training and inference footprint of the four, but also the highest accuracy, and was chosen as the system's backbone going forward.

Table 1. Training configuration and results for the two SetFit classifier iterations.

Parameter	Iteration 1	Iteration 2
Base model	BAAI/bge-m3 (568M params)	BAAI/bge-m3 (568M params)
OCR method	Docling	Qwen3-VL-8B-Instruct (FP8)
Documents	612 (17 classes)	2,427 (15 classes)
Docs per class	$\sim$50 (4 classes $<$ 50)	$\sim$200 (4 classes $<$ 200)
Train / test split	498 / 114	1,943 / 484
Max input tokens	512	768
Batch size	8	6
Epochs	3	2
Contrastive iterations	10	10
Loss function	CoSENT	CoSENT
Test accuracy	**98.25%** (114 held-out docs)	**98.87%** (484 held-out docs)
Training time	$\sim$30 min	$\sim$3 h
Training GPU	RTX 4090 (24 GB)	RTX 4090 (24 GB)

Table 2. Embedding backbones fine-tuned and compared for the SetFit classification stage, all trained on the same 612-document corpus.

Model	Params	Max tok	Test acc. (114 held-out)
MiniLM	118M	128	93.86%
E5-base	278M	512	97.37%
BGE-M3	568M	512	**98.25%**

The first iteration's few held-out errors were manually traced to incomplete OCR text rather than a genuine embedding or classifier failure; the same OCR-error-propagation pattern from Section 2.1 showed up again at a larger scale below. When we later evaluated this model against a larger pool of 2,436 documents, after rerunning OCR with Qwen3-VL-8B-Instruct at FP8 precision, accuracy dropped to 92.98%. This decrease was driven primarily by vehicle registration certificates, whose per-class accuracy fell to 49%. The first-iteration model had been trained on Docling text, where many vehicle registration certificates had little usable text; even after rerunning OCR at evaluation time, the class remained poorly learned. After retraining on VLM-read text, the second-iteration model reached 100% accuracy for vehicle registration certificates in the internal per-type breakdown, and none of the 6 external validation misclassifications (Section 5) involved this type either.

Second iteration. We reran text extraction on the whole corpus with the VLM-based reader, growing the processed pool to 2,436 documents; after excluding crop insurance (6 documents) and the input invoice ledger (3 documents) for the data-scarcity reasons given below, 2,427 documents across 15 trained classes remained for training, with about 200 documents per class otherwise, input truncated to 768 tokens, 2 training epochs, and the same CoSENT loss and contrastive-iteration count. This model reached 98.87% accuracy on its 484-document held-out closed-set test split. Both excluded classes fall below the 8-sample setting reported in the SetFit literature, and in a domain of closely similar classes a split this small produces too few contrastive pairs for stable learning and too few test documents for meaningful evaluation. They are therefore treated as gaps in data collection, not as validated classes in the reported accuracy across 15 classes. The health-insurance class in the corpus was also split during this growth, into voluntary and supplementary subtypes; both are trained separately here but reported jointly as "health insurance" in Section 5.

Error analysis. We went through the remaining misclassifications and low-confidence cases in the second iteration by hand and found that many reflected label ambiguity in the source corpus rather than a straightforward model failure. Of the 9 compulsory-liability documents assigned to the CASCO class in the full re-evaluated pool, at least 4 were manually confirmed to contain both coverages on a single policy despite having only one ground-truth label; the remaining cases were genuine misclassifications. A few documents are in German rather than Croatian, including a carrier's liability insurance policy that the model classified as CASCO insurance. Three documents labeled as commission lists are actually commission-posting notices. This matches the point raised in Section 2.4: taxonomies built on a single label per page often do not map cleanly onto real document collections (Van Landeghem et al., 2024). It also fits recent audits showing that error rates for individual categories reach double digits, even in benchmarks as established as RVL-CDIP (Larson et al., 2026). The clear next step is to define a labeling rule for policies that combine multiple types of coverage, rather than keep tuning the model.

Fig. 2 qualitatively illustrates the learned embedding space. The resulting neighborhoods are visually consistent with a well-separated CoSENT-trained representation; several points between clusters correspond to mislabeled or multi-coverage documents discussed above.

5 Pipeline Evaluation

We evaluated the pipeline on a held-out validation set of 107 documents covering 14 document types, provided by a single one of the project's partner companies and annotated with expected document type and

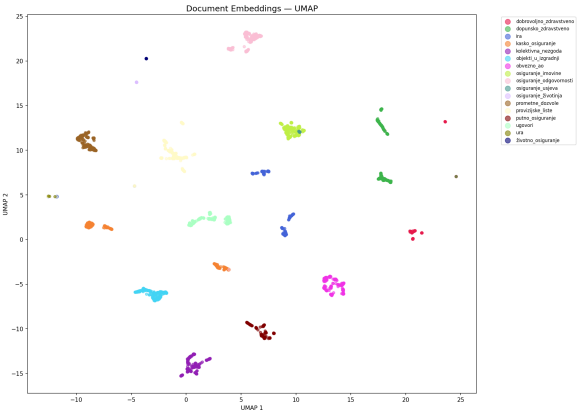


Figure 2. Exploratory UMAP projection produced with the second-iteration embedding model over the full labeled corpus, including classes excluded from training, colored by document class.

Table 3. Per-type structured-field extraction accuracy on the held-out validation set, with document counts, scored field counts, and 95% Wilson confidence intervals (CI).

Document type	n(doc)	n(field)	Struct. acc.	95% CI
Output invoice ledger (IRA)	10	40	100.00%	91–100%
Vehicle registration certificates	4	28	100.00%	88–100%
Compulsory motor liability	12	205	96.10%	92–98%
Health insurance	8	128	91.41%	85–95%
Group accident insurance	10	117	88.03%	81–93%
Liability insurance	8	112	86.61%	79–92%
Travel insurance	5	81	86.42%	77–92%
Property insurance	11	176	85.80%	80–90%
Life insurance	5	90	85.56%	77–91%
CASCO insurance	12	218	84.86%	80–89%
Contracts	4	16	81.25%	57–93%
Construction-risk insurance	12	125	76.00%	68–83%
Animal insurance	1	13	n/a[a]	—
Input invoice ledger (URA)	4	0	n/a[b]	—
Overall (scored rows)	101	1,336	**87.80%**	86–89%

[a]single document: too sparse for a reliable per-type rate
[b]undertrained: 3 documents, below SetFit's 8-sample minimum (Section 4)

expected field values per document. Although drawn from one partner company, the validation set includes documents from multiple insurers and brokers, so it tests unseen issuer-template variation but not full cross-partner generalization. The validation set covers 14 document types, an observed subset of the full classifier taxonomy. The first SetFit iteration used 17 historical labels, while the second trained on 15 after excluding crop insurance and the input invoice ledger because of data scarcity. The validation set did not include every trained label, and none of its 107 documents appeared in any training corpus. Classification was evaluated on all 107 documents; extraction was evaluated over the scored structured fields shown in Table 3. Expected field values were annotated by partner-company staff against the source documents; we did not measure inter-annotator agreement or perform double annotation on this set, which we note as a direction for more rigorous validation.

Classification. The system classified 101 of 107 documents correctly (94.4%), after correcting label-naming differences between the validation set and the production taxonomy. This compares with 98.87% on the internal held-out split from Section 4.

Extraction. Fields from the expected insights annotations were compared against extracted values using four matching criteria, ordered from strictest to loosest: (1) exact string match after Unicode normalisation; (2) substring containment; (3) numeric equivalence within 2% relative tolerance, for monetary amounts; (4) Jaccard token overlap above 0.65, for company names and addresses. Strict identifiers (OIB, policy numbers) required exact match.

Because rows without scored extraction fields were excluded, this is a scored-field extraction result, not a full-pipeline score over all documents. Across 1,336 scored structured fields, 87.80% matched (1,173/1,336); free-text description fields scored 60.87% (28/46), a result we treat as exploratory given the small sample rather than a reliable estimate. Table 3 shows per-type results, together with the number of validation documents of each type. Every scored type rests on a small sample, from 4 documents (contracts) to 12 (CASCO, compulsory motor liability, construction-risk), so per-type percentages are indicative rather than precise; Table 3 reports a 95% confidence interval alongside each rate. The overall structured-field score excludes rows marked n/a. The lowest-scoring type is construction-risk insurance (76.00%), where several fields require multi-part values, such as multiple insured parties or itemized sums insured. These are harder to match exactly and mirror the broader difficulty with multi-part liability limits discussed after Table 3. The next lowest reflects a narrower, single-cause failure: two free-text fields for contracts (wrong contracting party, a regrouped commission-rate table).

At the field level, personal-data and simple-amount fields (name, address, date of birth, total amount, policy number) matched all evaluated instances in this validation subset; this provides evidence that the LLM handles clearly defined, single-value fields well in the current sample. The hardest field was multi-part liability limits (25%), where ground truth and LLM group the same sub-amounts under different party labels. Covered risks (59%) and insured-party name (68%) were also difficult, the former because different insurers use different vocabulary for the same risk; these counts are similarly modest.

The input invoice ledger produced no predictions because it was excluded from training (too few documents available). None of the n/a rows count toward the overall structured-field accuracy.

The overall score is computed over the 101 scored validation documents; the 5 n/a-row documents are excluded. Being field-level, it does not yet capture document-level exact match, abstentions, or the full

cost of multi-value fields; a fuller end-to-end evaluation is a priority for future work.

6 Discussion and Future Work

The classification stage reaches 94.4% on truly held-out external documents, and the extraction stage achieves 87.80% structured-field accuracy over scored extraction rows. We report classification and scored-field extraction accuracy separately, not a single end-to-end field score across unscored documents. The main remaining failures are classification-related: 4 of 6 external misclassifications are input-invoice-ledger documents excluded from training, and the other 2 are plausible combined-coverage confusions; vehicle registration certificates had no external-set misclassifications. A related next step is defining a labeling convention for documents that span more than one class. Beyond that, collecting more data for underrepresented classes is the next step.

A second direction follows from the bounding-box point in Section 3. Vehicle registration certificates are the dataset's most visually complex class (dense tables, security-pattern backgrounds, stamps, phone-camera scans), so their OCR quality is more capture-sensitive than that of the mostly text-based document types. This class caused the most errors in the first classification iteration (Section 4), not because it is hard to learn, but because most scans yielded little usable text from the text-only OCR pipeline. Rerunning OCR with a vision-language model addressed most of that failure mode: per-class classification accuracy on the same re-evaluated pool rose from 49% (98/200) to 100% (200/200). We have not directly evaluated the reading stage itself against ground-truth transcriptions; the comparison between the VLM-based and Docling-based readers here is indirect, made through its effect on downstream classification accuracy, and a direct reading-stage evaluation remains open work. The broader lesson is that for visually dominant, low-text document types, the more robust fix may not be better text extraction, but using the VLM's own visual and layout understanding directly. Classification could use the page image and its grounded regions, rather than treating every document type as a text problem first and falling back to the image only when that fails.

The same VLM-based visual grounding, currently used only to detect non-textual regions for anonymization (Section 3), could be reused in the main pipeline to answer questions text alone cannot: whether a contract or policy is actually signed and where, whether an official stamp or seal appears on the document, or whether a payment QR code or barcode is present on an invoice, stored as structured metadata alongside the extracted text.

7 Conclusion

We presented AppAI, a hybrid pipeline for reading, classifying, and extracting structured attributes from Croatian bookkeeping and insurance documents, combining a vision-language reading stage, a few-shot SetFit classifier, and a self-hosted LLM extraction stage. Its main contribution is the applied integration of these existing techniques into a working pipeline for a concrete, previously unaddressed Croatian accounting and insurance document-processing setting, evaluated on genuinely held-out data. The classification stage is the most thoroughly evaluated part of the system: across two trained iterations, it reached over 98% accuracy on Croatian-specific closed-set test splits covering 15 to 17 classes, and the error analysis traced many remaining mistakes to OCR quality and source-label noise rather than to the classifier itself. On a single-partner held-out validation set (Section 5), the classification stage reached 94.4% accuracy on truly novel documents and the extraction stage reached 87.80% structured-field accuracy over scored validation fields. The reading stage already produces a layout and visual-grounding signal that classification and extraction do not yet use; making fuller use of it remains an open direction for future work.

Acknowledgments

This research was carried out under the project number IP.1.1.03.0138 within the Integrated Territorial Programme 2021–2027, co-financed by the Ministry of Regional Development and EU Funds.

References

Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., … Zhu, K. (2025). *Qwen3-VL Technical Report* (tech. rep.). Qwen Team. https://arxiv.org/abs/2511.21631

Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., … Lin, J. (2025). *Qwen2.5-VL Technical Report* (tech. rep.). Alibaba Group. https://arxiv.org/abs/2502.13923

Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., & Liu, Z. (2024). M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. *Findings of the Association for Computational Linguistics: ACL 2024*, 2318–2335.

Data Privacy Stack. (2024). Presidio: Data Protection and De-identification SDK. https://presidio.dataprivacystack.org/

Gao, T., Yao, X., & Chen, D. (2021). SimCSE: Simple Contrastive Learning of Sentence Embeddings. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Kim, G., Hong, T., Yim, M., Nam, J., Park, J., Yim, J., Hwang, W., Yun, S., Han, D., & Park, S. (2022). OCR-Free Document Understanding Transformer. *Proceedings of the European Conference on Computer Vision (ECCV)*, 498–517.

Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J. E., Zhang, H., & Stoica, I. (2023). Efficient Memory Management for Large Language Model Serving with PagedAttention. *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP)*. https://doi.org/10.1145/3600006.3613165

Larson, S., Lim, G., & Leach, K. (2023). On Evaluation of Document Classification with RVL-CDIP. *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Larson, S., Nagy, A., Desai, S., Desai, C., Lima, N. C., Yuan, Y., Betala, S., Prajapati, K. K., Suleiman, J. T., Duwal, S., & Leach, K. (2026). Revising RVL-CDIP: Quantifying Errors and Test-Train Overlap. *Proceedings of the ACM Symposium on Document Engineering (DocEng)*. https://arxiv.org/abs/2606.31446

Li, M., Lv, T., Chen, J., Cui, L., Lu, Y., Florencio, D., Zhang, C., Li, Z., & Wei, F. (2023). TrOCR: Transformer-Based Optical Character Recognition with Pre-Trained Models. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual Instruction Tuning. *Advances in Neural Information Processing Systems 36 (NeurIPS)*.

Livathinos, N., Auer, C., Lysak, M., Nassar, A., Dolfi, M., et al. (2025). Docling: An Efficient Open-Source Toolkit for AI-Driven Document Conversion. *AAAI-25 Workshop on Open-Source AI for Mainstream Use*.

Ljubešić, N., & Lauc, D. (2021). BERTić – The Transformer Language Model for Bosnian, Croatian, Montenegrin and Serbian. *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*.

OpenAI. (2023). GPT-4V(ision) System Card [System card]. https://openai.com/index/gpt-4v-system-card/

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., … Zoph, B. (2023). *GPT-4 Technical Report* (tech. rep.). OpenAI. https://arxiv.org/abs/2303.08774

Qwen Team. (2025). Qwen3-14B Model Card. https://huggingface.co/Qwen/Qwen3-14B

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992. https://doi.org/10.18653/v1/D19-1410

Scius-Bertrand, A., Jungo, M., Vögtlin, L., Spat, J.-M., & Fischer, A. (2024). Zero-Shot Prompting and Few-Shot Fine-Tuning: Revisiting Document Image Classification Using Large Language Models. *Proceedings of the 27th International Conference on Pattern Recognition (ICPR)*.

Smith, R. (2007). An Overview of the Tesseract OCR Engine. *Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR)*.

Štefanec, V., Farkaš, D., Thakkar, G., & Tadić, M. (2024). Building a Large Language Model for Croatian. *Proceedings of the International Conference New Trends in Translation and Technology (NeTTT)*.

Tunstall, L., Reimers, N., Jo, U. E. S., Bates, L., Korat, D., Wasserblat, M., & Pereg, O. (2022). Efficient Few-Shot Learning Without Prompts. *Proceedings of the NeurIPS 2022 Workshop on Efficient Natural Language and Speech Processing (ENLSP)*.

Van Landeghem, J., Biswas, S., Blaschko, M., & Moens, M.-F. (2024). Beyond Document Page Classification: Design, Datasets, and Challenges. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2962–2972. https://openaccess.thecvf.com/content/WACV2024/html/Van_Landeghem_Beyond_Document_Page_Classification_Design_Datasets_and_Challenges_WACV_2024_paper.html

Wang, D., Raman, N., Sibue, M., Ma, Z., Babkin, P., Kaur, S., Pei, Y., Nourbakhsh, A., & Liu, X. (2024). DocLLM: A Layout-Aware Generative Language Model for Multimodal Document Understanding. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8529–8548.

Learning Analytics, AI and Technologies in Education

Sam Sabah, Alexander Tillmann, David Weiß, Hendrik Drachsler
When the Map Becomes a Place: Immersive Virtual RealityTerrain Exploration with GeoVR

Sirje Virkus, Muhammad Shaheer Alam
Knowing but Not Doing: The Awareness-Practice Gap in Responsible and Green AI Use

Jasenka Begić, Tihana Banko, Ana Lokas Ćošković
Perceptual Indicators for Teaching Quality: Transforming SET Data into Learning Analytics Tools for Adult Online Instructors

Branislav Fecko, Jozef Dziak, Tibor Vince, Ján Molnár
Conversational AI for Electrical Measurement Education: An Open-Source LLM Framework

Nikola Kadoić, Blaženka Divjak, Barbi Svetec, Thomas Penniston, Mirza Žižak, Dino Vlahek
Review of Learning Outcomes Prioritisation Approaches

Sabina Rako
Students' Perceived Usefulness of Prompts in Blended Learning: Two Pilot Studies

Goran Hajdin
Pedagogical Friction in ICT-Integrated K-12 Classrooms: Cognitive Offloading, Technostress, and Shallow

Neven Pintarić, Ivan Peraić
Evaluation of Student Assignments Using Local Large Language Models and Graph-Based Evidence

Vjeran Bušelić, Dinko Horvat
Schema Before Output: Metacognitive Schema-First Scaffolding in Computing Higher Education in the GenAI Era

Vjeran Bušelić, Dinko Horvat
Skills Never Formed: AI Adoption and Career Formation Risk in IT Higher Education

Michael Eichhorn, Julia Schmitt, Alexander Tillmann, Josmario Albuquerque, Paul Astles, Catriona Matthews, Bart Rienties, Blazenka Divjak, Tanja Šestanj-Perić, Egle Gedrimiene, Hanni Muukkonen, Mirza Žižak
Exploring Authentic Learning Scenarios in Online and Blended Environments Across European Universities

Hanni Muukkonen, Egle Gedrimiene, Blaženka Divjak, Barbi Svetec, Julia Schmitt, Bart Rienties, Michael Eichhorn, Alexander Tillmann
Synergies of learning design and analytics: Design patterns of higher education courses

When the Map Becomes a Place: Immersive Virtual Reality Terrain Exploration with GeoVR

Sam Sabah, Alexander Tillmann, David Weiß
Goethe University Frankfurt and studiumdigitale
Eschersheimer Landstr. 155/157, 60323 Frankfurt am Main
{sabah, tillmann, weiss}@sd.uni-frankfurt.de

Hendrik Drachsler
Goethe University Frankfurt, studiumdigitale and DIPF
Eschersheimer Landstr. 155/157, 60323 Frankfurt am Main
h.drachsler@dipf.de

Abstract. *GeoVR is a Virtual Reality application for geographical and geomorphological education, built as a Learning Space within the fuelsME ecosystem. Users explore Digital-Elevation-Model-based terrain, switching between an immersive ground-level view and a scalable sandbox overview, supported by an analytical toolset (height indicator, compass, profiler, ruler) and a terrain layer configurable for any region without code changes. Guided by scale fluidity, analytical concreteness, and collaborative inquiry, we argue that dual-mode navigation supports scale-fluid geomorphological reasoning, and that layered VR architectures let domain-specific applications focus on differentiating affordances rather than generic infrastructure. We report deployment observations and a planned evaluation agenda.*

Keywords. Virtual Reality, Geography, Geomorphology, Educational Applications, Terrain Exploration, fuelsME, Field Trip Simulation, Multi-User VR

1 Introduction

Geography and the geosciences share a foundational pedagogical challenge: the phenomena students need to understand are written into landscapes distributed across the entire surface of the earth. Glacial cirques, river deltas, fault scarps, and volcanic calderas are the primary sources for a discipline whose core concepts are spatial, scalar, and temporal. Physical field trips are regarded as an important element in promoting geographical learning processes (Fuller et al., 2000, 2006).

However, even during physical field trips, it is often difficult to select key landscape features from the wealth of impressions and to grasp their spatial location. A major challenge is the identification of forms and their temporal and spatial changes, which in turn allow conclusions to be drawn about the dynamics of processes, such as the reconstruction of former valley floors and river courses, which are of fundamental importance for understanding geomorphodynamic systems such as the development of runoff regimes and flood development (Tillmann and Wunderlich, 2023). The varying scales and timeframes involved in observing these phenomena, as well as the translation of conventional two-dimensional representations, very quickly exceed learners' visual-spatial imagination (Kastens and Manduca, 2012). Photographs flatten depth, topographic maps demand cartographic literacy, and desktop GIS software places learners outside the data they are trying to understand.

Virtual Reality (VR) offers a fundamentally different relationship: the learner stands inside a three-dimensional landscape that responds to their movement and invites exploration. Recent evidence shows that well-designed VR learning environments match traditional instruction on academic outcomes while substantially raising engagement and motivation (Allcoat and von Mühlenen, 2018; Wright et al., 2023). More importantly, VR enables experiences fieldwork cannot: continuous scale change from satellite altitude to ground level within seconds, or direct superposition of terrain states from different geological periods.

This paper presents GeoVR, a VR application for geographical education that realises these possibilities through Digital Elevation Model (DEM)-based terrain rendering, dual-mode spatial navigation, and an interactive analytical toolset. GeoVR is one Learning Space within a larger VR ecosystem built under the fuelsME:Create framework (Mascher et al., 2024), which provides shared multi-user infrastructure and a collaborative co-

design process for technology and didactics. This paper treats that framework as given and focuses on what GeoVR contributes on top of it.

1.1 Contributions

This paper makes three contributions:

1. A design case study of GeoVR, documenting the pedagogical rationale, terrain pipeline, and dual-mode navigation design (micro/macro) of a VR learning environment for geomorphology.

2. An architectural argument for layered, reusable VR infrastructure in education: by inheriting shared multi-user capabilities rather than rebuilding them, domain-specific applications can direct most of their development effort toward the pedagogical affordances that differentiate the learning experience.

3. A preliminary deployment report and an evaluation plan intended for pre-registration, drawing on informal observations from early seminar use to surface emerging interaction patterns and to specify the controlled studies that will test GeoVR's learning claims.

GeoVR is presented here as a single instance of how a domain-specific VR learning environment can be built on shared infrastructure. Whether the same approach transfers to adjacent disciplines such as geology, environmental science, or urban planning is an open question that this paper does not attempt to settle.

2 The fuelsME:Create Context

GeoVR is one of several Learning Spaces (*fuelsME:Lernräume*) within the fuelsME:App, a VR application developed under the fuelsME:Create collaborative framework (Mascher et al., 2024) by a research consortium of several higher-education institutions.

The fuelsME:Create framework is relevant to this paper in two ways. First, it provides the organisational process within which GeoVR was designed: an interdisciplinary agile methodology in which educators act as Product Owners responsible for backlog prioritisation, ensuring that pedagogical requirements drive technical decisions rather than the reverse (Sabah et al., 2025). Second, it provides the shared technical infrastructure, fuelsME:Core, on which GeoVR is built.

2.1 fuelsME:Core as a Shared Infrastructure Layer

fuelsME:Core is a Unity-based multi-user VR runtime that provides the foundational capabilities required by any Learning Space in the fuelsME ecosystem: real-time network synchronisation across connected Head-Mounted Displays (HMDs), a configurable avatar system, reusable interaction primitives, and a token-based session isolation mechanism that allows multiple concurrent research groups to share server infrastructure without interfering with one another (Mascher et al., 2024).

The architectural consequence for GeoVR is significant: the application does not implement any of these capabilities. It inherits them. Concretely, the GeoVR codebase contains no networking, avatar, or session-management code; its implementation effort was spent almost entirely on the four domain-specific capabilities that differentiate geographical education (the terrain pipeline, the two view modes, the analytical toolset, and terrain-scale navigation). In the consortium's product backlog, all GeoVR-specific items addressed these pedagogical features; infrastructure items were handled once, at the fuelsME:Core level, and benefited every Learning Space simultaneously. The reuse also runs in the other direction: components developed for GeoVR, in particular the DEM-based heightmap import pipeline and the terrain-scaled teleportation, are candidates for adoption by other spatially oriented Learning Spaces in the ecosystem. Table 1 makes this division explicit.

Table 1. Division of responsibility between fuelsME:Core and GeoVR

Capability	fuelsME:Core (foundation)	GeoVR extension
Multi-user networking	Provided (built-in)	Inherited / no changes needed
Avatar system	Provided (built-in)	Inherited / no changes needed
Session isolation	Token-based rooms	Inherited / no changes needed
Terrain rendering	Not provided	DEM-based 3D terrain pipeline
View modes	Not provided	Micro (immersive) / Macro (sandbox)
Analytical tools	Not provided	Height indicator, compass, elevation profiler, height ruler
Spatial navigation	Basic locomotion	Teleportation tuned for terrain scale

2.2 The Learning Space

Within the fuelsME:Create framework, a distinction is drawn between a Learning Space (the technical and visual environment, authored once) and a Scenario (a specific deployment of that space for a defined course, group, and set of learning objectives)(Mascher et al., 2024). GeoVR is a Learning Space. It has been deployed in several distinct Scenarios: one focusing on glacial landform identification for a physical geography seminar, another centred on river morphology comparison for an environmental studies module (Sabah et al., 2025).

This distinction matters for the interpretation of the design decisions described in Section 3. Every design choice in GeoVR was made at the Learning Space level: it must serve the full range of scenarios for which the space will be used, not just the first one. This has pushed the design toward generality and configurability: terrain can be swapped by loading different DEM data, and tasks can be changed by reconfiguring scenario-level prompts, rather than toward narrow optimisation for a single course.

3 GeoVR: Design and Implementation

3.1 Pedagogical Design Principles

The design of GeoVR was governed by three pedagogical principles, developed iteratively through the interdisciplinary process described in fuelsME:Create (Mascher et al., 2024; Sabah et al., 2025) and informed by the geoscience education literature. Immersive virtual reality can enhance experiential understanding but may impose additional cognitive load if not carefully designed (Parong and Mayer, 2018; Mayer, 2002), and the three principles described below reflect deliberate choices made by the design team to manage this trade-off between immersion and cognitive demand.

The first principle is *scale fluidity*. Geomorphological reasoning requires the ability to move between spatial scales: from the regional overview that reveals drainage patterns to the local detail that reveals erosion mechanisms. A fixed perspective, whether map, photograph, or conventional three-dimensional model, privileges one scale at the expense of others. GeoVR's design makes scale switching a primary interaction rather than an afterthought (Lawson and Mayer, 2025).

The second principle is *analytical concreteness*. Students often struggle to connect the abstract conventions of topographic maps (contour lines, colour-coded elevation bands, gradient arrows) to felt spatial experience. GeoVR's toolset is designed to bridge this gap by overlaying quantitative data (elevation values, bearing angles, profile curves) directly onto the immersive terrain, so that students can read a number and simultaneously see and feel what it means.

The third principle is *collaborative inquiry*. Geography is fundamentally a discipline of shared observation and contested interpretation. GeoVR's multi-user design is not a convenience feature: it is a deliberate response to the evidence that shared spatial attention in VR supports richer discussion and deeper conceptual development than solo exploration (van der Meer et al., 2023). This principle is realised in GeoVR through three concrete mechanisms. First, the co-present avatars inherited from fuelsME:Core make each learner's position, orientation, and pointing gestures visible within the terrain, so that attention can be directed deictically ("this ridge here") rather than only verbally. Second, the two view modes support shared attention in complementary ways: in macro view, learners stand around the sandbox model as they would around a physical relief model and can jointly inspect the same feature from different angles, while in micro view, co-located exploration lets learners experience the same landform from within, at their own vantage points. Third, the analytical tools produce quantitative readings that give learners a shared vocabulary for coordinating attention across distance: a learner can announce their elevation from the height indicator and thereby anchor a joint reference point in the terrain. Section 4 reports an emergent instance of this last mechanism observed in early seminar use.

3.2 Terrain Pipeline

The technical foundation of GeoVR is a terrain pipeline that converts publicly available Digital Elevation Model (DEM) data into interactive Unity Terrain objects. The pipeline proceeds in three stages.

3.2.1 Data Acquisition and Preprocessing

GeoVR supports DEM sources that provide raster-format elevation data at sufficient resolution for educational use, including the Shuttle Radar Topography Mission (SRTM) dataset (approximately 30-metre horizontal resolution globally) and the Copernicus Digital Elevation Model (approximately 10-metre resolution for Europe). Raw DEM files are preprocessed using standard GIS tooling to clip the region of interest, fill data voids, and resample to the resolution appropriate for the target Unity Terrain size.

The choice of source dataset involves a trade-off between global coverage, resolution, and licence

compatibility. SRTM data is in the public domain and covers most of the earth's land surface, making it the default choice for curriculum flexibility. Copernicus DEM data offers significantly higher resolution for European locations, which is relevant for scenarios focused on Alpine glaciology or river morphology in densely studied European catchments.

3.2.2 Terrain Generation in Unity

Processed DEM data is imported into Unity as a heightmap and applied to a Unity Terrain object. The Terrain object's resolution parameters (heightmap resolution, detail resolution, and base map resolution) are tuned to balance visual fidelity against performance on the Meta Quest 2's mobile GPU. In practice, a heightmap resolution of 513×513 with detail and base map resolutions of 1024×1024 provides adequate detail for educational use without exceeding the Quest 2's memory budget.

Texturing is applied procedurally using Unity's Terrain layer system, with texture weights computed from elevation and slope at each surface point. This produces naturalistic colouration (green valley floors shading to grey rockfaces and white snowfields at altitude) without requiring manual authoring per terrain tile and without depending on satellite imagery that may be cloud-covered or unavailable for a given region.

3.2.3 Terrain Reuse and Configuration

Because terrain generation is separated from scenario configuration, the same GeoVR Learning Space can be used with any DEM dataset without code changes. Switching to a new geographic location requires only loading a different heightmap asset and adjusting the scale parameters that convert raw elevation units to Unity world-space metres. This configurability is the mechanism that makes GeoVR viable for the full range of scenarios in the curriculum rather than just the scenario it was first built for.

3.3 Dual-Mode Spatial Navigation

GeoVR offers two complementary view modes that give users fundamentally different analytical relationships to the terrain. The design of these modes is the central pedagogical affordance of the application and the aspect of the design that most directly required coordination between the educational geographers and the software engineers in the fuelsME:Create process (Sabah et al., 2025).

3.3.1 Micro View: Immersive Field Mode

In micro view, the user is scaled to human height and placed within the terrain. Navigation is achieved through short-distance teleportation: the user points a controller at a target location on the terrain surface and triggers a near-instantaneous transition, avoiding the nausea that continuous locomotion can induce on mobile VR hardware. The teleportation arc is visually rendered as a parabolic cursor that snaps to the terrain surface, giving users precise control over their destination (Fig. 1).

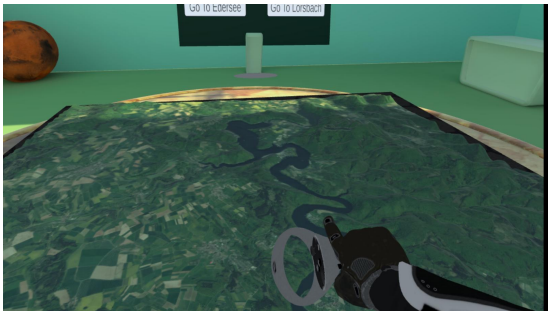


Figure 1. Micro view: the user is scaled to human height and placed inside the terrain. Navigation uses short-distance teleportation, visualised here as a parabolic cursor snapping to the terrain surface.

Micro view is designed for tasks that require felt spatial experience: identifying the cross-sectional shape of a glacial valley, appreciating the gradient of a river's long profile, or observing the relative heights of a terrace sequence from within the floodplain. These are tasks that no map or diagram can support in the same way, because they depend on the learner's proprioceptive sense of their own position within a three-dimensional space.

3.3.2 Macro View: Sandbox Mode

In macro view, the terrain is rendered at a reduced scale as a physical model on a horizontal surface at approximately waist height, analogous to a sandbox relief model in a geography classroom (Fig. 2). The user can walk around the model, lean over it, and view it from any angle. This mode supports tasks that require a synoptic perspective: identifying drainage divides, tracing the paths of major rivers from source to mouth, or comparing the morphology of adjacent valleys.

The distinction between micro and macro view is not merely cosmetic. In micro view, the user is inside the data; in macro view, the user is outside it looking in. Both perspectives are necessary for complete geomorphological understanding, and the ability to switch between them within a single session, without leaving the virtual environment,

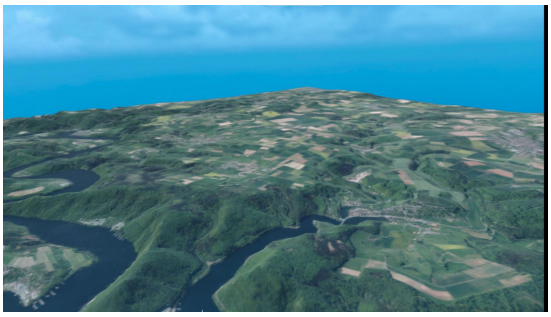


Figure 2. Macro view: the terrain is rendered at reduced scale as a physical model at waist height, analogous to a sandbox relief model. The user can walk around the model and view it from any angle.

is a capability that has no analogue in traditional fieldwork or GIS software.

3.4 Interactive Analytical Toolset

GeoVR provides a suite of interactive tools that extend the scope of analysis available within the virtual terrain. All tools are accessed via the Meta Quest 2 controller buttons and mapped to the interaction framework provided by fuelsME:Core, ensuring consistency with the other Learning Spaces in the fuelsME:App (Fig. 3).

- **Height indicator:** Displays the absolute elevation of the user's current position in real time. In micro view, this grounds the user's felt experience in quantitative data; in macro view, it allows point-to-point elevation comparison. The indicator updates continuously as the user moves, supporting the development of intuition about elevation gradients.
- **Compass:** Displays magnetic bearing as an overlay, supporting orientation exercises that require students to connect cardinal directions to terrain features. The compass is particularly useful in scenarios where students must read a topographic map alongside the VR environment and locate specific features by bearing.
- **Elevation profiler:** Allows the user to draw a transect line across the terrain surface using the controller trigger. The application computes the elevation profile along the transect and renders it as a two-dimensional graph overlaid on the terrain. This tool directly bridges the gap between immersive experience and cartographic representation: students can select two points they have physically visited in micro view and immediately see the quantitative cross-section connecting them (Sabah et al., 2025).
- **Height ruler:** A measuring tool that reports the real-world height of a selected terrain feature, expressed in actual metres above sea level rather than in-VR units. Whereas the height indicator reports the user's own elevation, the height ruler lets students measure features they observe, such as the relief of a valley wall, the prominence of a peak, or the depth of a gorge, and develop accurate intuitions about the real-world scale of what they are seeing.

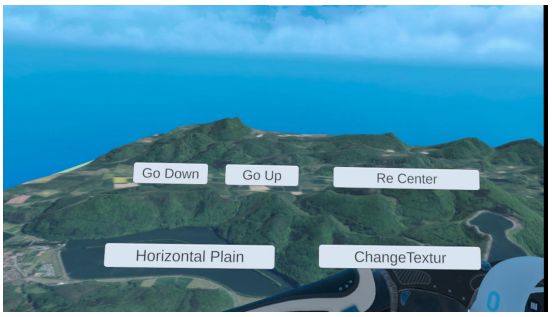


Figure 3. A subset of GeoVR's interactive tools, accessed through a menu attached to the user's virtual arm, keeping controls within easy reach during exploration.

4 Pedagogical Deployment and Early Experience

4.1 Multi-User Session Management

GeoVR is deployed as part of the fuelsME:App on Meta Quest 2 HMDs managed through Meta's App Lab distribution platform, which the fuelsME:Create framework uses to maintain separate release channels for development, educator, and public demo use (Mascher et al., 2024). This infrastructure is fully shared with the other Learning Spaces in the fuelsME ecosystem and requires no GeoVR-specific configuration.

Session isolation, the mechanism that prevents two concurrent seminar groups from sharing the same virtual room, is handled by the token system described in fuelsME:Create (Mascher et al., 2024). When a geography course session is registered, a unique token is generated and distributed to the enrolled students. Only HMDs presenting the same token join the same GeoVR session. This ensures that two courses using GeoVR simultaneously on the same server infrastructure cannot observe or disturb each other, which is essential for both the pedagogical experience and the integrity of any concurrent research data collection.

4.2 Observed Interaction Patterns

Early seminar deployments of GeoVR at *Anonymized for review* have allowed informal observation of how students interact with the application in practice. The first cohort consisted of twelve first-year geography students with high self-reported computer skills (M=4.25, SD=0.97 on a 5-point scale) and mixed prior VR familiarity (M=3.42, SD=1.51), with three students reporting little or no prior VR experience and six reporting high familiarity. We emphasise the methodological status of what follows: these are anecdotal impressions gathered by the attending educators during regular seminar sessions, without a systematic observation protocol, coding scheme, or formal analysis procedure. Given the small cohort and the informal setting, the patterns reported below should be read as formative feedback that motivates the evaluation programme in Section 5, not as empirical findings. With that caveat, several patterns are noteworthy.

Students in this cohort used the scale-switching affordance more than anticipated. Rather than spending extended periods in one mode before switching, they tended to alternate between micro and macro view as they tested ideas, forming an interpretation in macro view about the shape of a valley, descending into micro view to check it from the inside, and returning to macro to revise. This iterative pattern resembles the kind of inference cycle characteristic of geomorphological fieldwork. Whether it in fact reflects authentic disciplinary reasoning, rather than, for example, novelty-driven exploration of the mode-switching mechanic itself, cannot be determined from informal observation. We therefore treat it as a hypothesis: the planned evaluation (Section 5) will test it by correlating logged view-mode transitions with think-aloud data.

The collaborative dimension of GeoVR generated spontaneous interactions that the educators had not planned for. In one session, students began using the height indicator not as an individual measurement tool but as a way of coordinating joint attention: "I'm at 1,847 metres, can you see me from there?" This single episode is consistent with the form of collaborative sense-making that the literature on multi-user VR identifies as a distinctive affordance of shared virtual environments (van der Meer et al., 2023), and it illustrates the third design principle (Section 3) in use; whether such tool-mediated joint attention occurs systematically is a question for the planned evaluation.

The elevation profiler was the tool with the steepest learning curve in this cohort. Students required explicit instruction in how to interpret the profile graph, which is expected: the tool makes demands of cartographic literacy that not all first-year students possess. This observation has informed the design of the accompanying scenario materials, which now include a guided introduction to cross-section interpretation before students encounter the profiler in open exploration.

4.3 Limitations Encountered

Several practical limitations have emerged from early deployments. The Meta Quest 2's processing constraints impose a cap on terrain resolution that is occasionally visible as faceted geometry on steep slopes. This artefact does not impair the educational function of the application for most scenarios, but it reduces the credibility of the representation for students who are familiar with the actual terrain being modelled. Higher-resolution rendering on more capable hardware is a priority for future development.

The token distribution process, though straightforward from a technical standpoint, adds an administrative step for educators that is unfamiliar and has caused confusion in several sessions. A self-service booking interface that automates token generation and distributes it via the university's learning management system is under development.

Finally, GeoVR currently supports only static terrain: the temporal change tool described in the original feature specification has not yet been implemented. Dynamic comparison of terrain states across geological time periods remains a high-priority item in the product backlog and is expected to be completed in the next development cycle.

5 Research Agenda for Empirical Evaluation

GeoVR has been evaluated informally through observation and educator feedback, but a rigorous empirical evaluation programme has not yet been conducted. This section outlines the research questions that the consortium intends to pursue and the methodological approaches planned for each. The studies described here have not yet been pre-registered; the consortium intends to pre-register the study designs, hypotheses, and analysis plans on a public registry before data collection begins.

The primary research question concerns learning outcomes: does GeoVR-mediated terrain exploration produce measurably superior spatial cognition and geomorphological reasoning relative to traditional map-and-photograph instruction? Addressing this question requires a controlled study design comparing GeoVR and non-VR conditions on pre- and post-tests of spatial reasoning and dis-

ciplinary knowledge. (Gielstra et al., 2024) provide a useful methodological precedent for this kind of evaluation in a VR geoscience context.

A secondary research question concerns the specific contribution of the collaborative, multi-user design: does joint exploration of virtual terrain produce different learning outcomes than solo exploration? The literature on collaborative VR (van der Meer et al., 2023) suggests it should, but empirical evidence from geoscience contexts is sparse. GeoVR's architecture, which allows the multi-user feature to be selectively enabled or disabled at the scenario level, makes it unusually well-suited for an experiment that isolates this variable. A third research question concerns long-term retention: are spatial understanding gains from VR-based learning maintained over weeks and months, or do they decay at the same rate as gains from conventional instruction? (Allcoat and von Mühlenen, 2018) note that most VR learning studies measure immediate post-test outcomes; retention data is a gap in the literature that GeoVR evaluations will be designed to address.

Methodologically, the evaluations will combine quantitative outcome measures with qualitative data from think-aloud protocols during GeoVR sessions, post-session interviews, and analysis of in-session interaction logs. The log data, which records each user's position, tool activations, and view-mode transitions throughout a session, is a distinctive advantage of VR over traditional fieldwork: it provides a continuous, objective record of how students engaged with the terrain, which can be analysed for patterns that correlate with learning outcomes. This log analysis will also test the hypothesis raised by the informal observations in Section 4: if rapid alternation between micro and macro view reflects disciplinary inference rather than novelty-driven exploration, mode transitions should co-occur with interpretive talk in the think-aloud data and should persist beyond students' initial familiarisation with the application.

6 Conclusion

This paper has presented GeoVR as a design case study of a pedagogically grounded VR Learning Space built on the fuelsME:Create framework (Mascher et al., 2024). The framework provided both the organisational process for co-designing technology and didactics and the shared fuelsME:Core infrastructure for multi-user VR. Within that scaffolding, GeoVR's development could be directed at affordances specific to geographical education rather than at the generic infrastructure any multi-user VR application requires.

What this paper demonstrates are GeoVR's specific contributions: a DEM-based terrain pipeline that supports any geographic region without code changes, a dual-mode spatial navigation design that pairs immersive field experience with synoptic overview, and an interactive toolset that bridges immersive observation and cartographic representation. Its effects on learning outcomes, long-term retention, and collaborative learning remain open questions. The informal seminar observations are consistent with authentic geomorphological reasoning and have already shaped the accompanying scaffolding materials, but establishing these effects requires controlled studies. The next step is therefore a pre-registered evaluation combining outcome measures, think-aloud protocols, and in-session interaction logs.

The broader implication of this work is architectural as much as pedagogical. Building GeoVR on shared infrastructure rather than from scratch allowed development effort to remain focused on the discipline-specific affordances that differentiate the learning experience. Whether this layered approach scales to other domains and other institutions remains an open question, and one that will be tested as fuelsME:Core moves toward public release as an open foundation for use beyond the original consortium.

References

Allcoat, D. and von Mühlenen, A. (2018). Learning in virtual reality: Effects on performance, emotion and engagement. *Research in Learning Technology*, 26.

Fuller, I., Edmondson, S., France, D., Higgitt, D., and Ratinen, I. (2006). International perspectives on the effectiveness of geography fieldwork for learning. *Journal of Geography in Higher Education*, 30(1):89–101.

Fuller, I., Rawlinson, S., and Bevan, R. (2000). Evaluation of student learning experiences in physical geography fieldwork: Paddling or pedagogy? *Journal of Geography in Higher Education*, 24(2):199–215.

Gielstra, D., Moorman, L., Kelly, J., Schulze, U., Resler, L. M., Cerveny, N. V., Gielstra, J., Bryant, A., Ramsey, S., and Butler, D. R. (2024). Designing virtual pathways for exploring glacial landscapes of Glacier National Park, Montana, USA for physical geography education. *Education Sciences*, 14(3):272.

Kastens, K. A. and Manduca, C. A. (2012). *Earth and mind II: A synthesis of research on thinking and learning in the geosciences*, volume 486. Geological Society of America.

Lawson, A. P. and Mayer, R. E. (2025). Effect of pre-training and role of working memory characteristics in learning with immersive vir-

tual reality. *International Journal of Human–Computer Interaction*, 41(4):2523–2540.

Mascher, U., Weiß, D., Fuchs, A., Appel, S., Fernes, D., and Sabah, S. (2024). fuelsme: Create-ein framework zur kollaborativen entwicklung einer vr-anwendung fur die hochschullehre. In *Proceedings of DELFI Workshops 2024*. Gesellschaft für Informatik eV.

Mayer, R. E. (2002). Multimedia learning. In *Psychology of learning and motivation*, volume 41, pages 85–139. Elsevier.

Parong, J. and Mayer, R. E. (2018). Learning science in immersive virtual reality. *Journal of educational psychology*, 110(6):785.

Sabah, S., Tillmann, A., Schneider, J., and Drachsler, H. (2025). Enhancing user onboarding in virtual reality educational applications: Evaluating the effectiveness of pre-training user onboarding method. In *CSEDU (1)*, pages 364–371.

Tillmann, A. and Wunderlich, J. (2023). Forschendes lernen–vor ort und virtuell. virtual reality (vr)-anwendungen im kontext von exkursionen am beispiel der mittelgebirgslandschaft kellerwald/edersee. *GW Unterricht*, 171:54–65.

van der Meer, N., van der Werf, V., Brinkman, W.-P., and Specht, M. (2023). Virtual reality and collaborative learning: A systematic literature review. *Frontiers in Virtual Reality*, 4.

Wright, P. N., Whitworth, M., Tibaldi, A., Bonali, F., Nomikou, P., Antoniou, V., Vitello, F., Becciani, U., Krokos, M., and Van Wyk de Vries, B. (2023). Student evaluations of using virtual reality to investigate natural hazard field sites. *Journal of Geography in Higher Education*, 47(2):311–329.

Knowing but Not Doing: The Awareness-Practice Gap in Responsible and Green AI Use

Sirje Virkus, Muhammad Shaheer Alam
School of Digital Technology
Tallinn University
Narva mnt 25, 10120 Tallinn, Estonia
{sirje.virkus, shaheer}@tlu.ee

Abstract. *Generative artificial intelligence is now routine in university work, but responsible use remains inconsistent. This paper examines how students in AI-related programmes at Tallinn University understand and practise Green AI and Responsible AI. In this exploratory study of 57 students and six interview participants from two AI-related programmes at Tallinn University, participants generally recognised ethical, privacy and environmental concerns, but reported that efficiency and convenience sometimes took priority over verification, transparency, privacy protection and sustainable use. The findings suggest that universities may need to move beyond awareness raising by embedding responsible AI routines, institutional guidance and supportive tool design into academic practice.*

Keywords. Green AI, Responsible AI, Generative AI, higher education, student behaviour, awareness-practice gap, AI ethics, sustainability

1 Introduction

Generative artificial intelligence (GenAI) tools such as ChatGPT, virtual assistants and writing support tools are now embedded in the ordinary workflow of many university students. They are used to summarise readings, draft text, generate ideas, explain difficult concepts and accelerate repetitive academic tasks (Athanassopoulos et al., 2026; Deng et al., 2025; Xiao et al., 2025). This uptake has produced an important pedagogical tension. On one side, artificial intelligence (AI) tools can improve access, productivity and confidence. On the other side, they introduce risks for academic integrity, privacy, intellectual dependence, bias, accountability and environmental sustainability (Francis et al., 2025).

This paper is based on the study that examined this tension through the combined lens of Green AI and Responsible AI. Green AI refers to the development and use of AI technologies that minimise environmental impact by reducing energy consumption and carbon emissions (Tabbakh et al., 2024). Responsible AI addresses the ethical and societal conditions under which AI should be designed and used, including fairness, accountability, transparency, privacy and human oversight (Raman et al., 2024). In higher education these two areas intersect because student behaviour is not merely a private matter. Repeated daily prompting, uncritical reliance on outputs and casual sharing of data can accumulate into broader institutional and societal consequences.

A central problem is that knowledge of AI risks does not automatically produce responsible action. Students may know that AI outputs should be checked, that private data should not be uploaded, or that unnecessary use has environmental costs, while still bypassing these practices when under pressure. The empirical question is therefore not only whether students have heard of Green AI or Responsible AI, but how such awareness appears in actual academic routines and what conditions enable or inhibit responsible behaviour.

This paper focuses on four research questions: 1) What level of awareness do students report regarding Green AI and Responsible AI? 2) How do students report applying Green AI and Responsible AI principles in academic work? 3) What gaps emerge between students' awareness and reported practice? and 4) What contextual factors influence responsible AI behaviour?

This paper contributes to higher education AI research in three ways: 1) It connects Responsible AI and Green AI through the concept of routine academic practice. 2) It provides exploratory evidence of an awareness-practice gap among students in two AI-related programmes at one university. 3) It identifies contextual factors that participants associated with this gap, including time pressure, institutional ambiguity and tool design. The paper contributes to discussions on AI literacy by arguing that literacy must include situated behavioural competence, not only conceptual familiarity.

2 Related Work

The rapid development of GenAI in higher education has generated substantial interest in its pedagogical

opportunities, institutional implications and ethical risks. Data from the Web of Science underscores this trend: between 1981 and May 2026, 35,815 publications were indexed under 'artificial intelligence and education.' The higher education sector mirrors this growth, contributing 6,323 publications to the field.

Recent studies show that students use GenAI tools for a wide range of academic purposes, including writing support, summarisation, programming assistance, idea generation and personalised explanation (Athanassopoulos et al., 2026; Deng et al., 2025; Xiao et al., 2025). These uses can support productivity, accessibility and confidence, particularly when students face complex tasks or time constraints. At the same time, the literature identifies significant concerns, including plagiarism, over-reliance, reduced critical engagement, inaccurate or fabricated outputs, unequal access and uncertainty about acceptable academic practice (Kasneci et al., 2023; Dwivedi et al., 2023). A recurring issue is that student adoption often develops faster than institutional policy and classroom guidance, leaving students to interpret expectations through peers, lecturers and tool affordances (Nikolic et al., 2024; Stracke et al., 2026).

Responsible AI provides an important framework for addressing these challenges. Foundational AI ethics frameworks emphasise principles such as fairness, accountability, transparency, privacy, explicability, human oversight and non-maleficence (European Commission, 2019; Floridi et al., 2018; Floridi & Cowls, 2019; Jobin et al., 2019; OECD, 2026). In higher education, these principles become relevant not only at the level of institutional governance but also in students' everyday academic practices. Responsible student use of AI may involve checking AI-generated claims, recognising possible bias, avoiding the upload of sensitive data, disclosing AI assistance where required, and retaining human judgement and authorship over academic work (European Commission, 2024; Wang et al., 2024). However, a persistent challenge is operationalisation. Ethical principles are often easier to endorse in general terms than to apply consistently in concrete, time-pressured academic situations (Mittelstadt, 2019; Morley et al., 2023).

Green AI adds a further dimension to responsible AI use by foregrounding the environmental costs of AI systems. Green AI research has moved from a model-centred concern with reducing the computational cost of training toward a broader lifecycle perspective. Early work challenged accuracy-at-any-cost AI and promoted reporting energy and compute use. Subsequent research developed optimisation methods such as pruning, quantisation, compression and efficient algorithms. More recent work expands the field to hardware, data centres, renewable energy and system-level design, while also shifting attention from training to inference and large-scale real-world use. This evolution shows that Green AI is no longer only a technical efficiency agenda, but an emerging sociotechnical concern involving developers, infrastructure providers, institutions and users (Tabbakh et al., 2024). This broader lifecycle perspective is particularly relevant for higher education, where frequent low-stakes AI use may appear individually insignificant but can become consequential when repeated across large student populations. For students, however, the environmental cost of AI use is usually invisible, indirect and difficult to estimate, which may weaken its influence on everyday decisions. At the student-use level, Green AI is operationalised in this study through observable or reportable choices that may reduce avoidable computational demand: selecting a lightweight tool for simple tasks, avoiding unnecessary regeneration, combining related requests, limiting redundant prompts, reusing adequate outputs, and considering energy efficiency when choosing between tools.

The relationship between awareness and behaviour is a useful lens for understanding this problem. Environmental and ethical behaviour research has long shown that knowledge, concern or positive attitudes do not automatically translate into consistent action. The value-action gap, or attitude-behaviour gap, describes the mismatch between what individuals know or value and what they actually do (Blake, 1999; Kollmuss & Agyeman, 2002). Similar patterns have been identified in ethical consumption, where expressed ethical intentions often fail to produce corresponding behaviour (Carrington et al., 2014). Behavioural theories suggest that action is shaped not only by attitudes, but also by subjective norms, perceived behavioural control, habits, convenience and feedback mechanisms (Ajzen, 1991; Stern, 2000; Verplanken & Wood, 2006). Applied to AI use, students may recognise that outputs should be verified, sensitive data should be protected and unnecessary AI use should be avoided, yet still fail to act accordingly when deadlines, unclear expectations or convenient interfaces make responsible behaviour more difficult.

This perspective is especially important because much of the existing research on generative AI in higher education focuses on perceptions, attitudes, benefits, risks, intentions or institutional policies (Almassaad et al., 2024; Chan & Hu, 2023; Fošner, 2024). Related work has also developed AI ethics guidelines for education, emphasising transparency, privacy, accountability, academic integrity, human oversight, and responsible pedagogical integration (European Commission, 2022; Holmes & Miao, 2023). In parallel, Green AI and sustainable computing research has examined the environmental implications of AI systems, including energy consumption, carbon emissions, infrastructure and technical efficiency (Dalvi-Esfahani et al., 2020; Raman et al., 2024; Schwartz et al., 2020; Tabbakh et al., 2024; Verdecchia et al., 2023). These bodies of work are valuable, but they have often developed separately. Less attention has been given to how ethical and environmental

responsibility intersect in students' everyday academic AI use.

Therefore, the key gap addressed in this study is not simply whether students are aware of Responsible AI or Green AI, but whether such awareness is translated into reported academic behaviour. Existing studies often examine what students think about AI, how they perceive its benefits and risks, or how institutions should regulate its use. Less is known about the relationship between students' awareness and their routine practices, including prompting, verification, disclosure, privacy protection and tool selection. This gap is especially important among students in AI-related programmes, who are frequent users of AI tools and potential future developers, analysts or decision-makers. This exploratory study addresses this gap by examining how students enrolled in two AI-related programmes at Tallinn University report understanding and enacting Responsible AI and Green AI in academic work, and by identifying the contextual factors that enable or constrain responsible and sustainable behaviour.

3 Methodology

This study used an exploratory mixed-methods design to examine how students understand and apply principles of Green AI and Responsible AI in their academic use of AI tools. A mixed-methods approach was appropriate because the study aimed not only to describe students' levels of awareness and reported practices, but also to understand why responsible behaviours may or may not occur in everyday academic contexts. The quantitative component provided a descriptive overview of students' AI use, awareness and self-reported practices, while the qualitative component offered deeper insight into students' reasoning, experiences and perceived barriers to responsible AI use.

The study was conducted among students enrolled in AI-related programmes at Tallinn University. The participants included bachelor's-level Informatics students and master's students in the Artificial Intelligence for Sustainable Societies programme. These groups were selected because they are likely to use AI tools regularly in academic work and may also become future developers, analysts or decision-makers in AI-related fields. The survey sample consisted of 57 respondents, including 28 bachelor's students and 29 master's students. In addition, six students participated in semi-structured interviews. The sample was exploratory and context-specific rather than statistically representative, but it was suitable for identifying patterns of awareness, practice and contextual influence among frequent academic users of AI tools.

Data were collected through an online questionnaire and semi-structured interviews in spring 2026. The questionnaire examined students' frequency of AI use, types of AI tools used, academic use cases, awareness of Green AI and Responsible AI, knowledge of AI-related energy consumption, environmentally responsible AI use, privacy-related behaviour, verification practices, disclosure of AI use and perceptions of institutional guidance. Several questions used scenario-based items to examine how students responded to concrete academic situations involving privacy, academic integrity, sustainability and tool selection. This helped move beyond general attitudes by asking students to evaluate practical choices in realistic contexts.

The questionnaire contained 46 items organised into five domains: AI-use frequency and purposes; Responsible AI awareness; Green AI awareness; reported responsible practices; and perceptions of institutional and tool-level support. Attitudinal and self-reported behavioural items used a five-point Likert scale ranging from 1 ('strongly disagree'/'never') to 5 ('strongly agree'/'always'), as appropriate. Knowledge and scenario items used categorical response options with one or more responses identified as reflecting the responsible choice.

The items were developed from the study's conceptual framework and informed by literature on Responsible AI, Green AI, privacy, verification, disclosure and the awareness-behaviour gap. The instrument was reviewed by two experts for content relevance.

The semi-structured interviews were used to explore selected survey themes in greater depth. Participants were recruited voluntarily through course-related communication channels and student networks within the relevant programmes. Bachelor's students were primarily approached through course communication channels, while master's students were contacted through programme-related communication and peer networks. The six interview participants had also completed the survey and were selected purposively based on their familiarity with AI tools and willingness to reflect on their personal experiences. Interview questions focused on students' academic use of AI tools, verification of AI-generated outputs, disclosure of AI assistance, handling of sensitive or private data, consideration of environmental impact and perceptions of institutional expectations. The interviews also examined contextual factors such as time pressure, convenience, trust in AI outputs, uncertainty about university rules and the design of AI tools. Before participation, interviewees received a consent form and provided written informed consent. Participants were informed about the purpose of the research, the voluntary nature of participation, anonymity and secure data handling. This qualitative component was important for interpreting the awareness-practice gap, as it allowed students to explain why responsible practices may be difficult to maintain in routine academic work.

Survey data were analysed descriptively. The analysis focused on identifying patterns in students' AI

use, awareness of Green AI and Responsible AI, and self-reported responsible practices. Bachelor's and master's students were compared across Green AI awareness, Responsible AI awareness, privacy-related behaviour and environmentally responsible AI use. Group means and response distributions were examined to identify possible differences between the two academic levels. Because several constructs were measured using single items and the subgroup sizes were small, the comparisons were restricted to descriptive statistics and should not be interpreted as evidence of statistically reliable group differences. The comparative findings were therefore treated as exploratory and were not intended to support broad statistical generalisation. All behavioural measures concerned reported rather than directly observed practices. The study therefore captures participants' accounts and perceptions of their behaviour, not independently verified AI-use routines.

Interview transcripts were analysed using a hybrid deductive–inductive thematic procedure. Initial deductive codes were derived from the research questions and survey domains, including verification, privacy, disclosure, environmental consideration, time pressure and institutional guidance. Additional codes were developed inductively when recurrent ideas emerged from the transcripts. Codes were grouped into broader themes through iterative comparison across participants. Survey and interview findings were then integrated using a theme-by-source matrix that identified convergence, complementarity and divergence between the two datasets.

4 Findings

This section presents the main findings on students' awareness and reported practices concerning Green AI and Responsible AI in academic work.

4.1 AI Use as a Routine Academic Practice

AI use was a regular part of academic work for the majority of respondents. Of the 57 survey participants, 40.4% reported using AI tools daily and 45.6% reported using them several times a week. A further 8.8% used AI tools once a week, while only a very small proportion reported using them a few times a month or rarely. Taken together, more than 85% of respondents used AI tools at least several times a week, indicating that AI was not an occasional or experimental resource but part of students' routine study practices.

ChatGPT was the most widely used tool, selected by 80.7% of respondents. Other tools were also used, including Google Gemini by 47.4%, GitHub Copilot or similar AI code assistants by 35.1%, Grammarly or similar writing tools by 15.8%, and Microsoft Copilot by 7.0%. A relatively large group, 54.4%, selected "other" tools, suggesting that students' AI use extended beyond the most visible commercial platforms.

Students used AI across a broad range of academic tasks. The most common use case was coding, debugging or code generation, reported by 78.9% of respondents. This was followed by summarising readings or research papers, reported by 77.2%. Writing essays, reports or assignments and using AI for exam preparation or concept explanation were each reported by 68.4% of respondents. Translation or grammar correction was reported by 57.9%, while 35.1% used AI for data analysis or statistics. These results show that AI tools were embedded in both technical and non-technical academic activities.

The interview data reinforced this pattern of routinisation. Students described AI as part of their normal study workflow rather than as a special or exceptional tool. One participant stated, "It's part of my workflow… I use it every day" (Int 2), while another described it as habitual: "It's like a habit… like brushing" (Int 5). This routine use is important because responsible AI behaviour is shaped through repeated small decisions, such as whether to verify an answer, disclose AI assistance, avoid sharing data or choose a lower-impact tool.

4.2 Awareness of Green AI and Responsible AI

The survey results indicate that students had a generally positive level of awareness of Green AI. In response to the Green AI awareness item, 43.9% of respondents selected "strongly agree" and 31.6% selected "agree," while 17.5% selected a neutral response. This suggests that approximately three-quarters of the respondents expressed agreement that they were aware of Green AI concepts. However, the presence of a notable neutral group also indicates that this awareness was not uniform across the sample.

Knowledge of AI energy consumption was more mixed. In the knowledge-based item on AI training and usage, 63.2% of respondents selected the correct answer: that AI training consumes energy during model development, while AI usage consumes energy continuously. However, 33.3% selected the option that training and usage consume the same amount of energy, 22.8% indicated that usage always consumes more energy than training, and 1.8% selected the option that energy used during training does not affect environmental impact. Because respondents could select more than one option, the percentages exceed 100%. These results suggest partial environmental literacy: many students understood the basic distinction between training and usage-related energy consumption, but misconceptions remained.

Students also showed awareness that AI use has environmental consequences. In relation to the perceived environmental impact of AI use, the largest

response categories were "agree" and "strongly agree." However, perceived personal responsibility was more evenly distributed, with agreement, disagreement and neutral responses all present. This indicates a distinction between recognising that AI has environmental impact and feeling personally responsible for reducing that impact.

Interview findings also reflected this uneven awareness. Some students clearly recognised that AI systems require substantial resources. One participant stated, "I know it's bad… data centres use a lot of water and energy" (Int 6). Another explained, "It definitely consumes power… I don't know exact numbers, but I'm aware" (Int 2). These comments suggest that students often had a general awareness of environmental impact, but not necessarily detailed or systematic knowledge.

Awareness of Responsible AI was stronger. In the Responsible AI awareness item, 61.4% of respondents selected "strongly agree" and 26.3% "agree," while 8.8% selected a neutral response. This indicates that nearly nine in ten respondents expressed agreement that they were aware of Responsible AI concepts. Awareness of data privacy risks was similarly high: 43.9% agreed and 42.1% strongly agreed that they understood the privacy risks of entering sensitive information into commercial AI tools.

Knowledge-based and scenario-based items further confirmed that students could identify responsible options in structured situations. In a Responsible AI principles knowledge check, 86.0% of respondents correctly identified human oversight and verification as the key issue. In a privacy-related knowledge item, 89.5% correctly identified that a dataset containing participant names and demographic information should not be uploaded to a public AI chatbot without institutional approval. These findings suggest that students are able to recognise responsible AI choices when risks are made explicit. Taken together, the survey and interview evidence converged in showing relatively high conceptual recognition of Responsible AI risks. However, the interviews qualified the survey findings by indicating that awareness was often general rather than detailed, especially in relation to environmental impacts.

4.3 Responsible Practice and the Awareness-Practice Gap

The central finding is that awareness did not consistently translate into responsible academic behaviour. Students were often able to identify responsible choices in abstract or scenario-based questions, but their reported everyday practices were less consistent. This awareness-practice gap was visible in three main areas: verification of AI outputs, privacy protection and environmentally restrained use.

In the academic integrity scenario, 42.1% of respondents selected rewriting the AI-generated content in their own words, while 33.3% indicated that they would add a brief note acknowledging AI use. Smaller groups indicated that they would submit the work without mentioning AI use or ask the lecturer, while 10.5% selected "not sure." These results suggest that many students could identify responsible options, but uncertainty and variation remained around disclosure and academic transparency.

Interview data showed that disclosure was often conditional rather than routine. One participant stated, "If they don't ask, I wouldn't mention it" (Int 6). Another explained, "If I wrote it and improved it with AI, it's still my work" (Int 5). These comments illustrate that students may interpret AI assistance differently depending on whether they see it as support, collaboration or authorship. Disclosure therefore appears to depend strongly on perceived expectations rather than on a stable norm. The two forms of evidence were complementary: the survey demonstrated variation in chosen disclosure responses, while the interviews explained that disclosure decisions depended on perceived authorship and lecturer expectations.

Verification practices also showed an awareness-practice gap. Students recognised that AI outputs can be inaccurate, fabricated or biased. One participant noted, "It makes up sources… they don't even exist" (Int 6), while another observed, "Sometimes it just accepts my ideas even if they are wrong" (Int 2). Despite this awareness, verification was not always systematic. Students reported checking more carefully for important assignments, but using AI outputs with less scrutiny for lower-stakes work. As one participant explained, "If it's important, I check everything… if not, I just wing it" (Int 5). Another stated, "For less relevant work, I don't always check" (Int 6). The survey and interview data together suggest that verification was not absent but conditional. Students reported applying stronger checks to high-stakes work and lighter checks to routine or lower-stakes tasks.

Privacy behaviour followed a similar pattern. In a privacy scenario, 64.9% of respondents selected removing identifiers before using an AI tool, while 28.1% indicated that they would avoid uploading the data altogether. This shows that students could identify safer privacy practices when the issue was clearly presented. However, routine privacy-related practices were weaker. Survey findings on checking privacy policies before using new AI tools showed that a large proportion of respondents selected "never" or "rarely," with only a small group indicating that they did this often or always. The datasets showed a discrepancy between students' ability to identify a safe response in a structured scenario and their reported reluctance to examine privacy conditions during routine tool selection.

The interviews help explain this gap. Students were aware that AI tools may store or reuse inputs, but this awareness did not always prevent data sharing. One participant reflected, "It came to my mind… but I just ignored it" (Int 5). Another stated more directly,

"You're basically giving away your data" (Interview 6). These comments suggest that privacy awareness exists, but students may treat privacy checks as additional effort rather than as a normal part of AI-supported academic work.

Environmental responsibility was also inconsistent. In the environmentally responsible use scenario, 80.7% of respondents selected the use of a built-in spell-checker or lightweight grammar tool for a simple spelling correction task. This indicates that students could identify a lower-impact option when the choice was made explicit. However, environmental considerations were much less visible in routine tool selection. Survey results on whether students considered privacy policies or energy efficiency when choosing AI tools showed that "rarely" and "never" were common responses, while "often" and "always" were selected by only a limited number of respondents. Interview data supported this interpretation. Students recognised environmental impact but did not usually treat it as a priority in everyday use. One participant stated, "It's in the back of my mind, but I don't think about it daily" (Int 2). Another commented, "When I need information, I don't think about the environment" (Int 5). A third summarised the tension clearly: "I know it's bad… but it's convenient" (Int 6). These findings indicate that environmental responsibility is particularly vulnerable to being displaced by speed, convenience and task completion. The scenario result therefore indicates recognition of a comparatively lower-impact option, whereas the interviews suggest that environmental considerations rarely functioned as a spontaneous criterion in everyday use.

4.4 Contextual Influences on Responsible AI Behaviour

The findings show that the awareness-practice gap was not simply caused by lack of knowledge. Instead, students' behaviour was shaped by contextual factors, including time pressure, perceived task importance, institutional ambiguity and tool design.

Time pressure was one of the strongest barriers. In the survey item on whether deadlines make responsible AI use more difficult, agreement-based responses were common, indicating that many students perceived time pressure as a constraint on responsible behaviour. Interview participants also described deadlines as increasing dependence on AI. One student stated, "Closer to deadlines, it's more tempting to use AI" (Int 6). Another explained that responsible checking takes additional effort: "It costs too much time to check everything yourself" (Int 6). These findings suggest that responsible AI practices are often perceived as desirable but costly.

Task importance also influenced behaviour. Students reported that they were more likely to verify outputs and act carefully in high-stakes situations, while shortcuts were more common in routine or lower-stakes tasks. This indicates that responsibility was conditional rather than habitual. Students did not necessarily apply a fixed responsible AI routine across all academic work; instead, they adjusted their behaviour according to perceived risk and importance.

Institutional and lecturer guidance also shaped behaviour. Survey responses on institutional clarity showed a high proportion of neutral responses, suggesting uncertainty about whether university rules on student AI use were clear. Responses on lecturer guidance showed a similar pattern, with many students selecting neutral or mixed responses. Interview findings confirmed this ambiguity. One participant stated, "No course directly covered this" (Int 6), while another said, "It's something I just kind of knew" (Int 2). These comments suggest that students often rely on personal judgement rather than structured institutional guidance.

Tool design was another important influence. Survey results on whether AI tools make it easy to choose options that protect privacy or use less energy showed limited confidence in current tool support. Many respondents selected neutral, disagreement or "not sure" responses. Interview participants suggested that AI tools rarely make environmental or privacy consequences visible at the point of use. One participant argued, "If it showed I wasted 100 liters of water, I might think twice" (Int 6). Another stated, "Systems don't make you conscious about it" (Int 2). These findings suggest that responsible behaviour is affected not only by student awareness but also by the design of AI interfaces.

The comparative analysis between bachelor's and master's students showed only small differences. Bachelor's students reported slightly higher mean scores for Green AI awareness (M = 4.36) than master's students (M = 4.12), and also higher awareness of AI energy consumption (M = 4.46 compared with M = 4.07). Responsible AI awareness was high in both groups, with only a minimal difference between bachelor's students (M = 4.59) and master's students (M = 4.50). Privacy behaviour was also similar, although bachelor's students reported slightly higher data-avoidance behaviour (M = 4.10) than master's students (M = 3.96). In contrast, master's students reported stronger privacy policy checking behaviour (M = 2.68) than bachelor's students (M = 1.62). Thus, academic level did not strongly differentiate awareness or practice. Because these comparisons were descriptive and based on small subgroups, the observed differences should be interpreted as exploratory patterns rather than statistically established differences between bachelor's and master's students.

Within this sample, students in the two AI-related programmes were frequent users of AI tools and generally reported recognising ethical, privacy and environmental concerns. However, responsible behaviour remains inconsistent. Students are often able to identify responsible choices in structured scenarios, but everyday practice is shaped by convenience, time

pressure, unclear expectations and limited tool-level support. The central pattern observed in this exploratory sample was not a simple absence of awareness, but a gap between knowing what responsible AI use requires and consistently doing it in routine academic work.

5 Discussion

The findings challenge a simple information-deficit model of AI responsibility. Students' inconsistent responsible AI practices cannot be explained only by a lack of awareness. On the contrary, the results suggest that many students recognise key ethical, privacy and environmental concerns associated with AI use. However, this awareness competes with other factors in everyday academic work, including efficiency, convenience, time pressure, perceived task importance, confidence in AI outputs, institutional ambiguity and the design of AI tools. This supports behavioural perspectives that view action as shaped not only by knowledge and attitudes, but also by norms, habits and perceived behavioural control.

The awareness-practice gap should therefore be interpreted carefully. Inconsistent verification of AI outputs, for example, does not necessarily mean that students are unaware of AI risks or indifferent to academic quality. It may also reflect pragmatic judgement. Students may verify outputs more carefully when the task is high stakes, factually sensitive, unfamiliar or formally assessed, while applying lighter checks to low-stakes tasks such as brainstorming, language refinement or preliminary explanation. From this perspective, selective verification may be a form of risk-based decision-making rather than simple irresponsibility. Nevertheless, such pragmatism becomes problematic when students overestimate their ability to judge when verification is necessary, or when fluent AI outputs create a false sense of reliability. The issue is therefore not only whether students verify, but whether they possess sufficiently robust criteria for deciding when and how verification should occur.

A similar point applies to privacy behaviour. The finding that students do not routinely read privacy policies should not be interpreted only as a failure of individual responsibility. Privacy policies are often lengthy, legalistic and difficult to compare across tools. For students working under time constraints, reading such documents may appear unrealistic or disproportionate to the immediate academic task. This suggests that the gap between privacy awareness and privacy practice is partly structural. Students may understand that sensitive data should be protected, but lack accessible, actionable information about what different tools do with uploaded content. In this sense, privacy responsibility cannot be placed entirely on students; it also depends on institutional guidance, tool transparency and the availability of safer alternatives.

The Green AI findings raise a related issue. Environmental responsibility is particularly difficult to operationalise because the environmental impact of AI use is remote, distributed and invisible. Unlike academic integrity violations, which are directly connected to assessment rules, the energy and water costs of AI use are not visible in the interface or in the assignment process. Students may therefore understand that AI systems have environmental consequences while still failing to consider these consequences during ordinary tool selection. This does not necessarily indicate lack of concern. Rather, it reflects weak feedback mechanisms and limited perceived control. When students cannot see, estimate or compare environmental costs, sustainability becomes an abstract concern rather than a practical decision criterion.

These findings suggest that Responsible AI and Green AI should be understood as sociotechnical practices rather than only individual choices. Students make decisions within environments shaped by course expectations, institutional policies, assessment pressure and commercial platform design. When university guidance is ambiguous, students must infer what counts as acceptable assistance, collaboration, authorship or misconduct. When AI interfaces prioritise speed and ease of use, they encourage immediate task completion rather than reflection. When tools do not display privacy risks, uncertainty levels or environmental signals, responsibility is delegated to users at the moment when convenience is most salient. The result is a mismatch between the expectations placed on students and the supports available to help them act responsibly.

This distinction between individual, institutional and platform responsibility is important. Students remain responsible for how they use AI in academic work, particularly in relation to verification, authorship, disclosure and data protection. However, their responsibility should not be treated as isolated from the systems in which AI use occurs. Universities have responsibility for establishing clear norms, providing practical guidance and designing assessment conditions that encourage transparent and critical AI use. Platform providers also have responsibility for making risks visible, offering privacy-preserving defaults and supporting users in making proportionate choices. Responsible AI behaviour is therefore not produced by awareness alone, but by the alignment of student competence, institutional expectations and tool design.

The comparison between bachelor's and master's students did not reveal a uniform academic-level pattern. Awareness scores were broadly similar, while the clearest descriptive difference concerned privacy-policy checking. This suggests that progression from bachelor's to master's study may not, by itself, produce consistently more responsible AI routines. However, the subgroup sizes were too small to support a firm conclusion, and programme-specific curriculum,

research experience or assessment practices may be more relevant than study level alone. Future studies should examine these factors using larger stratified samples.

For Responsible AI, the practical challenge is operationalisation. Principles such as transparency, accountability, privacy and human oversight are widely accepted, but they must be translated into repeatable academic routines. A responsible student workflow might include defining the task before prompting, avoiding the upload of sensitive data, checking factual claims against reliable sources, revising outputs in one's own words, documenting the role of AI assistance and confirming whether AI use is permitted for the assignment. However, these routines should not simply be assumed. They need to be taught, practised, assessed and reinforced across courses. Otherwise, responsible AI remains an abstract ideal rather than a stable academic habit.

For Green AI, students need practical heuristics rather than only general environmental warnings. Since students cannot realistically calculate the precise environmental cost of each interaction, guidance should focus on proportionate and actionable choices. Examples include using built-in spellcheckers or lightweight grammar tools for simple tasks, avoiding unnecessary regeneration of responses, batching related prompts, selecting smaller or institutionally recommended tools where appropriate, and reserving large generative systems for tasks where they provide meaningful value. Such heuristics would make sustainability more usable in everyday academic decision-making.

The results also imply that universities should avoid treating AI policy primarily as a disciplinary instrument. Strict prohibition may drive AI use underground and increase non-disclosure, while vague permissiveness may leave students uncertain about acceptable practice. A more productive approach is to distinguish between acceptable, conditional and unacceptable uses of AI. For example, brainstorming, language refinement or debugging may be permitted under specified conditions, while undisclosed generation of substantive assessed content may be prohibited. Clear disclosure templates and course-level examples can reduce uncertainty and normalise transparency.

This study suggests that AI literacy should be extended to include both responsibility literacy and sustainability literacy. For students in AI-related programmes, this is especially important because their habits as users may shape their future assumptions as developers, analysts or decision-makers. If AI is experienced primarily as a tool for speed and convenience, students may reproduce systems that externalise ethical and environmental costs. Conversely, if verification, transparency, privacy protection and sustainability are embedded into academic routines, responsible AI can become part of students' professional identity. The central implication is therefore that responsible AI use should not be framed merely as individual compliance, but as a shared academic and sociotechnical competence supported by education, policy and design.

6 Limitations and Future Research

The study is exploratory and context-specific. The survey sample of 57 students and six interviews provide rich insight into patterns at one university, but they do not support broad statistical generalisation. The study relies on retrospective self-reports rather than direct observation, system logs or diary records. This introduces social desirability, recall and interpretation biases. Participants may overstate practices perceived as responsible, such as verification or privacy protection, and may differ in how they understand response categories such as 'often' or 'rarely.' Accordingly, the study cannot determine how frequently these behaviours occurred in actual AI interactions or establish the exact magnitude of the awareness-practice gap. The findings should be interpreted as evidence of perceived and reported practice.

Future research should use larger and more diverse samples, compare disciplines, and examine how policy changes affect behaviour over time. Observational or diary-based methods could capture actual AI-use routines more accurately than retrospective self-reporting. Experimental work could test whether interface prompts, disclosure templates, energy feedback or privacy warnings change student behaviour. Longitudinal research is also needed to understand whether responsible AI habits developed during university persist into professional practice.

7 Conclusion

This study examined how students in AI-related programmes understand and practise Green AI and Responsible AI in their academic use of AI tools. Based on survey and interview data, the findings suggest that participants in these two programmes were frequent AI users and generally recognised ethical, privacy and environmental concerns. However, this awareness does not consistently translate into everyday academic practice, particularly in relation to verification, disclosure, privacy protection and environmentally restrained tool use.

The principal pattern identified in this exploratory, self-reported dataset is a possible awareness-practice gap: students often know what responsible and sustainable AI use requires, but their reported behaviour appears to be associated with time pressure, convenience, task importance, unclear institutional expectations and limited tool-level support. This suggests that responsible AI use cannot be explained

only through individual knowledge or intention; it must also be understood as a sociotechnical practice shaped by educational and platform environments.

The paper offers a context-specific contribution to higher education AI research by connecting Responsible AI and Green AI through the lens of routine academic practice, and by showing that AI literacy should include situated behavioural competence rather than only conceptual awareness. Responsible AI use in higher education depends on making responsible behaviour ordinary, visible and institutionally supported. These implications should be interpreted as propositions for further investigation rather than as general conclusions about university students. Future efforts should therefore focus on embedding verification, transparency, privacy protection and sustainability into teaching, assessment, institutional guidance and AI tool design.

References

Ajzen, I. (1991). The Theory of Planned Behavior. *Organizational Behavior and Human Decision Processes*, 50(2), 179-211. doi: 10.1016/0749-5978(91)90020-T

Almassaad, A., Alajlan, H., & Alebaikan, R. (2024). Student Perceptions of Generative Artificial Intelligence: Investigating Utilization, Benefits, and Challenges in Higher Education. *Systems*, 1*2*(10), 385.doi: 10.3390/systems12100385

Athanassopoulos, S., Tzavara, A., Aravantinos, S., Lavidas, K., Komis, V., & Papadakis, S. (2026). Teacher Education Students' Practices, Benefits, and Challenges in the Use of Generative AI Tools in Higher Education. *Education Sciences*, 16(2), 228. doi:10.3390/educsci16020228

Blake, J. (1999). Overcoming the 'Value-Action Gap' in Environmental Policy: Tensions Between National Policy and Local Experience. *Local Environment*, 4(3), 257-278. doi: 10.1080/13549839908725599

Carrington, M. J., Neville, B. A., & Whitwell, G. J. (2014). Lost in Translation: Exploring the Ethical Consumer Intention–Behavior Gap. *Journal of Business Research*, 67(1), 2759-2767.doi: 10.1016/j.jbusres.2012.09.022

Chan, C. K. Y., & Hu, W. (2023). Students' Voices on Generative AI: Perceptions, Benefits, and Challenges in Higher Education. *International Journal of Educational Technology in Higher Education*, 20(1), 43. doi: 10.1186/s41239-023-00411-8

Dalvi-Esfahani, M., Alaedini, Z., Nilashi, M., Samad, S., Asadi, S., & Mohammadi, M. (2020). Students' Green Information Technology Behavior: Beliefs and Personality Traits. *Journal of Cleaner Production*, 257, 120406. doi: 10.1016/j.jclepro.2020.120406

Deng, R., Jiang, M., Yu, X., Lu, Y., & Liu, S. (2025). Does ChatGPT Enhance Student Learning? A Systematic Review and Meta-analysis of Experimental Studies. *Computers & Education*, 227, 105224. doi:10.1016/j.compedu.2024.105224

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., … Wright, R. (2023). Opinion Paper: "So What if ChatGPT Wrote it?" Multidisciplinary Perspectives on Opportunities, Challenges and Implications of Generative Conversational AI for Research, Practice and Policy. *International Journal of Information Management*, 71, 102642. doi: 10.1016/j.ijinfomgt.2023.102642

European Commission. Directorate General for Education, Youth, Sport and Culture. (2022). *Ethical guidelines on the use of artificial intelligence (AI) and data in teaching and learning for educators.* Publications Office. https://data.europa.eu/doi/10.2766/153756

European Commission. (2024). *Living guidelines on the responsible use of generative AI in research*. Directorate-General for Research and Innovation. https://research-and-innovation.ec.europa.eu/document/download/2b6cf7e5-36ac-41cb-aab5-0d32050143dc_en

European Commission. Directorate General for Communications Networks, Content and Technology & High Level Expert Group on Artificial Intelligence. (2019). *Ethics guidelines for trustworthy AI.* Publications Office. https://data.europa.eu/doi/10.2759/346720

Floridi, L., & Cowls, J. (2019). A Unified Framework of Five Principles for AI in Society. *Machine learning and the city: Applications in Architecture and Urban Design*, 535-545. doi: 10.1162/99608f92.8cd550d1

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People - An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, 28(4), 689-707. doi: 10.1007/s11023-018-9482-5

Fošner, A. (2024). University Students' Attitudes and Perceptions Towards AI Tools: Implications for Sustainable Educational Practices. *Sustainability*, 16(19), 8668. doi: 10.3390/su16198668

Francis, N.J., Jones, S., & Smith, D.P. (2025) Generative AI in Higher Education: Balancing Innovation and Integrity. *British Journal of Biomedical Science*, 81:14048. doi:10.3389/bjbs.2024.14048

Holmes, W., & Miao, F. (2023). *Guidance for Generative AI In Education and Research*. UNESCO Publishing.

Jobin, A., Ienca, M., & Vayena, E. (2019). The Global Landscape of AI Ethics Guidelines. *Nature Machine Intelligence*, 1(9), 389-399. doi: 10.1038/s42256-019-0088-2

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., ... & Kasneci, G. (2023). ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. *Learning and Individual Differences*, 103, 102274. doi: 10.1016/j.lindif.2023.102274

Kollmuss, A., & Agyeman, J. (2002). Mind the Gap: Why Do People Act Environmentally and What are the Barriers to Pro-Environmental Behavior? *Environmental Education Research*, 8(3), 239-260. doi: 10.1080/13504620220145401

Mittelstadt, B. (2019). Principles Alone Cannot Guarantee Ethical AI. *Nature Machine Intelligence*, 1(11), 501-507. doi: 10.1038/s42256-019-0114-4

Morley, J., Kinsey, L., Elhalal, A., Garcia, F., Ziosi, M., & Floridi, L. (2023). Operationalising AI Ethics: Barriers, Enablers and Next Steps. *AI & Society*, 38(1), 411-423. doi: 10.1007/s00146-021-01308-8

Nikolic, S., Wentworth, I., Sheridan, L., Moss, S., Duursma, E., Jones, R. A., ... & Middleton, R. (2024). A Systematic Literature Review of Attitudes, Intentions and Behaviours of Teaching Academics Pertaining to AI and Generative AI (GenAI) in Higher Education: An Analysis of GenAI Adoption Using the UTAUT Framework. *Australasian Journal of Educational Technology*, 40(6), 56-75. doi: 10.14742/ajet.9643

OECD (2026). *OECD Due Diligence Guidance for Responsible AI*, OECD Publishing, Paris, doi: 10.1787/41671712-en.

Raman, R., Pattnaik, D., Lathabai, H. H., Kumar, C., Govindan, K., & Nedungadi, P. (2024). Green and Sustainable AI Research: An Integrated Thematic and Topic Modeling Analysis. *Journal of Big Data*, 11(1), 1-28. doi:10.1186/s40537-024-00920-x

Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. *Communications of the ACM*, 63(12), 54–63. doi:10.1145/3381831

Stracke, C. M., Nahar, N., Punzo, V., Massaro, S., Pappa, D., Di Grassi, A., ... & Burgos, D. (2026). Artificial Intelligence Policies for Higher Education: Manifesto for Critical Considerations and a Roadmap. *MAP Education and Humanities*, 6, 61-73. doi: 10.53880/2744-2373.2025.6.61

Stern, P. C. (2000). New Environmental Theories: Toward a Coherent Theory of Environmentally Significant Behaviour. *Journal of Social Issues*, 56(3), 407-424. doi: 10.1111/0022-4537.00175

Tabbakh, A., Al Amin, L., Islam, M., Mahmud, G. I., Chowdhury, I. K., & Mukta, M. S. H. (2024). Towards Sustainable AI: A Comprehensive Framework for Green AI. *Discover Sustainability*, 5(1), 408. doi:10.1007/s43621-024-00641-4

Verdecchia, R., Sallou, J., & Cruz, L. (2023). A Systematic Review of Green AI. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 13(4), e1507. doi: 10.1002/widm.1507

Verplanken, B., & Wood, W. (2006). Interventions to break and create consumer habits. *Journal of Public Policy & Marketing*, *25*(1), 90-103. doi: 10.1509/jppm.25.1.90

Wang, H., Dang, A., Wu, Z., & Mac, S. (2024). Generative AI in Higher Education: Seeing ChatGPT Through Universities' Policies, Resources, and Guidelines. *Computers and Education: Artificial Intelligence*, 7, 100326. doi: 10.1016/j.caeai.2024.100326

Xiao, L., Pyng, H. S., Ayub, A. F. M., Zhu, Z., Gao, J., & Qing, Z. (2025). University Students' Usage of Generative Artificial Intelligence for Sustainability: A Cross-Sectional Survey from China. *Sustainability*, 17(8), 3541. doi:10.3390/su17083541

Perceptual Indicators for Teaching Quality: Transforming SET Data into Learning Analytics Tools for Adult Online Instructors

Jasenka Begić, Tihana Banko, Ana Lokas Ćošković
Algebra Bernays University
Department of Interdisciplinary Sciences
Gradišćanska Street 24, 10 000 Zagreb, Croatia
{jasenka.begic, tihana.banko, ana.lokascoskovic}@algebra.hr

Abstract. *This study examines a modified Students' Evaluation of Educational Quality (SEEQ)-based questionnaire adapted for adult online learners as a potential perceptual layer for learning analytics. Conducted at Algebra Bernays University in Croatia with 161 adult learners in online professional programs, the research analyzes relationships between teacher interactivity dimensions (motivation, constructive feedback, availability, discussion initiation, communication frequency) and overall satisfaction. Results show high satisfaction levels ($M = 4.18$, $SD = 0.55$) and identify teacher motivation ($\beta = .411$, $p < .001$) and constructive feedback ($\beta = .207$, $p = .010$) as the strongest unique predictors of satisfaction, together explaining significant variance ($R^2 = .607$, adjusted $R^2 = .594$). Building on these findings, we propose a proof-of-concept framework wherein SET-based interactivity scores function as interpretable, teacher-centered indicators for learning analytics dashboards and rule-based early-warning systems. The framework demonstrates how perceptual data from established evaluation instruments can be transformed into actionable metrics (e.g., Motivation Score, Feedback Quality Score) with traffic-light categorization to support timely pedagogical interventions in adult online education.*

Keywords. learning analytics, Student Evaluation of Teaching (SET), perceptual indicators, motivation, constructive feedback

1 Introduction

Online education for adult learners has expanded rapidly in recent years, creating flexible pathways for upskilling and reskilling, but also raising new questions about how to ensure and monitor the quality of online teaching. In this context, teacher interactivity – including motivation, constructive feedback, availability and use of interactive methods – has emerged as a key determinant of learners' satisfaction and engagement. Student Evaluation of Teaching (SET) instruments are widely used to capture learners' perceptions of teaching quality, yet they are still predominantly treated as isolated, summative quality assurance tools rather than as explicit inputs to learning analytics pipelines. Building on this gap, the present study examines a modified SET questionnaire developed for adult online programs as a potential perceptual signal for learning analytics and proposes a proof-of-concept framework in which SET-based measures of teacher interactivity are conceptualised as interpretable indicators for future dashboards and early-warning models.

2 Background and Literature Review

A substantial body of research shows that the quality of teacher–student interaction is a key predictor of engagement, satisfaction and learning outcomes in online environments. Large-scale evidence from 334 universities indicates that **teaching presence** – particularly the facilitation of discourse and provision of feedback – is a central determinant of students' online learning experience and satisfaction (Li & Wang, 2024). Recent work on instructor presence and interactive tools similarly demonstrates that perceived instructor presence, together with the use of interactive technologies, significantly predicts student engagement and satisfaction in hybrid learning environments (Park & Kim, 2020). Conceptual and empirical syntheses emphasise that **instructor interactivity** – through timely responses, personalised feedback and proactive communication is strongly associated with online student satisfaction and persistence (e.g., Roque-Hernández et al., 2024), converging with earlier findings that interactive teaching behaviours shape both learners' affective

experience and cognitive outcomes in digital environments (Sun & Chen, 2022).

Recent literature further underlines the importance of motivational and relational dimensions of online teaching. For instance, Pham and Nguyen (2025) explain that teaching presence in virtual learning environments is strongly connected with learners' perceptions of support, communication quality and overall course satisfaction. Furthermore, many authors emphasise that meaningful interaction, timely feedback and motivational support remain among the strongest predictors of persistence and engagement in online education (Bergdahl et al., 2024).

Student Evaluation of Teaching (SET) instruments are among the most widely used tools for capturing learners' perceptions of teaching quality at course and program level in higher education (Spooren et al., 2013). Within this family, the Students' Evaluation of Educational Quality (SEEQ) developed by Marsh (1982, 1984) is recognised as a reliable, valid and multidimensional instrument covering aspects such as organisation/clarity, enthusiasm, group interaction, individual rapport, breadth, examinations and overall rating (Marsh, 1982, 1984). Adaptations of SEEQ in different languages and systems have confirmed its factor structure and psychometric robustness, supporting its use as a basis for localised evaluation instruments (e.g. Daumiller et al., 2021 for the German adaptation). In online contexts, SET/SEEQ-derived evaluations have shown that students particularly value well-organised courses, engaging activities and instructors who are responsive and courteous in digital communication (Bolliger & Martindale, 2004; Watermark Insights, 2024). Despite this, SET results are still predominantly treated as end-of-course feedback rather than as structured data streams that could be integrated with other learning analytics signals.

In line with a widely used definition, **learning analytics** refers to systematically measuring, collecting, analysing, and reporting data about learners and their contexts to better understand and improve learning processes and environments (Lang, Siemens, Wise, Gašević & Merceron, 2022). Building on this perspective, the *Handbook of Learning Analytics (2nd ed.)* emphasises the integration of diverse data sources and methods to support both learners and educators, highlighting the need for interpretable indicators that can inform pedagogical decision-making (Lang et al., 2022). In practice, learning analytics implementations typically rely on log-level traces from learning management systems (logins, clicks, time on task), assessment data and indicators of participation in online activities (Papamitsiou & Economides, 2014), while predictive approaches increasingly apply machine-learning algorithms to forecast outcomes such as completion, grades or dropout risk in higher education (Čotić Poturić, 2024). Empirical work has shown that multiple data sources – LMS logs, assessment results and survey data – can be combined to assess student learning and course effectiveness in online courses (Harrati et al., 2016), yet perceptual data such as course evaluations and satisfaction surveys remain underused as first-class input variables and are often only reported descriptively alongside behavioural indicators (Lang et al., 2022; Čotić Poturić, 2024). This limitation has become increasingly evident in many learning analytics research studies, which argue that many existing analytics implementations primarily focus on behavioural traces and may overlook the perceptual and affective dimensions of learning experiences (Bergdahl et al., 2024).

Research in human-centred learning analytics further highlights this issue. Alfredo et al. (2024) explain that learning analytics and AI-supported educational systems should focus on interpretability, transparency and pedagogical usefulness, rather than depend only on predictive optimisation. Furthermore, de Vreugd et al. (2024) argue that instructors require actionable and pedagogically meaningful indicators instead of isolated behavioural metrics, while Herodotou et al. (2025) show that both students and instructors benefit from analytics systems that communicate engagement, support and learning progress in a comprehensive manner.

Bridging this gap requires reconceptualising SET not only as a summative evaluation instrument, but as a structured, learner-centred source of perceptual indicators that can complement behavioural and performance data within learning analytics frameworks (Lang et al., 2022). Dimensions measured by SEEQ-derived instruments – such as good teaching, clear goals and standards, assessment quality, workload and general skills – as well as items explicitly targeting teacher interactivity, offer interpretable variables that align closely with pedagogical concepts and instructor decision-making (Marsh, 1982, 1984; Daumiller et al., 2021).

In the context of online learning, where teaching presence and instructor interactivity are empirically linked to engagement and satisfaction (Li & Wang, 2024; Park & Kim, 2020; Roque-Hernández et al., 2024), SET-based measures of motivation, feedback and interaction are particularly promising as teacher-facing analytics indicators. Building on examples where survey data are combined with LMS and performance data in online courses (Harrati et al., 2016; Papamitsiou & Economides, 2014) and on regional work demonstrating the use of machine-learning methods in predictive learning analytics in higher education (Čotić Poturić, 2024), the present study examines a SEEQ-based instrument adapted for adult online programs as a candidate perceptual layer in learning analytics, and explores how its interactivity scales can be transformed into interpretable indicators and rule-based scenarios for dashboards and early-warning models.

3 Research Problem and Objectives

Learning analytics has advanced considerably in leveraging behavioural and performance data to monitor and support students, but there is still a lack of interpretable, teacher-centred indicators that can feed into models of satisfaction and risk in adult online programs. Existing systems predominantly rely on log data and grades, which reveal what learners do but say little about how they experience teaching quality and teacher interactivity in digitally mediated environments (Siemens & Baker, 2012; Papamitsiou & Economides, 2014). At the same time, Student Evaluation of Teaching (SET) instruments are widely used to capture perceptions of teaching in online and on-campus courses, yet they are rarely integrated into learning analytics pipelines or used as structured signals in predictive or diagnostic models (Spooren et al., 2013; Ifenthaler & Yau, 2020). This creates a disconnect between student-centred evaluation practices and data-driven infrastructures: teachers typically receive static, end-of-course SET reports, while analytics dashboards omit perceptual indicators that would make patterns of engagement and satisfaction more pedagogically interpretable. In the context of adult online education – where teacher-student interaction is a proven determinant of engagement and satisfaction (Sun & Chen, 2022; Yuan et al., 2022) – this lack of integrated, teacher-centred indicators limits the potential of learning analytics to support timely, targeted improvements in teaching practice.

In response to this problem, the present study has four objectives:

1. to examine adult learners' satisfaction with online teaching in professional programs at a Croatian higher education institution using a modified SET questionnaire tailored to adult online learners;
2. to analyse the relationships between different aspects of teacher interactivity (e.g., motivation, constructive feedback, availability, facilitation of discussion, use of online tools) and overall satisfaction with online teaching;
3. to assess the predictive power of teacher-interactivity aspects for overall satisfaction by means of multiple regression analysis, thereby identifying which dimensions function as the strongest perceptual predictors; and
4. to propose a proof-of-concept framework in which SET-based measures of teacher interactivity are conceptualised as candidate input variables for learning analytics dashboards and rule-based early-warning indicators in adult online programs.

4 Research Hypotheses

Based on the outlined research problem and objectives, and building on prior empirical findings on teacher-student interaction, satisfaction and SET in online learning (Alfredo et al., 2024; Sun & Chen, 2022; Yuan et al., 2022; Spooren et al., 2013), the study tests the following hypotheses:

- H1: Adult learners report a significantly high level of satisfaction with online teaching, as measured by the modified SET questionnaire.
- H2: Aspects of teacher interactivity (motivation, constructive feedback, availability, facilitation of discussion, communication frequency, use of online tools and interactive methods) are significantly positively associated with adult learners' overall satisfaction with online teaching.
- H3: A combination of teacher-interactivity aspects significantly predicts adult learners' overall satisfaction with online teaching.

5 Methodology

The study employed a cross-sectional, quantitative research design based on an online questionnaire, with a subsequent proof-of-concept interpretation in terms of learning analytics. Data were collected at a single point in time from adult learners enrolled in online professional programs and analysed using descriptive statistics, correlation analysis, and multiple regression to address the stated hypotheses. In addition to the conventional psychometric and predictive analyses, the results were re-interpreted through a learning analytics lens by identifying which questionnaire scales could serve as interpretable, teacher-centred indicators for future dashboards and early-warning rules.

5.1 Sampling and Participants

The target population comprised adult learners enrolled in fully online programs of adult education at Algebra Bernays University in Croatia. All participants were engaged in online professional programs leading to partial qualifications or micro-credentials in fields such as graphic design, UX design, web development and related digital competencies. A non-probability convenience sampling approach was used: an invitation to participate in the study was distributed via email to all adult learners registered in online programs and the survey was also advertised through the institution's website and social-media channels. Participation was voluntary and anonymous.

A total of 161 participants took part in the study. The majority of participants identified as female (60.9%), with 36.0% male and 3.1% choosing not to disclose their gender. The most represented age group was 25-34 years (44.7%), followed by 35-44 years

(30.4%), 45-54 years (14.3%) and 18-24 years (9.3%), with a small proportion aged 55-64 years (1.2%). Regarding educational attainment, 41.6% held a Master's degree, 16.8% a Bachelor's degree and 34.2% had completed secondary school, with smaller proportions holding specialist or doctoral degrees. Most respondents were currently employed (72.7%), reflecting the profile of working adult learners, while 19.3% were unemployed and smaller groups reported other statuses (e.g., self-employed, student, retired). In terms of experience with online education, 58.2% had been involved in online learning for less than six months, 35.4% for 6-12 months, and a minority for longer periods (1-2 years or more than two years).

5.2 Instrument: Modified Student Evaluation of Teaching (SET)

For this study, we used a student-evaluation instrument based on the Students' Evaluation of Educational Quality (SEEQ) questionnaire developed by Marsh (1982), which is a reliable, valid and useful tool for collecting students' evaluations of university teaching across multiple dimensions (Marsh, 1982, 1984). SEEQ has been adapted and validated in various languages and higher-education systems, confirming its multidimensional structure and psychometric robustness (Daumiller et al., 2021). Building on this foundation, the SEEQ-based instrument was translated into Croatian and linguistically adapted to the context of adult learners in online programs at Algebra Bernays University, with items kept as close as possible to the original content while ensuring clarity and relevance for the local context.

In addition to the core SEEQ dimensions, two global items were added: an overall rating of the instructor and an overall rating of the online tools and platforms used in the course. These additions were motivated by evidence that online learners' satisfaction is shaped not only by traditional aspects of teaching quality, but also by their holistic perception of the instructor and the digital environment (Bolliger & Martindale, 2004; Park & Kim, 2020). Furthermore, a set of items explicitly targeting teacher interactivity in online settings (e.g. motivation, constructive feedback, availability, initiating discussions, use of interactive tools) was included to reflect recent work on teaching presence and instructor interactivity in online classrooms (Li & Wang, 2024; Roque-Hernández et al., 2024). The adapted instrument thus retains the theoretically grounded structure of SEEQ while extending it with context-sensitive items that capture the specific demands of adult online education.

The final questionnaire comprised several multi-item scales grouped into two main blocks. The first block captured different aspects of satisfaction with teaching through the dimensions of good teaching (e.g., clarity of presentation, organisation, preparedness), clear goals and standards (clarity of learning objectives and expectations), assessment (perceived fairness and appropriateness of assessment methods), workload (perceived appropriateness of the amount and difficulty of work) and general skills (perceived contribution to transferable skills such as problem-solving or communication). Two additional single-item global measures assessed the overall instructor rating and the overall rating of online tools. The second block focused on teacher interactivity in the online environment and included items such as: the teacher motivates participants, the teacher provides constructive feedback, feedback on tests or short quizzes is valuable, the teacher responds only to the most essential questions (reverse-coded), the teacher is regularly available for additional consultations, the teacher regularly initiates discussions and encourages participation, satisfaction with the frequency of communication with the teacher, the teacher's skill in using online tools and platforms, and the use of interactive tools and methods during classes (e.g., group projects, quizzes, simulations). All items were rated on a five-point Likert-type scale ranging from 1 ("strongly disagree") to 5 ("strongly agree"), and scale scores were computed as mean values of the relevant items.

Overall satisfaction was operationalised as a composite score derived from satisfaction-related scales. For each participant, scale scores were calculated as the arithmetic mean of the items within each block (good teaching, clear goals and standards, assessment, workload, general skills, overall rating of the instructor, and overall rating of online tools). The overall satisfaction composite was then calculated as the mean of these seven scale scores, while higher values indicate higher perceived overall satisfaction.

Beyond its primary function as a summative and formative evaluation tool, the modified SET instrument was conceptualised in this study as a potential artefact for learning analytics in adult online education. Specifically, its dimensions - good teaching, clear goals and standards, assessment, workload, general skills and, most importantly, teacher-interactivity items – were treated as **perceptual indicators that** capture how learners experience teaching quality and interaction at a given point in time. In a learning-analytics framework, such perceptual indicators can complement behavioural data from learning management systems (e.g., logins, participation in activities, time on resources) by adding a "perceptual layer" that reflects the learner's viewpoint (Siemens & Baker, 2012; Ifenthaler & Yau, 2020). The present study, therefore, not only analyses these scales using traditional statistical techniques but also explores how they could be transformed into candidate indicators for teacher-facing dashboards and early-warning rules in future analytics-enhanced implementations.

Lastly, the adapted instrument thus combines established SEEQ dimensions of teaching quality with newly developed items tailored to online adult learning environments. Furthermore, in this study, the adapted

instrument is used in an exploratory manner. While it provides a structured framework for capturing adult learners' perceptions of teaching quality and interactivity, the measurement properties of the adapted scales should be regarded as preliminary, and interpretations of the results are made with this limitation in mind.

6 Results

All analyses were conducted using standard procedures for quantitative questionnaire data. Descriptive statistics were first computed for all scales and key items of the modified SET instrument. For each satisfaction-related dimension (good teaching, clear goals and standards, assessment, workload, general skills) and for the global items (overall satisfaction, overall instructor rating, overall rating of online tools), means and standard deviations were calculated to summarise central tendency and variability. As described in Methodology, overall learner satisfaction was calculated as the arithmetic mean of seven scales that measure different aspects of satisfaction.

Descriptive results showed generally high perceived teaching quality and satisfaction: **good teaching** had the highest mean (M = 4.49, SD = 0.72), followed by **clear goals and standards** (M = 4.43, SD = 0.93) and **assessment** (M = 4.25, SD = 0.88), while **workload** (M = 3.65, SD = 0.60) and **general skills** (M = 3.53, SD = 0.98) were rated somewhat lower but still above the scale midpoint. The **overall instructor rating** (M = 3.95, SD = 0.55) and **overall rating of online tools** (M = 3.74, SD = 1.14) indicated generally positive, though more varied, perceptions of the instructor and technological environment, and the **composite overall satisfaction score** confirmed a high satisfaction level (M = 4.18, SD = 0.55).

To test H1 (that adult learners report a significantly high level of satisfaction), satisfaction ratings were categorised (2, 3, 4, 5) and a chi-square test was used to examine whether the distribution differed from a reference distribution. The analysis confirmed H1 as it showed a statistically significant deviation, with ratings of 4 and 5 occurring much more frequently (N = 140) than ratings of 3 (N = 19) or 2 (N = 2), $\chi^2(3) = 109.48$, $p < .001$, supporting the conclusion that satisfaction was significantly high. The distribution of satisfaction categories and chi-square test results is presented in Table 1.

Table 1. Chi-square test results

Participant satisfaction	Observed N	X^2	p value
2	2	109.484	.000
3	19		
4	89		
5	51		

To address H2 (associations between teacher interactivity and satisfaction), Pearson correlations were computed between overall satisfaction and each **instructor interactivity** item. All correlations were statistically significant at $p < .001$ thus confirming H2. The strongest positive associations were observed for "The teacher motivates participants" (r = .645), "The teacher provides constructive feedback" (r = .642) and "Feedback on tests or short quizzes is valuable" (r = .661), indicating that motivation and feedback quality are closely linked to satisfaction. Medium to large positive correlations were also found for availability for consultations (r = .442), initiating discussions and encouraging participation (r = .487), satisfaction with the frequency of communication (r = .500), skill in using online tools (r = .412) and the use of interactive tools and methods (r = .352). The only negatively keyed item, "The teacher responds only to the most essential questions of participants", showed a medium negative correlation with satisfaction (r = −.322), suggesting that minimal responsiveness is associated with lower satisfaction. Pearson correlations between all teacher-interactivity indicators and overall satisfaction are reported in Table 2.

Table 2. Correlation matrix (interactivity vs. satisfaction)

Item	Correlation with Satisfaction	p-value	N
Teacher motivates participants.	.645**	.000	161
Teacher provides constructive feedback.	.642**	.000	161
Feedback on tests or short quizzes is valuable.	.661**	.000	161
Teacher responds only to the most essential questions of participants.	-.322**	.000	161
Teacher is regularly available for additional consultations.	.442**	.000	161
Teacher regularly initiates discussions and encourages participation.	.487**	.000	161
Satisfied with the frequency of communi	.500**	.000	161

cation with the teacher.			
Teacher is skilled in using online tools and platforms.	.412**	.000	161
During classes we use interactive tools and methods (e.g., group projects, quizzes)	.352**	..	161

To test H3 (predictive power of teacher interactivity aspects), a multiple linear regression analysis was conducted with overall satisfaction as the dependent variable and five interactivity indicators as predictors: teacher motivates participants, teacher provides constructive feedback, teacher is regularly available for consultations, teacher regularly initiates discussions and encourages participation, and satisfaction with the frequency of communication.

The choice of five interactivity indicators as predictors in the primary regression model was based on theoretical and practical considerations. Specifically, teacher motivation, constructive feedback, availability, discussion facilitation and communication frequency were considered to represent the key interpersonal dimensions of teacher presence. The remaining items were retained as complementary descriptive indicators of online teaching practice and were therefore analysed through descriptive statistics and correlation analyses.

Multiple linear regression was conducted to examine the extent to which selected indicators of instructor interactivity are statistically associated with overall satisfaction with distance teaching. Considering the exploratory nature of this study and the constraints of the available dataset, the formal diagnostic tests of regression assumptions, such as normality of residuals, homoscedasticity, linearity, or multicollinearity, were not conducted. The regression results are interpreted with caution, and the estimated regression coefficients should be viewed primarily as descriptive indicators of the relative strength of associations between interactivity indicators and overall satisfaction.

The model showed a strong overall fit, $R = .779$, $R^2 = .607$, adjusted $R^2 = .594$, $F(5, 155) = 47.91$, $p < .001$, indicating that approximately 59.4% of the variance in satisfaction was explained by the set of interactivity predictors and thus confirming H3. Table 3 shows the regression model summary and Table 4 the ANOVA results.

Table 3. Regression model summary

Model	R	R^2	Adjusted R^2	Std. error of the estimate
1	.779	.607	.594	.34992

Table 4. ANOVA table of regression analysis

Model		Sum of Squares	df	Mean Square	F	p-value
1	Regression	29.331	5	5.866	47.908	.000
	Residual	18.979	155	.122		
	Total	48.310	160			

Inspection of individual coefficients revealed that only teacher motivation ($\beta = .411$, $p < .001$) and constructive feedback ($\beta = .207$, $p = .010$) made statistically significant unique contributions, whereas availability ($\beta = .067$, $p = .344$), initiating discussions ($\beta = .127$, $p = .126$) and satisfaction with communication frequency ($\beta = .088$, $p = .367$) were not significant when considered alongside the other variables. This pattern suggests that, although many aspects of interactivity correlate positively with satisfaction, **the quality of motivation** and **constructive feedback** are the most critical predictors when examined jointly. Table 5 details the regression coefficients for each interactivity predictor.

Table 5. Regression coefficients

Coefficients						
		Unstandardized Coefficients	Standardized Coefficients		t	p-value
Model		B	Std. Error	Beta		
1	(Constant)	1.859	.167		11.116	.000
	Teacher motivates participants.	.230	.043	.411	5.323	.000
	Teacher provides constructive feedback.	.136	.052	.207	2.615	.010
	Teacher is regularly available for additional consultations.	.040	.042	.067	.948	.344
	Teacher regularly initiates discussions and encourages participation.	.069	.045	.127	1.540	.126
	Satisfied with the frequency of communication	.048	.053	.088	.906	.367

	with the teacher.					

7 Discussion

7.1 Interpretation of Key Findings

The results indicate a consistently high level of satisfaction with online teaching among adult learners, with overall satisfaction averaging above 4 on a 5-point scale and the "good teaching" dimension receiving the highest ratings. These findings are in line with recent large-scale studies showing that when teaching presence is strong – particularly in terms of organisation, facilitation and feedback – students report positive online learning experiences and satisfaction (Li & Wang, 2024). In this sample, learners perceived instructors as well prepared, clear and organised, which likely contributed to the positive overall evaluations and aligns with the dimensions emphasised in SEEQ-based evaluations (Marsh, 1982, 1984).

Additionally, the findings contribute to recent discussions in learning analytics research by showing that perceptual indicators derived from SET instruments capture dimensions of learning experience that behavioural traces alone do not entirely clarify. Since learning analytics systems mostly rely on clickstream data, attendance metrics and activity logs, affective and perceptual dimensions of learning may remain underrepresented (Bergdahl et al., 2024). These findings hence support the argument that students' perceptions of teaching quality serve as an important data source within learning analytics systems.

The correlation and regression analyses highlight the central role of teacher motivation and constructive feedback in shaping satisfaction. Large positive correlations between these aspects and satisfaction, combined with their status as the only significant predictors in the regression model, suggest that they are the primary perceptual drivers of how adult learners experience online teaching quality. This pattern resonates with recent empirical work demonstrating that teaching presence, especially facilitation, motivation and timely feedback, is a strong predictor of online satisfaction (Li & Wang, 2024, Pham & Nguyen, 2025), and with studies showing that instructor presence and interactive tools together support engagement and satisfaction (Park & Kim, 2020; Roque-Hernández et al., 2024). Importantly, not all interactivity aspects uniquely predicted satisfaction once motivation and feedback were included in the model: availability for consultations, initiating discussions and communication frequency showed positive bivariate correlations but did not contribute significantly in the multivariate analysis. This suggests that while these aspects are valued, their effects may be mediated through or overshadowed by the perceived quality of motivational climate and feedback, which is informative when prioritising indicators for both teaching practice and learning analytics design.

7.2 Proof-of-Concept Learning Analytics Framework

The SEEQ-based SET instrument used in this study provides a structured set of perceptual indicators that capture how adult learners experience key aspects of teaching, particularly teacher interactivity. In line with calls from the *Handbook of Learning Analytics* to integrate diverse, interpretable indicators into analytics workflows (Lang et al., 2022), these scales can function as a perceptual layer that complements behavioural and performance data typically used in learning analytics (Siemens & Baker, 2012; Papamitsiou & Economides, 2014). Because SEEQ-derived dimensions such as good teaching, clear goals, assessment quality, and especially motivation and feedback are directly interpretable by instructors, they are well-suited as teacher-facing dashboard indicators that translate complex learner experiences into concise, pedagogically meaningful metrics (Marsh, 1984; Daumiller et al., 2021; de Vreugd et al., 2024).

The empirical patterns showing strong predictive power of teacher motivation and constructive feedback can be readily implemented as simple dashboard widgets and rule-based alerts within existing LMS, which already employ traffic-light visualizations for risk indication. For example, Moodle plugins and platforms like IntelliBoard use red-amber-green indicators to flag at-risk learners based on behavioral traces. However, the novelty of this approach lies in applying such familiar traffic-light categorization to perceptual data derived from modified SET questionnaires. A teacher dashboard could display a Motivation Score and Feedback Quality Score (on a 5-point scale) alongside overall satisfaction (see Fig. 1). Courses could be classified using thresholds informed by observed distributions – green (low risk) when both scores are ≥4.0 (indicating high predicted satisfaction), yellow (moderate risk) when either score falls between 3.0 and 3.5 (suggesting targeted improvement), and red (high risk) when both scores drop below 3.0 (triggering flags for pedagogical support). This repurposing of established evaluation instruments as perceptual indicators aligns with prior work integrating survey and behavioural data (Harrati et al., 2016; Herodotou et al., 2025) and guidance on designing interpretable analytics for educators (Lang et al., 2022, Pan et al., 2024), demonstrating how existing SET results can be transformed into actionable, teacher-centered learning analytics tools.

The specific cut-points used in this traffic-light scheme (3.0 and 4.0) are heuristic and intended to provide an intuitive, interpretable visual summary for educators. In the present study these bands were not validated against external outcomes such as course

completion, performance, or retention, so the colour coding should be interpreted as a descriptive aid, and not as a prescriptive standard.

Figure 1. Teacher-facing learning analytics dashboard showing low motivation and feedback scores (yellow/red-risk indicators)

In a more advanced learning analytics implementation, these perceptual indicators could be combined with behavioural data (e.g., declining engagement in forums, reduced attendance in live sessions, lower completion of online activities) to refine early-warning logic. For instance, a course with borderline motivation/feedback scores and a downward trend in engagement metrics could be prioritised for timely support, whereas a course with similar SET scores but stable or improving engagement might require less urgent action. Although such integration lies beyond the scope of the present dataset, the scenarios presented here demonstrate how a theoretically grounded exploratory adaptation of SET dimensions – particularly teacher motivation and constructive feedback – can function as interpretable, teacher-centred signals within future learning analytics dashboards for adult online programs.

8 Limitations and Implications for Future Research

This study was conducted in a single higher-education institution and within a specific context of online adult-education programs, which limits the generalisability of the findings to other institutional settings, disciplines or learner populations. Future research should therefore replicate and extend the study across multiple institutions and diverse program types to examine whether the identified patterns – particularly the prominence of teacher motivation and constructive feedback – hold in different organisational and cultural contexts.

The cross-sectional design and exclusive reliance on self-report data introduce additional limitations. Because all data were collected at one point in time using a questionnaire, the study cannot capture temporal dynamics or establish causal relationships between teacher interactivity and satisfaction.

Even though the questionnaire is created on a well-established framework for student evaluations of teaching, we did not estimate reliability coefficients or examine the factor structure of the adapted satisfaction scales and newly introduced interactivity items. The absence of this evidence means that the dimensional validity and internal consistency of some scales in the adult online education context remain uncertain, and the present findings should therefore be interpreted as an initial, exploratory contribution.

Moreover, SET responses may be affected by social-desirability tendencies, response styles or contextual factors unrelated to teaching quality. These constraints point towards the need for longitudinal designs, the inclusion of behavioural and performance data from LMSs, and the use of more advanced modelling approaches (e.g., predictive or multimodal learning analytics models) that triangulate perceptual indicators with objective traces of learner engagement and outcomes.

A methodological limitation of the present study is the absence of formal diagnostic checks for the assumptions underlying multiple linear regression. As a result, the regression findings should be considered provisional and interpreted as exploratory evidence regarding patterns of association. Future studies should also compare alternative regression models and evaluate potential multicollinearity of conceptually related predictors.

The learning analytics component remains conceptual in the absence of real LMS log data. Therefore, the next step is to collect longitudinal SET data (e.g. mid-course and end-of-course) and align them with time-stamped LMS traces such as forum activity, attendance in synchronous sessions and assignment submissions. This would allow researchers to test whether low perceptual scores on motivation and feedback co-occur with specific behavioural patterns and whether integrated models outperform models based on behavioural or perceptual data alone. In this way, SET can evolve from a static evaluation tool into a dynamic, learner-centred component of multimodal learning analytics pipelines for adult online programs.

9 Conclusion

The study shows that adult learners in online programs report high levels of satisfaction with teaching and that among various aspects of teacher interactivity, teacher motivation and constructive feedback emerge as the strongest statistical predictors of reported satisfaction within the present sample. Considering the cross-sectional nature of the study, these findings reflect predictive associations rather than causal effects. References in this article to “predictors” or to “key areas” for professional development should therefore be understood as aspects of interactivity that are strongly linked to satisfaction in the context of this

study. Conceptually, the work positions modified SET interactivity scales not only as evaluation tools, but as candidate learning analytics indicators, demonstrating how they can be used to derive interpretable teacher-centred metrics and simple rule-based scenarios for dashboards and early-warning rules in adult online programs.

References

Alfredo, R., Echeverria, V., Jin, Y., Yan, L., Swiecki, Z., Gašević, D., & Martinez-Maldonado, R. (2024). Human-centred learning analytics and AI in education: A systematic literature review. *Computers and Education: Artificial Intelligence, 6,* 100215. https://doi.org/10.1016/j.caeai.2024.100215

Bergdahl, N., Bond, M., Sjöberg, J., Dougherty, M., & Oxley, E. (2024). Unpacking student engagement in higher education learning analytics: A systematic review. *International Journal of Educational Technology in Higher Education, 21,* 63. https://doi.org/10.1186/s41239-024-00493-y

Bolliger, D. U., & Martindale, T. (2004). Key factors for determining student satisfaction in online courses. *International Journal on E-Learning, 3*(1), 61–67.

Čotić Poturić, V. (2024). Prediktivne metode analitika učenja i rudarenja obrazovnih podataka u visokom obrazovanju temeljene na algoritmima strojnog učenja. *Zbornik Veleučilišta u Rijeci, 12*(1), 1–23.

Daumiller, M., Grassinger, R., Engelschalk, T., & Dresel, M. (2021). SEEQ-DE: Konstruktion und Überprüfung einer d eutschsprachigen Adaption des Instruments „Student Evaluation of Educational Quality“ (SEEQ; Marsh, 1982, 2007). *Diagnostica, 67*(4), 176–188. https://doi.org/10.1026/0012-1924/a000274

de Vreugd, L., van Leeuwen, A., Jansen, R., & van der Schaaf, M. (2024). Learning Analytics Dashboard Design and Evaluation to Support Student Self-Regulation of Study Behaviour. *Journal of Learning Analytics*, *11*(3), 249-262. https://doi.org/10.18608/jla.2024.8529

Harrati, N., Bouchrika, I., Kamoun, F., & Krichene, S. (2016). Using learning analytics to assess student learning in online courses. *International Journal of Emerging Technologies in Learning, 11*(10), 36–44. https://doi.org/10.3991/ijet.v11i10.5881

Herodotou, C., Shrestha, S., Comfort, C., Andrews, H., Mulholland, P., Bayer, V., Maguire, C., Lee, J., & Fernandez, M. (2025). A Participatory Approach to Designing a Student-Facing Dashboard for Online and Distance Education. *Journal of Learning Analytics*, *12*(2), 158-174. https://doi.org/10.18608/jla.2025.8481

Ifenthaler, D., & Yau, J. Y.-K. (2020). Utilising learning analytics to support study success in higher education: A systematic review. *Educational Technology Research and Development, 68*(4), 1961–1990. https://doi.org/10.1007/s11423-020-09788-z

Lang, C., Siemens, G., Wise, A. F., Gašević, D., & Merceron, A. (Eds.). (2022). *Handbook of learning analytics* (2nd ed.). Society for Learning Analytics Research.

Li, W., & Wang, W. (2024). The impact of teaching presence on students' online learning experience: Evidence from 334 Chinese universities during the pandemic. *Frontiers in Psychology, 15*, 1291341. https://doi.org/10.3389/fpsyg.2024.1291341

Marsh, H. W. (1982). SEEQ: A reliable, valid, and useful instrument for collecting students' evaluations of university teaching. *British Journal of Educational Psychology, 52*(1), 77–95. https://doi.org/10.1111/j.2044-8279.1982.tb02505.x

Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology, 76*(5), 707–754. https://doi.org/10.1037/0022-0663.76.5.707

Papamitsiou, Z., & Economides, A. A. (2014). Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. *Educational Technology & Society, 17*(4), 49–64.

Pan, Z., Biegley , L., Taylor, A., & Zheng, H. (2024). A Systematic Review of Learning Analytics: Incorporated Instructional Interventions on Learning Management Systems. *Journal of Learning Analytics*, *11*(2), 52-72. https://doi.org/10.18608/jla.2023.8093

Park, J., & Kim, J. (2020). Perception of instructor presence and its effects on learning experience in online classes. *Journal of Information Technology Education: Research, 19*, 475–488. https://doi.org/10.28945/4570

Pham, A. T., & Nguyen, T. B. (2025). Students' perceptions of teaching presence in the virtual learning environment: A community of inquiry case study in Vietnam. *Acta Psychologica, 261*, 105810. https://doi.org/10.1016/j.actpsy.2025.105810

Roque-Hernández, R. V., López-Mendoza, A., & Salazar-Hernández, R. (2024). Perceived instructor presence, interactive tools, student

engagement, and satisfaction in hybrid education post-COVID-19 lockdown in Mexico. *Heliyon, 10*, e27342. https://doi.org/10.1016/j.heliyon.2024.e27342

Siemens, G., & Baker, R. S. J. d. (2012). Learning analytics and educational data mining: Towards communication and collaboration. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge (LAK '12)* (pp. 252–254). ACM. https://doi.org/10.1145/2330601.2330661

Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the validity of student evaluation of teaching: The state of the art. *Review of Educational Research, 83*(4), 598–642. https://doi.org/10.3102/0034654313496870

Sun, S.-Y., & Chen, C.-L. (2022). The influence of teacher–student interaction on the effects of online learning: The mediating role of learning engagement and psychological atmosphere. *Frontiers in Psychology, 13*, 779217. https://doi.org/10.3389/fpsyg.2022.779217

Watermark Insights. (2024, September 24). *How online course evaluations help improve instructional course quality*. https://www.watermarkinsights.com/resources/blog/how-online-course-evaluations-help-improve-instructional-course-quality/

Yuan, M., Zhang, Y., & Li, H. (2022). Enhancing teacher–student interactions and student online engagement in an online learning environment. *Frontiers in Psychology, 13*, 983995. https://doi.org/10.3389/fpsyg.2022.983995

Conversational AI for Electrical Measurement Education: An Open-Source LLM Framework

Branislav Fecko, Jozef Dziak, Tibor Vince, Ján Molnár
Technical University of Košice
Department of Theoretical and Industrial Electrical Engineering
Letná 1/9, 042 00 Košice, Slovakia
{branislav.fecko, jozef.dziak, tibor.vince, jan.molnar}@tuke.sk

***Abstract**. Large language models are increasingly used in engineering education, yet their deployment in measurement laboratories remains limited by the lack of grounding in actual instrument data. This paper identifies three scenarios for LLM integration into electrical measurement education, defines seven criteria for their systematic comparison, and presents a functional open-source implementation of Scenario 2. The implemented framework connects two digital multimeters and a programmable power supply to a Custom GPT via a FastAPI backend and Cloudflare Tunnel, grounding conversational feedback in student-measured values.*

Keywords. electrical measurement, laboratory education, instrument integration, conversational AI

1 Introduction

Large language models (LLMs) have been adopted rapidly in higher education. A systematic review of 88 empirical studies confirmed their positive effects on academic performance and student engagement, while also highlighting the risks of overreliance and the need for longitudinal evaluation (Shi et al., 2026). In engineering education, these tools support dynamic content generation and personalized explanations.

Students naturally turn to AI tools because of their immediate availability and the ability to ask questions at their own pace (Shi et al., 2026). However, growing interest in the use of LLMs in laboratory courses suggests that this potential can also be extended to the practical component of instruction (Wang, 2026). As class sizes grow and individualized instructor feedback becomes less available, students naturally turn to LLMs as a source of assistance (Hasan et al., 2025). A key limitation, however, is that LLMs generate incorrect or misleading responses in more than one out of ten cases when the experimental context is not explicitly provided (Kregear et al., 2025). Unlike an instructor, an LLM cannot infer the course of an experiment from the surrounding situation and therefore requires the experimental context to be explicitly included in the prompt. One solution is to ground the LLM automatically in values measured directly by laboratory instruments, supplemented with relevant theoretical background and procedural guidance. This principle forms the core of the proposed system.

This paper addresses three questions: What are the main scenarios for integrating LLMs into electrical measurement education? What criteria enable their systematic comparison? And what architectural design can connect laboratory instruments with a conversational LLM to provide data-grounded feedback? The paper also presents a functional open-source implementation of one of these scenarios. Quantitative pedagogical validation with students is beyond the scope of the present study and is planned for the second phase of the project.

The remainder of the paper is organized as follows. Section 2 reviews the related literature and formulates the research questions. Sections 3 and 4 identify and compare the proposed integration scenarios. Section 5 describes the architectural design and implementation of the framework. Finally, Sections 6 and 7 discuss the results and conclude the paper.

2 Background and Related Work

The proposed system lies at the intersection of three domains: the application of generative AI in engineering education, laboratory automation, and efforts to integrate LLMs with physical laboratory instruments. The convergence of these domains reveals a research gap that this paper seeks to address.

2.1 Generative AI in Engineering Education

Generative AI is making its way into engineering and STEM education in several forms. The first is the direct support of conceptual learning. For example, in thermodynamics, ChatGPT-assisted instruction led to

measurable improvements in student understanding compared to traditional teaching approaches (El Fathi et al., 2025). The second is the enhancement of learning materials themselves. The concept of intelligent textbooks uses AI methods to enrich digital content with interactive modules, automated assessment, embedded chatbots, and adaptive student-content interaction (Sosnovsky et al., 2025).

A specific example of such an intelligent textbook in electrical engineering education is the GenAI-T system, built on GPT-4 and integrated into an introductory electrical engineering course. A four-year comparative study involving 736 students demonstrated improved performance in formative assessments (Fecko et al., 2026). Efforts to extend the use of LLMs beyond self-directed learning and into formal instruction have also reached laboratory education. In a quasi-experimental study, the integration of LLMs into an undergraduate laboratory course improved students' target competency, while also revealing several practical challenges associated with their implementation (Wang, 2026).

Across these approaches, AI operates exclusively at the level of text and pre-prepared content. It works with what students write or read, rather than with what they actually measure. As a result, these solutions lack feedback grounded in the real measurement values produced by the laboratory instruments.

2.2 Instrument Control and Automation in Measurement Laboratories

Software-based control of laboratory instruments is a well-established area in engineering practice. Commercial tools such as LabVIEW and MATLAB are powerful and widely used, but they are also expensive and difficult to transfer across different instruments and platforms (Letowski et al., 2020). These limitations have motivated a shift toward open-source solutions. For example, the remote laboratory project Laborem was originally built on National Instruments' LabVIEW platform but was later redesigned as a lightweight open-source system due to data communication issues (Letowski et al., 2020). Similarly, the high cost of commercial solutions motivated the development of a fully open-source remote laboratory based on Raspberry Pi and Python and support for real-time web-based control (Álvarez & Nomesqui, 2023).

Open-source approaches rely almost exclusively on Python and its ecosystem. For example, measurement automation implemented in Python has been shown to improve the reproducibility of experimental results by continuously recording all relevant information about the experiment (Koerner et al., 2019). Instrument communication is typically based on the Standard Commands for Programmable Instruments (SCPI) protocol and the PyVISA library, which provides a unified Python interface for instruments from different manufacturers (Koerner et al., 2019,Fébba et al., 2024). Libraries such as NumPy and SciPy then provide tools for the subsequent analysis of measurement data.

From an educational perspective, a meta-analysis found that virtual and automated laboratories contribute significantly to student learning outcomes, particularly by increasing motivation and engagement. However, they cannot fully replace hands-on work with physical instruments (Li & Liang, 2024). The combination of low cost, reproducibility, and vendor-independent instrument access therefore provides a mature technological foundation. Operating such systems still requires programming expertise, and the measured data remains largely devoid of pedagogical interpretation. Integrating this infrastructure with a conversational LLM would enable not only automated measurement execution but also immediate feedback on measured results within the context of the ongoing laboratory exercise.

2.3 AI-Driven Instrument Control and Conversational Laboratory Assistants

The integration of LLMs with experimentation is a rapidly evolving research area. The GPT-4–driven Coscientist system demonstrated the ability to autonomously design, plan, and execute complex experiments, by leveraging tools such as documentation retrieval, code execution, and laboratory automation (Boiko et al., 2023). Traditionally, developing control software for existing laboratory instruments has required substantial programming expertise, creating a barrier for researchers and educators without a computational background. LLMs have the potential to significantly lower this barrier (Xie et al., 2025).

This shift is increasingly reflected in laboratory measurement systems. Through interactions with ChatGPT-4, researchers developed a control interface for the Keithley 2400 source-measure unit, a widely used instrument for current–voltage (I–V) measurements, with only minimal manual code revisions. The resulting framework included both a graphical user interface and prompt templates for additional instruments (Fébba et al., 2024). The LABIIUM system extends this concept by integrating an LLM-based code generation assistant with Lab-Automation-Measurement Bridges, enabling zero-configuration instrument connectivity through standard tools such as Python and VSCode (Olowe & Chitnis, 2025). Remotely accessible platforms have also emerged as environments for prototyping automation and LLM workflows within self-driving laboratories (Kitchin, 2025).

These studies connect LLMs with laboratory instruments, yet their primary focus remains code generation and autonomous experimentation rather than pedagogically oriented feedback. The Adaptive Lab Mentor framework represents a notable exception by providing real-time personalized feedback through sensor fusion and Internet of Things (IoT)

technologies. Even in this case, however, the conversational language model is proposed only as a future extension of the framework (Hasan et al., 2025). Evidence from a physics laboratory setting also suggests that the reliability of feedback declines when it is not grounded in data describing the measurement process (Kregear et al., 2025).

2.4 Research Gap

The literature review reveals that these components currently exist in isolation: pedagogically effective AI without access to measured data (2.1), instrument automation without a conversational layer (2.2), and LLM–instrument integration without a pedagogical focus (2.3). The Adaptive Lab Mentor framework comes closest to a comprehensive solution by combining real-time diagnostics with pedagogical feedback. Its conversational language model component, however, remains a proposed direction for future development. Based on the available literature, no existing system integrates a conversational LLM with physical measurement instruments to provide pedagogically oriented feedback specifically for electrical measurement education. These components can also be combined in multiple ways, differing primarily in the extent to which feedback is grounded in real measurement data and in the pedagogical transparency of the interaction. These observations give rise to three research questions:

- RQ1: What are the main scenarios for LLM integration into electrical measurement laboratory education and how do they differ in terms of data grounding and pedagogical transparency?
- RQ2: What criteria enable systematic comparison of LLM integration scenarios in the context of hands-on engineering education?
- RQ3: What architectural design connects laboratory instruments with a conversational LLM to deliver data-grounded student feedback?

3 LLM Integration Scenarios in Measurement Education

Three scenarios for integrating LLMs into electrical measurement laboratory education were identified and are described in terms of their degree of data grounding and pedagogical transparency. The scenarios are derived from two sources. The first is the literature review presented in Chapter 2, which identified three distinct layers: pedagogically oriented generative AI, instrument automation, LLM–instrument integration. The second is an informal analysis of student needs in the target course, conducted through a voluntary anonymous survey with open-ended questions about the challenges students encounter during laboratory measurements and how AI could support them. The purpose of this data collection was not to test hypotheses but to ground the scenario design in authentic user needs. A rigorous evaluation involving students is planned for the second phase of the project.

Three recurring categories of needs emerged from the survey responses. Students commonly rely on general-purpose conversational AI tools without providing structured measurement context, manually entering measured values into the chat or attaching photographs of circuit diagrams. At the same time, they repeatedly expressed a need for support during the measurement process itself, including verification of measured values, explanations of why a circuit behave as expected or unexpectedly, and assistance in identifying potential errors. Several students explicitly stated that they wanted the assistant to explain the procedure rather than simply provide the answer. This expectation highlights a fundamental pedagogical tension shared across all three scenarios. A higher degree of automation can improve measurement efficiency, yet it may also reduce the student's role to that of a passive observer. The goal of laboratory education is not merely to obtain a measured value but to develop an understanding of the measurement process. For this reason, the extent to which a system preserves students' active cognitive engagement is as important a dimension of comparison as its level of technical integration with laboratory instruments.

Based on the literature review and the identified student needs, seven criteria were defined for the systematic comparison of integration scenarios: context grounding, pedagogical transparency, degree of automation, deployment complexity, wiring error detection, longitudinal learning analytics, and instructor analytics. Three scenarios are analyzed in this paper: unstructured LLM use with manual data entry (3.1), an instrument-integrated LLM providing context-aware feedback (3.2), and an autonomous laboratory framework with interaction logging and adaptive analytics (3.3). Each scenario is illustrated by a block diagram showing the flow of information between the student, the laboratory instrument, and the language model.

3.1 Scenario 1 - Unstructured LLM Use

The first scenario represents the way LLMs are currently used by students in laboratory courses, as also reflected in the informal needs analysis. In this scenario, the laboratory instrument is not connected to the system. Measurements are performed independently, and the measured values are read directly from the instrument display and manually entered into a general-purpose conversational AI tool, often together with a photograph of the circuit diagram. The language model generates its response solely from the information provided by the student. The flow of information is illustrated in Figure 1.

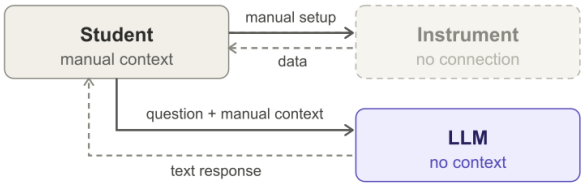


Figure 1. Unstructured LLM Use

3.2 Scenario 2 - Instrument-Integrated LLM via OpenAI Actions

The second scenario represents the main contribution of this paper. Students interact directly with a Custom GPT based on GPT-4. When a question is submitted, the Custom GPT invokes the appropriate backend endpoint through an OpenAI Action. The backend is implemented as a FastAPI application using the PyVISA library and serves as an intermediary between the language model and the laboratory instrument. The server runs locally and is exposed through a Cloudflare Tunnel under a public address, enabling communication with OpenAI servers without requiring a public IP address. Measurements are executed through SCPI commands, after which the instrument returns the measured values. These values are then passed back to the Custom GPT as part of the context of the ongoing conversation.

Before the conversation begins, a system prompt is loaded into the Custom GPT from a configuration block. This block contains the theoretical background of the course, instructions for the laboratory exercise, and guidance directing the model to ask students guiding questions rather than provide direct answers. The Custom GPT therefore generates responses grounded in the values measured by the student and returns them as part of the ongoing interaction. The flow of information is illustrated in Figure 2.

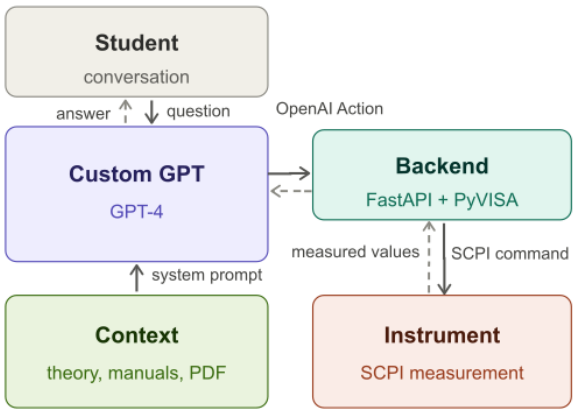


Figure 2. LLM via OpenAI Actions

3.3 Scenario 3 - Autonomous AI Laboratory Framework

The third scenario extends Scenario 2 with a longitudinal dimension and serves as a conceptual framework for the second phase of the research project. The entire system is encapsulated within a dedicated web application through which students submit questions and receive responses. As in Scenario 2, the backend layer provides programmatic communication with laboratory instruments through SCPI. Unlike Scenario 2, however, communication with the language model is handled directly through an API interface, enabling the deployment of a locally hosted model without dependence on external services.

A central component of this scenario is a database that serves two functions. The first is Interaction Logging, where every exchange between the student and the language model is recorded together with the measured values and their deviation from the expected results. The second is the storage of contextual resources and configuration data, including system prompts, theoretical materials, laboratory instructions, and model settings. Before each conversation, these resources are provided to the language model as a system prompt. Model responses are returned to the student through the backend layer. The recorded logs remain accessible to the backend and are used to generate adaptive feedback for students as well as aggregated reports for instructors. Empirical validation of this scenario is planned for the second phase of the project. The flow of information is illustrated in Figure 3.

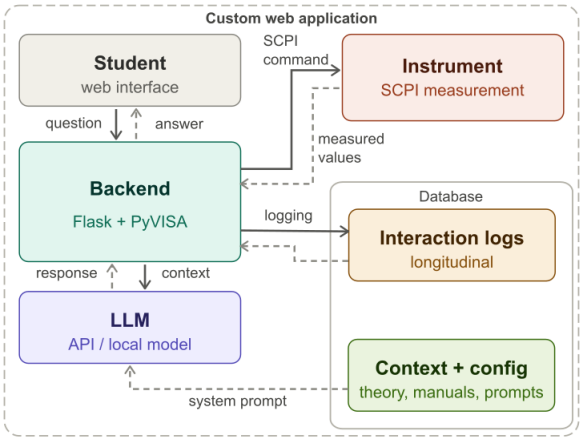


Figure 3. Autonomous AI Laboratory Framework

4 Comparative Analysis

Based on the literature review and the analysis of student needs, seven criteria were defined for comparing the proposed scenarios: Data Grounding, Pedagogical Transparency, Degree of Automation, Deployment Complexity, Wiring Error Detection, Longitudinal Learning Analytics, and Instructor Analytics. The first two criteria directly address RQ1 and are derived from the requirements identified in 3. The remaining five reflect the practical constraints and requirements of deployment in a university laboratory environment.

Scenario 1 requires no additional infrastructure and reflects current informal practice. The quality and relevance of feedback depend entirely on the context provided by the student, whether in the form of text, a circuit diagram image, or manually entered measurement values.

Scenario 2 addresses Data Grounding through an SCPI/PyVISA backend and provides partial support for Wiring Error Detection, at the cost of moderate Deployment Complexity. Before the conversation begins, contextual information is loaded into the Custom GPT through a system prompt containing the theoretical background of the course and instructions for the laboratory exercise. The prompt also directs the model to ask guiding questions rather than provide direct answers, thereby preserving the student's active role in the measurement process. Dialogue history is retained within a single session, enabling context-aware follow-up questions but not longitudinal tracking of student progress.

Scenario 3 extends Scenario 2 by introducing long-term Interaction Logging and Instructor Analytics. This added functionality requires a more robust server infrastructure. The recorded interactions make it possible to identify recurring errors across an entire student cohort, creating opportunities for adaptive instruction.

Table 1. Comparison of LLM Integration Scenarios

Criterion	S1	S2 ★	S3
Data grounding	Partial	**Full**	Full
Pedagogical transparency	Low	**Moderate**	High
Degree of automation	None	**Full**	Full
Deployment complexity	None	**Moderate**	High
Connection error detect.	None	**Partial**	Full
Longitudinal analytics	None	**Partial**	Full
Instructor analytics	None	**None**	Full

5 Framework Design and Implementation - Scenario 2

The implemented framework realizes Scenario 2 described in 3.2 and directly addresses RQ3. Its primary pedagogical value lies not in measurement automation itself, but in the dialogue-driven configuration of measurements grounded in an understanding of the physical phenomena occurring within the investigated circuit. The framework is demonstrated through the measurement of the I–V characteristic of an unknown component. The task involves identifying a black box selected from a set of three components with distinct electrical characteristics: a resistor, a diode, and a capacitor.

5.1 System Architecture

The framework is organized into three layers. The physical layer consists of two Rigol DM3068 digital multimeters and a Hameg HMP4040 programmable power supply connected to a host computer via USB. The data layer is implemented as a FastAPI-based backend server responsible for instrument control, measurement configuration management, and the provision of measurement-specific endpoints for the AI layer. The server runs locally and is securely exposed through a Cloudflare Tunnel under a public address, enabling communication with OpenAI servers without requiring a public IP address or port forwarding. The AI layer is implemented as a GPT-4–based Custom GPT that communicates with the backend through OpenAI Actions. Through a sequence of guiding questions, the system encourages students to interpret and understand the measured results rather than merely receive them. The flow of information across all three layers is illustrated in Figure 4.

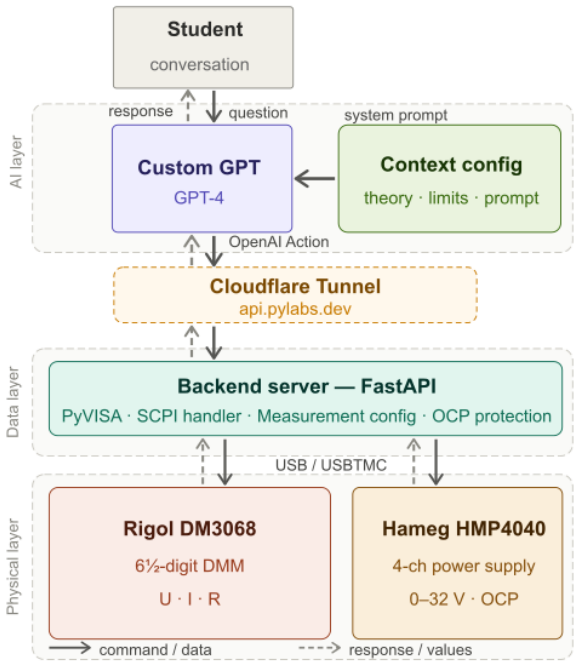


Figure 4. Three-Layer Framework Architecture

The framework is designed with extensibility in mind. Each laboratory instrument is represented by a dedicated driver, while each measurement task is implemented through a separate configuration file and FastAPI endpoint. As a result, new measurement tasks can be added without modifying the core architecture.

5.2 Instrument Communication Layer

The Hameg HMP4040 programmable power supply and two Rigol DM3068 digital multimeters are controlled via the PyVISA library using the USBTMC

protocol. One multimeter measures voltage, while the other measures current. Each instrument is identified at server startup using a standard SCPI query, which verifies availability and correct assignment before any measurement is performed.

Before each measurement, the backend configures the overcurrent protection of the HMP4040 power supply according to the limits defined for the active measurement task. This hardware-level protection prevents damage to the component under test regardless of the course of the dialogue. The sweep sequence used for measuring the I–V characteristic is implemented as a FastAPI endpoint (Listing 1).

```
@app.post("/measure_va_characteristic")
async def measure_va_characteristic(
        params: MeasurementParams):
    psu.write(f":CURR {params.i_limit}")
    psu.write(":OUTP ON")
    dmm_v.write(":FUNC:VOLT:DC")
    dmm_i.write(":FUNC:CURR:DC")
    results = []
    for v in params.voltage_steps:
        psu.write(f":VOLT {v}")
        await asyncio.sleep(params.step_interval)
        U = float(dmm_v.query(":READ?"))
        I = float(dmm_i.query(":READ?"))
        results.append({"U": round(U, 4), "I":
round(I, 6)})
    psu.write(":OUTP OFF")
    return {"points": results,
            "config": params.model_dump()}
```

Listing 1. I–V characteristic sweep, FastAPI endpoint

5.3 Measurement Configuration and Guided Execution

Each measurement task is defined through a configuration file prepared by the instructor. The file contains a measurement identifier, instrument safety limits, a system prompt for the Custom GPT, and reference values used for anomaly detection.

The measurement workflow consists of three phases. In the first phase, students identify the task through a natural-language request. The Custom GPT loads the corresponding configuration and guides them in defining the measurement parameters in the context of the physical properties of the component under investigation. This phase represents the pedagogical core of the system: students actively make methodological decisions, developing an understanding of the underlying physical phenomena before the measurement begins. In the second phase, the backend verifies the instrument configuration through SCPI queries and prompts the student to make corrections if inconsistencies are detected. In the third phase, the backend executes the measurement and returns the results as a structured JSON object.

The following example illustrates the first phase of the workflow for measuring the I–V characteristic of an unknown component:

***Student**: "I want to identify an unknown component by measuring its voltage-current characteristics."*
***Custom GPT**: "What current limit will you choose, and why?"*
***Student**: "I'll set it to 10mA as a safe limit."*
***Custom GPT**: "What voltage range and step size do you propose?"*
***Student**: "0 to 2V, 0.1V steps—for an initial exploration of the characteristics."*
***Custom GPT**: "Measurement complete. The measured data points are available in Table 2."*
***Student**: "Exponential curve — looks like a diode. I'll run a new sweep from 0 to 1 V, step 0.02 V, limit 15 mA to verify the threshold voltage."*

The JSON payload returned by the backend and processed by the Custom GPT is shown in Listing 2.

```
{
  "measurement": "VA characteristic",
  "parameters": {
    "U_max": 2.0,
    "step": 0.1,
    "i_limit": 0.010
  },
  "points": [{"U": 0.0, "I": 0.000},
             "...",
             {"U": 2.0, "I": 0.008}],
  "analysis": {
    "threshold_voltage": 0.62,
    "type_hint": "exponential"
  }
}
```

Listing 2. JSON payload returned to Custom GPT

5.4 Open-Source Availability

The Scenario 2 implementation will be made publicly available under the PolyForm Non-Commercial License 1.0.0 upon acceptance. Educational and research use is unrestricted; commercial use requires a separate license agreement. The repository link is omitted to preserve double-blind review. Scenario 3 will be released in a subsequent research phase.

6 Discussion

The three research questions form a progression from the conceptual to the technical level. A common theme runs through all three: to what extent should the system take over tasks traditionally performed by students and where should space remain for their own decision-making. The findings suggest that this boundary is not technical but pedagogical. The system should automate data collection, not thinking.

6.1 LLM Integration Scenarios

The three scenarios differ in the extent to which feedback is automatically grounded in measurement data, in how effectively they preserve students' active cognitive engagement, and in the reliability of the model's output. Reliability depends directly on whether the context is provided by the instructor or assembled independently by the student (Kregear et al., 2025). A higher degree of instrument integration does not necessarily translate into greater pedagogical value. A system that performs measurements automatically and simply returns the result may replace the instructor without contributing to the development of student

competence. The added value of Scenario 2 therefore lies not in automated data collection, but in requiring students to actively determine measurement parameters before the experiment and justify their choices in the context of the physical properties of the component under investigation.

Greatest benefits emerge in measurement tasks where time-consuming automated data collection directly illustrates fundamental principles of electrical engineering. Examples include the nonlinear I–V characteristic of an incandescent lamp, which demonstrates the relationship between resistance and filament temperature, and transient phenomena during capacitor charging. In such cases, the conversational assistant transforms raw measurement data into a pedagogical dialogue that would be impractical to sustain through manual data collection alone. The level of system autonomy is not fixed. It can range from step-by-step verification of instrument settings through SCPI commands before measurement execution to fully automated data collection followed by result interpretation. This flexibility enables multiple implementation approaches tailored to the objectives of a particular course and student cohort.

6.2 Criteria for Systematic Comparison

Systematic comparison of scenarios in hands-on laboratory education requires criteria that capture not only the technical level of integration but also the pedagogical quality of the interaction. Criteria derived exclusively from the literature on LLM-based tutors are insufficient in this context because they overlook the physical measurement layer. It is precisely this layer that introduces two criteria absent from existing frameworks: Data Grounding in real measurement data and Wiring Error Detection. Deviations between measured and expected values make it possible to identify issues such as instrument misconfiguration or an open circuit without requiring visual input. Deployment Complexity also distinguishes Scenario 2 from Scenario 3. Scenario 2 depends on OpenAI's Custom GPT platform, including its availability and service conditions. Scenario 3 removes this dependency through direct API-based integration, at the cost of increased infrastructure requirements.

6.3 Architectural Design of the Proposed Framework

Integrating laboratory instruments with a conversational LLM requires an architecture that separates the physical measurement layer from the AI interaction layer while supporting extensibility without modifications to the system core. The proposed three-layer architecture, built around measurement-specific endpoints, satisfies these requirements and enhances operational safety by preventing students from interacting directly with low-level SCPI commands. The demonstrated use case of identifying an unknown component serves as a deliberate proof of concept that validates all layers of the system while minimizing hardware dependencies. The same architecture can be scaled to measurement tasks involving multiple instruments, where automation offers substantially greater benefits. Empirical validation of Scenario 3 is planned as part of the next phase of the research project.

7 Conclusion and Future Work

Three scenarios for integrating LLMs into electrical measurement education were identified, seven comparison criteria were defined, and an open-source implementation of Scenario 2 was presented. The framework connects two Rigol DM3068 digital multimeters and a Hameg HMP4040 programmable power supply to a GPT-4–based Custom GPT through a FastAPI backend, enabling conversational feedback grounded in student-generated measurement data.

The primary contribution of the proposed framework is not measurement automation itself, but the dialogue-driven configuration of measurement parameters prior to experiment execution. Students' active cognitive engagement is preserved by encouraging them to justify methodological decisions in the context of the physical properties of the component under investigation. The demonstrated use case of identifying an unknown component through measurement of its I-V characteristic validates all layers of the system and illustrates its potential for extension to more complex diagnostic tasks, where automation offers substantially greater benefits than manual data collection. Wiring Error Detection was also included among the comparison criteria. This criterion is systematically absent from existing frameworks for LLM-based tutoring systems and emerged directly from practical experience gained during implementation.

The current implementation has two main limitations. Dependence on OpenAI’s Custom GPT platform introduces operational risks related to service availability and platform policies. In addition, quantitative pedagogical validation involving students enrolled in the target course has not yet been conducted.

Future research will focus on three areas: empirical validation with students, implementation of Scenario 3 with support for a locally hosted language model, and expansion of the measurement portfolio to include tasks requiring coordination among multiple laboratory instruments.

Acknowledgments

Supported by the Cultural and Educational Grant Agency of the Ministry of Education, Science,

Research and Sport of the Slovak Republic (KEGA) under the project No. 038TUKE-4/2025.

References

Álvarez Ariza, J., & Nomesqui Galvis, C. (2023). RaspyControl Lab: A fully open-source and real-time remote laboratory for education in automatic control systems using Raspberry Pi and Python. *HardwareX*, *13*. https://doi.org/10.1016/j.ohx.2023.e00396

Boiko, D. A., MacKnight, R., Kline, B., & Gomes, G. (2023). Autonomous chemical research with large language models. *Nature*, *624*(7992), 570–578. https://doi.org/10.1038/s41586-023-06792-0

El Fathi, T., Saad, A., Larhzil, H., Lamri, D., & Al Ibrahmi, E. M. (2025). Integrating generative AI into STEM education: enhancing conceptual understanding, addressing misconceptions, and assessing student acceptance. *Disciplinary and Interdisciplinary Science Education Research*, *7*(1). https://doi.org/10.1186/s43031-025-00125-z

Fébba, D., Egbo, K., Callahan, W. A., & Zakutayev, A. (2024). From text to test: AI-generated control software for materials science instruments. *Digital Discovery*, *4*(1), 35–45. https://doi.org/10.1039/d4dd00143e

Fecko, B., Dziak, J., Vince, T., & Molnár, J. (2026). Generative AI interactive textbook in electrotechnics: A four-year comparative study on student performance and inclusion. *Computers and Education: Artificial Intelligence*, *10*. https://doi.org/10.1016/j.caeai.2026.100611

Hasan, M. S., Ahmed, A., Rasool, N., Jabe, M. M. N., Zeng, X., & Ali Pirzado, F. (2025). The Adaptive Lab Mentor (ALM): An AI-Driven IoT Framework for Real-Time Personalized Guidance in Hands-On Engineering Education. *Sensors*, *25*(24). https://doi.org/10.3390/s25247688

Kitchin, J. R. (2025). The evolving role of programming and LLMs in the development of self-driving laboratories. *APL Machine Learning*, *3*(2). https://doi.org/10.1063/5.0266757

Koerner, L. J., Caswell, T. A., Allan, D. B., & Campbell, S. I. (2019). A Python Instrument Control and Data Acquisition Suite for Reproducible Research. In *IEEE TRANSACTIONS ON INSTRUMENTATION AND MEASUREMENT*.

Kregear, T., Babayeva, M., & Widenhorn, R. (2025). Analysis of Student Interactions with a Large Language Model in an Introductory Physics Lab Setting. *International Journal of Artificial Intelligence in Education*, *35*(5), 2993–3016. https://doi.org/10.1007/s40593-025-00489-3

Letowski, B., Lavayssière, C., Larroque, B., Schröder, M., & Luthon, F. (2020). A fully open source remote laboratory for practical learning. *Electronics (Switzerland)*, *9*(11), 1–16. https://doi.org/10.3390/electronics9111832

Li, J., & Liang, W. (2024). Effectiveness of virtual laboratory in engineering education: A meta-analysis. *PLoS ONE*, *19*(12). https://doi.org/10.1371/journal.pone.0316269

Olowe, E. A., & Chitnis, D. (2025). *LABIIUM: AI-Enhanced Zero-configuration Measurement Automation System*. https://doi.org/10.1109/I2MTC62753.2025.11079200

Shi, Y., Yu, K., Dong, Y., & Chen, F. (2026). Large language models in education: a systematic review of empirical applications, benefits, and challenges. In *Computers and Education: Artificial Intelligence* (Vol. 10). Elsevier B.V. https://doi.org/10.1016/j.caeai.2025.100529

Sosnovsky, S., Brusilovsky, P., & Lan, A. (2025). Intelligent Textbooks. In *International Journal of Artificial Intelligence in Education* (Vol. 35, Number 3, pp. 967–986). Springer. https://doi.org/10.1007/s40593-024-00451-9

Wang, Y. (2026). Integrating large language models into medical undergraduate laboratory course to enhance bioethical competence: a quasi-experimental study. *Frontiers in Medicine*, *12*. https://doi.org/10.3389/fmed.2025.1745975

Xie, Y., He, K., & Castellanos-Gomez, A. (2025). Toward Full Autonomous Laboratory Instrumentation Control with Large Language Models. *Small Structures*, *6*(8). https://doi.org/10.1002/sstr.202500173

Review of Learning Outcomes Prioritisation Approaches

Blaženka Divjak, Nikola Kadoić, Barbi Svetec, Dino Vlahek
University of Zagreb Faculty of organization and informatics
Pavlinska 2, 42000 Varaždin, Croatia
{nkadoic, bdivjak, bsvetec, dvlahek}@foi.unizg.hr

Thomas Penniston
University of Maryland, Baltimore County (UMBC)
1000 Hilltop Circle, MD 21250, United States
penniston@umbc.edu

Mirza Žižak
University of Zagreb Faculty of Medicine
Šalata 3, 10000 Zagreb, Croatia
mirza.zizak@mef.hr

Abstract. *This paper reviews approaches to prioritising learning outcomes (LOs) by defining LO prioritisation, identifying existing methods, and evaluating their applicability in educational contexts. The study combines a literature review with a qualitative expert-based analysis of multi-criteria decision-making (MCDM) methods. Of 286 papers analysed, only 12 addressed LO prioritisation and were classified into four categories. Experts proposed and compared six methods: simple additive weighting, analytic hierarchy process, analytic network process, SWING, decision making trial and evaluation laboratory, and social network analysis process. Their applicability and complexity were assessed, demonstrating how LO prioritisation can support learning design and educational alignment.*

Keywords. Learning outcomes, weights, prioritisation, multi-criteria decision-making methods

1 Introduction

In contemporary education systems, courses and study programmes are typically defined through intended learning outcomes (LOs), referring to explicit statements describing what learners are expected to know, understand, or be able to do upon completion of a learning process. The increasing emphasis on outcome-based education, particularly within frameworks such as the Bologna Process (European Commission, 2015), highlights the importance of clearly defined and measurable LOs in curriculum design, implementation, and evaluation.

However, not all LOs are equally important. Some LOs represent core knowledge and competencies essential for achieving the overall purpose of a course, while others play a more supportive or complementary role. Despite this, LOs are often treated as having equal importance in practice, which may lead to suboptimal curriculum design, assessment strategies, and resource allocation. Consequently, prioritising LOs - i.e., determining their relative importance and assigning weights (Divjak et al., 2021) - emerges as a meaningful and necessary process for improving the effectiveness and coherence of teaching and learning. In this paper, prioritisation is operationalised as determining the relative importance of learning outcomes through weighting methods. Consequently, weighting is considered the quantitative implementation of prioritisation rather than an alternative decision task.

This paper addresses the challenge of LO prioritisation by exploring existing approaches and identifying suitable methods for its implementation. First, it provides the findings of a systematic review of the literature conducted to identify how LOs and related constructs such as learning objectives, goals, and competencies have been prioritised in the reviewed research. Second, it presents a qualitative analysis of potential LO prioritisation methods from the perspective of experts in decision-making, recognising that this problem can be framed as a multi-criteria decision-making task. Based on this, we propose a set of applicable methods for LOs prioritization and discuss their characteristics, particularly in terms of complexity and practical usability.

This research is conducted within the framework of the Erasmus+ project Streamlining AI in Learning Reimagination (SAILeR). The project aims to empower educators to enhance the quality of teaching and learning in the era of artificial intelligence by supporting the development of innovative learning design practices, including prioritization of LOs. Within the project, there is also a plan to enhance the Balanced Design Planning (BDP) learning design tool (Divjak et al., 2022) by integrating one of the identified prioritisation methods, taking into account their complexity and the level of understanding among educators who are not experts in multi-criteria decision-making.

2 Background

2.1 Learning Outcome Prioritisation

Learning outcomes (LOs) have a central role in learning and curriculum design. Terminology varies across contexts and learning objectives, learning goals, or learning competencies are also used when describing courses and curricula. These constructs are not synonymous, but in practice they often overlap and are sometimes used interchangeably. However, the central construct in contemporary education is that of LOs, referring to statements describing what learners are expected to know, understand, and be able to do upon completion of a learning process (European Commission, 2015). LOs describe observable and measurable learner achievements, while in competence-based education, competencies integrate knowledge, skills, and attitudes into demonstrable capabilities applied in professional and real-world contexts (Mulder, 2014).

Within a curriculum or a course, not all LOs are equally important. In particular, prior research has demonstrated that LOs differ in their relevance depending on multiple criteria, such as their importance for future professional practice, cognitive level (e.g., based on Bloom's taxonomy), contribution to the development of generic skills, and the required student workload (Divjak et al., 2021). This leads to LO prioritisation, defined as the process of determining the relative importance of LOs within a course or curriculum. This process typically results in assigning weights to LOs, where LO weight represents a quantitative or qualitative measure reflecting the importance of an LO relative to other LOs and its relative contribution to the overall educational goals.

The need for prioritisation has been recognised in prior research on assessment validity and learning design (Divjak et al., 2023), showing that determining LO weights is a prerequisite for constructing valid assessment plans and ensuring constructive alignment between LOs and assessment tasks. Furthermore, recent research in learning design has emphasised that effective course design requires considering the relative importance of LOs and aligning both formative and summative assessment with the LOs to support student mastery. This paper builds on these research streams, extending prior work on LO weighting and decision-making approaches (Divjak et al., 2021; Divjak et al., 2025; Divjak et al., 2024).

Prioritising LOs provides several key benefits:

- First, it enhances constructive alignment, supporting coherence between LOs, teaching and learning activities and assessment (Biggs, 1996). When LOs are prioritised, assessment tasks can be designed to reflect their relative importance. For example, in Divjak et al. (2021), LO weights were used to distribute assessment points across tasks, ensuring that more important outcomes contributed more significantly to final grades. This leads to improved validity and transparency of assessment.
- Second, prioritisation enables the organisation of time and resources helping educators allocate effort toward the most critical LOs. For instance, LOs associated with higher cognitive levels (e.g., analysis, evaluation, creation) can be allocated more instructional time and more complex learning activities, while lower-level LOs receive proportionally less emphasis.
- Third, prioritisation supports the appropriate distribution of workload (reflected in ECTS credits) across student activities (Divjak et al., 2023). The ECTS framework is based on student workload required to achieve LOs (European Commission, 2015). Since LOs differ in complexity, their weights can inform how workload, and consequently credits, should be distributed. For example, outcomes requiring project-based or collaborative work (often associated with higher weights) correspond to a larger share of student effort (Kumpas-Lenk et al., 2018).
- Finally, prioritisation of LOs enables more effective learning design (Divjak et al., 2023). Contemporary learning design approaches emphasise structured, data-informed planning of learning experiences. Learning analytics can be used to analyse how assessment design and its alignment with prioritized LOs influence students' mastery of those LOs (Divjak et al., 2025).

In summary, prioritising LOs is an important step in improving the quality and coherence of learning design and its implementation. It provides a foundation for valid assessment, efficient resource allocation, appropriate workload distribution, and advanced learning design practices. This research builds on prior work in this area and extends it by systematically reviewing prioritisation approaches and analysing suitable decision-making methods.

2.2 Prioritisation Approaches

As presented in Fig. 1, Divjak et al. (2024) classify approaches to LO prioritization according to two key criteria: influence on learners (stakes associated with educational context - i.e., the potential consequences of prioritization decisions for learners, institutions, or professional practice) and the complexity of LOs and their context (e.g., the number of stakeholders, interdependencies, and decision criteria).

Four approaches are proposed:

- The Simple approach is appropriate for educational contexts characterised by low-complexity LOs and low stakes for learners. Typical examples include non-formal learning, short professional development courses, or personal learning contexts where prioritisation decisions have limited consequences for learners. In such situations, LOs can be prioritised using intuitive methods, such as

direct estimation or simple weighting, requiring relatively little time and effort from educators and other stakeholders.

- The Standard approach is suitable when a more structured prioritisation process is desirable. This approach can rely on methods such as ranking, scoring, or other straightforward decision-making procedures. It is practical and not overly demanding in terms of resources and stakeholder involvement.
- The Complicated approach is intended for contexts involving either more complex LOs or a greater need to consider multiple stakeholder perspectives. Such contexts require a systematic and participatory prioritisation process, but they do not necessarily require highly sophisticated techniques. Structured expert consensus methods and stakeholder-informed prioritisation are characteristic of this approach.
- The Complex approach is appropriate for contexts characterised by highly complex LOs and high stakes for learners, such as formal degree programmes, certifications, or learning programmes with significant implications for learners' careers. Prioritisation requires substantial time, expertise, and analytical rigour. Divjak et al. explicitly recommend the use of structured decision-making methods, such as the analytic hierarchy process (AHP), to support comprehensive analysis from multiple expert perspectives and to ensure that the resulting priorities adequately reflect the complexity and importance of the educational context.

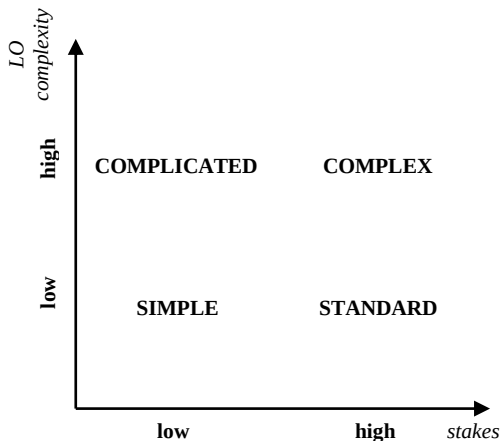


Figure 1. LO prioritisation approaches (adapted from Divjak, 2024)

Detailed guidance for classifying educational contexts according to LO complexity and stakes is provided by Divjak et al. (2024), while this paper focuses on selecting suitable prioritisation methods for each category.

2.3 Multi-Criteria Decision-Making Methods

Multi-Criteria Decision-Making (MCDM) refers to a family of methods designed to support decision-making in situations in which multiple, often conflicting criteria must be considered simultaneously. Depending on the decision context, MCDM approaches can be used for selection, ranking, classification, or prioritisation of alternatives based on predefined criteria and stakeholder preferences (Saaty, 2008; Hwang & Yoon, 1981).

LO prioritisation shows many characteristics of a typical MCDM problem. Educational stakeholders frequently need to determine which LOs are more important than others when designing curricula, allocating teaching time, constructing assessment plans, or distributing student workload. Such decisions rarely depend on a single factor. In this context, prioritisation of LOs can be conceptualised as a prioritisation-oriented MCDM problem rather than a selection problem, since all LOs remain relevant but differ in their relative importance. Consequently, LO prioritisation can be analysed using MCDM methods according to the following structure:

- Goal: determine the relative weights of LOs;
- Alternatives: LOs;
- Criteria: characteristics used to evaluate LOs (e.g., complexity/Bloom's level, contribution to programme outcomes, required workload, and professional relevance). The final selection of criteria depends on the educational context and should involve the relevant stakeholders (e.g., teachers, programme coordinators, employers or accreditation bodies).

3 Methodology

This study adopts a qualitative research approach to investigate LO prioritisation from two complementary perspectives: existing practices in the literature and potential methods derived from decision-making theory. Accordingly, the study is guided by the following research questions (RQs):

- RQ1: Which methods and approaches are currently used for LO prioritisation, and do they align with respective educational contexts?
- RQ2: Which prioritisation methods are most suitable for educational contexts characterised by different combinations of LO complexity and decision stakes?

To address the RQs, two methodological approaches were employed: (1) systematic literature review (SLR) and (2) qualitative expert-based analysis.

3.1 Literature Review (RQ1)

RQ1 was addressed through an SLR aimed at identifying scientific contributions that explicitly or implicitly deal with the prioritisation or weighting of LOs. The selection of papers is presented in Fig. 2 based on the PRISMA approach (Page et al., 2021). The literature search was conducted in two major scientific databases: Web of Science (WoS) and Scopus. The search queries were designed to capture variations in terminology related to LOs. The following keywords and combinations were used:
OR((("learning outcome" OR "learning objective*" OR "learning goal*" OR "learning competence*" OR "learning competenc*") NEAR/3 weight), (("learning outcome*" OR "learning objective*" OR "learning goal*" OR "learning competence*" OR "learning competenc*") NEAR/3 prioritisation), ("learning outcome*" OR "learning objective*" OR "learning goal*" OR "learning competence*" OR "learning competenc*") NEAR/3 importance), ("learning outcome* assessment" NEAR/3 weight), ("competenc* weighting"), ("outcomes-based education" NEAR/3 weight), ("MCDM" NEAR/3 "learning outcome*"). "educational measurement" NEAR/3 weighting)).*

In the Scopus database, the proximity operator W/3 was used instead of NEAR/3. The search scope included Title, Abstract, and Keywords fields in WoS and Topic in Scopus.

The criteria applied to ensure the relevance and quality of the selected studies were focused on: language (English), publication type (journal articles, conference papers, book chapters, reports), and accessibility (full-text availability).

The initial search resulted in 214 records from WoS and 295 records from Scopus. After removing duplicates, a total of 286 unique records remained for further analysis. The literature review was conducted in several phases.

- Phase 1. Titles and abstracts of all 286 records were analysed by six reviewers to determine whether the study addressed LO prioritisation. Finally, 61 papers were selected for further analysis. Each paper was independently reviewed by two reviewers, while disagreements were resolved through discussion until consensus was reached.
- Phase 2. A full-text analysis was conducted to determine whether papers are in fact relevant and to identify methods and approaches used for prioritising LOs. An additional filter of full-text accessibility was applied, resulting in a final sample of 50 papers for in-depth analysis. During deep analysis of papers, 38 papers were excluded since they do not deal with LO prioritisation.
- Phase 3. The remaining 12 papers were classified according to the approach by Divjak et al. (2024). Given the exploratory nature of this review and the heterogeneity of the included studies, no formal risk-of-bias assessment was conducted.
- Phase 4. A list of prioritisation methods identified in the literature was compiled.

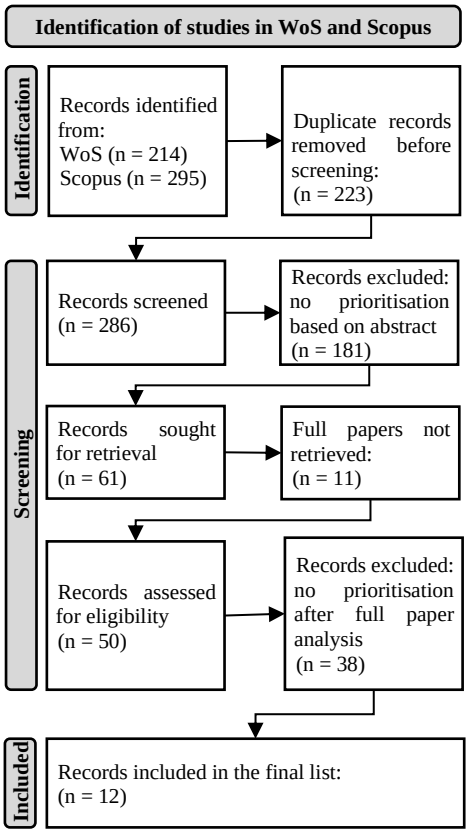


Figure 2. SLR flow (PRISMA)

3.2 Expert-Based Analysis (RQ2)

RQ2 was addressed through qualitative research, in two phases. Experts were selected purposively based on their research and practical experience in MCDM and educational decision-making as well as their availability in the SAILeR project. The discussions were analysed using qualitative thematic synthesis focused on identifying common viewpoints regarding applicability, complexity and usability of the proposed methods.

First, an expert-based analysis was conducted, grounded in the assumption that LO prioritisation can be conceptualised as an MCDM problem. In this context, LOs represent alternatives, while various criteria (e.g., importance, complexity, and workload) can be used to evaluate and prioritise them. To explore suitable methods for this purpose, an initial expert group analysis was conducted involving three experts (authors of this paper) in decision-making and MCDM methods experienced in applying decision-making methods in educational and organisational contexts. The experts discussed the characteristics of the LO prioritisation problem and identified an initial set of six MCDM methods that can be used for LO prioritisation.

Second, under the scope of SAILeR project, two focus groups were held in December 2025, with a total of 15 participants. During the focus group session, the participants were asked to (a) reflect on the LO prioritisation problem, (b) discuss the six MCDM methods identified by the experts, (c) evaluate the suitability of these methods in terms of complexity, interpretability, and practical applicability in educational settings, (d) discuss the applicability for implementation in the BDP tool for learning design, for educators that are typically not experts in MCDM.

4 Results

4.1 Literature Review Results

Table 1 presents an analysis of the approaches used in the final 12 papers. The analysis is focused on LO complexity, influence on learners (educational stakes), prioritisation approach, prioritisation method and appropriateness of the prioritisation method for the educational context, based on Divjak et al. (2024).

With respect to the classification of prioritization approaches proposed by Divjak et al. (2024), the majority of the studies were classified as Complex (n = 8), followed by Simple (n = 2), Standard (n = 1), and Complicated (n = 1).

Table 1. Categorisation of LO prioritisation papers according to Divjak et al. (2024)

Study	Topic	LO Complexity	Influence on Learners	Approach	Method	Appropr.
Sullivan et al. (2025)	Communication curriculum for dementia care partners	High	High stake – affects care quality and healthcare communication	Complex	Three-round Delphi process, ranking and priority ranking of learning objectives (Likert scale)	Medium
Persell (2010)	Introductory sociology learning goals	Low	Low stake – introductory course	Simple	Expert survey and ranking of learning goals (Likert scale)	High
Andayani et al. (2016)	Weighting learning competencies for student assessment	High	High stake – credentialing and certification	Complex	Rating scale combined with Fuzzy Analytic Network Process (FANP)	High
Ali (2024)	Outcome-based education (OBE) assessment framework	High	High stake – formal student assessment	Complex	Weighted outcome-based education framework with outcome and module weighting	High
Yoon et al. (2013)	IFRS knowledge, skills and abilities in accounting education	Low–Medium	High stake – professional accounting practice	Complicated	Employer survey of relative importance of learning objectives (Likert scale followed by ranking)	Low
Duke (2002)	Marketing learning outcomes assessment	Medium	Medium stake – curriculum improvement	Simple	Importance ratings, skill ratings, and priority index calculation (Likert scale followed by raking)	High
Baker et al. (2016)	Relative importance of criminal justice learning objectives	High	High stake – accreditation and professional preparation	Complex	Conjoint analysis and stakeholder ranking exercises	High
Eckel et al. (2017)	Scientific competencies in medical education	High	High stake – medical education and evidence-based practice	Complex	Curriculum inventory, expert consultation, and LO prioritisation process (Likert scale)	High
Boettcher et al. (2024)	Importance of laboratory learning objectives in engineering education	Medium	Medium stake – preparation for engineering work	Standard	Likert-scale survey of perceived importance of LOs	High
Beaudoin et al. (2013)	Biomedical informatics learning objectives in medicine	High	High stake – medical practice	Complex	Student survey of LO importance and desired emphasis	Low
Varga et al. (2020)	LOs and pedagogical skills in medical education	High	High stake – medical education quality	Complex	Importance–delivery gap analysis using questionnaires	Low
Moore et al. (2017)	Neonatal ethics teaching programme	High	High stake – neonatal clinical decision-making	Complex	Explicit weighting of LOs to construct a knowledge test	High

This distribution suggests that most published research on LO prioritisation can be classified as Complex (Sullivan et al., 2025; Andayani et al., 2016; Ali, 2024; Baker et al., 2016; Eckel et al., 2017; Beaudoin et al., 2013; Varga et al., 2020; Moore et al., 2017), meaning they are situated in educational contexts characterised by high influence on learners (stakes) and/or high complexity LOs and learning environments, particularly in professional disciplines such as medicine, healthcare, engineering, and criminal justice. These studies involve professional competencies, certification, clinical practice, or curriculum-wide decision-making. The two studies classified as Simple (Persell, 2010; Duke, 2002) addressed relatively straightforward curricular decision-making problems with limited consequences and comparatively low levels of outcome interdependence. One study, Boettcher et al. (2024), where prioritisation was undertaken in moderately complex educational settings using conventional survey-based approaches. Only one study, Yoon et al. (2013), was classified as Complicated, reflecting a

context characterised by multiple stakeholders and complex communication competencies, addressed through a structured expert consensus process.

With respect to the prioritization methods, the reviewed studies employed a diverse range of methods that can be grouped into five broad categories.

- First, several studies relied on survey- and rating-based approaches, including importance ratings, Likert-scale assessments, and stakeholder surveys (Persell, 2010; Yoon et al., 2013; Boettcher et al., 2024; Beaudoin et al., 2013; Varga et al., 2020). These approaches represent the least formalised forms of LO prioritisation and are primarily used to elicit the perceived importance of LOs.
- Second, expert consensus methods were represented by Sullivan et al. (2025), who employed a multi-round Delphi process to establish LO priorities. Delphi methods introduce structure and transparency into the prioritisation process while maintaining reliance on expert judgement.
- Third, Baker et al. (2016) employed conjoint analysis, a preference-modelling technique that derives the relative importance of LOs from stakeholders' evaluations of alternative competency profiles. Unlike direct importance ratings, conjoint analysis allows priorities to emerge from structured trade-offs among competing attributes.
- Fourth, Eckel et al. (2017) utilised curriculum inventory, mapping, and expert prioritisation procedures to identify and organise scientific competencies within a medical curriculum. Although not based on formal MCDM, this approach represented a systematic and curriculum-wide prioritisation process.
- Finally, the most sophisticated approaches were found in Andayani et al. (2016), Ali (2024), and Moore et al. (2017), who employed formal weighting procedures. These included the Fuzzy Analytic Network Process (FANP), weighted outcome-based education (OBE) frameworks, and explicit weighting of LOs within assessment blueprints. Such approaches closely resemble MCDM methodologies and provide transparent, quantitative prioritisation mechanisms.

To evaluate the appropriateness of the selected prioritisation methods, each study was examined for alignment between the characteristics of the educational context (i.e., LO complexity and influence on learners) and the sophistication of the prioritisation method employed. Of the 12 studies analysed, nine demonstrated strong alignment, one demonstrated partial aligment, and two demonstrated weak alignment.

- Strong alignment was observed when the chosen prioritisation method appeared appropriate for the complexity of the educational context and stakes for learners. This group included Persell (2010), Duke (2002), Yoon et al. (2013), Boettcher et al. (2024), Andayani et al. (2016), Ali (2024), Baker et al. (2016), Eckel et al. (2017), and Moore et al. (2017). In lower-complexity contexts, simple ranking procedures and importance ratings were considered adequate. Conversely, in highly complex and high-stakes settings, studies employed formal weighting methods, preference-modelling techniques, or structured curriculum-wide prioritisation processes. In particular, Andayani et al. (2016), Ali (2024), and Moore et al. (2017) represent exemplary applications of advanced methods in contexts where educational decisions carry substantial consequences.
- Only Sullivan et al. (2025) was classified as demonstrating partial alignment. The study addressed highly complex communication competencies in a high-stakes healthcare context and employed a rigorous Delphi process involving multiple rounds of expert consensus. However, although the Delphi method provides a structured prioritisation process, it does not offer the formal weighting and modelling capabilities associated with advanced MCDM approaches. Consequently, the method was considered appropriate but not fully aligned with the expectations of a Complex prioritisation context.
- Weak alignment was identified in Beaudoin et al. (2013) and Varga et al. (2020). Both studies addressed highly complex LOs within medical education and therefore operated in contexts characterised by both high complexity and high stakes. Nevertheless, prioritisation was based primarily on importance ratings and questionnaire responses. While such methods are useful for identifying stakeholder perceptions, they provide limited support for systematic prioritisation in contexts where educational decisions may have significant implications for professional competence and practice. These studies illustrate situations in which the sophistication of the prioritisation method appears insufficient relative to the complexity of the educational context.

4.2 Qualitative Analysis Results

RQ2 aimed to identify which methods and approaches could be used for LO prioritisation according to the educational context. This question was addressed through a qualitative analysis involving experts in MCDM. The expert analysis confirmed that LO prioritisation can be effectively conceptualised as a MCDM problem. The experts emphasised that LO prioritisation is a prioritisation problem rather than a selection problem, meaning that all alternatives (LOs) remain relevant but differ in importance.

The initial expert discussion was focused on analysing the MCDM methods, resulting in a proposal of six methods, which were later discussed from user perspective in larger focus groups. These methods differ primarily in terms of their structural assumptions

(e.g., independence vs. interdependence of elements) and computational complexity. The six methods are:

1. Simple Additive Weighting (SAW): The simplest approach, based on a flat structure consisting of a goal, criteria, and alternatives. Each LO is evaluated against predefined criteria, and weights are calculated through aggregation. This approach is intuitive and easy to apply, making it particularly suitable for educational contexts where users may not have expertise in decision-making methods. However, it assumes independence among criteria and LOs. (Fishburn, 1967; Hwan et al., 1981)
2. Analytic Hierarchy Process (AHP): Introduces a hierarchical structure, allowing criteria to be decomposed into sub-criteria. For example, LO complexity can be further divided into cognitive demand (level of LO) and content specifics (context). AHP enables pairwise comparisons and consistency checks, providing a more structured and theoretically grounded approach than SAW. It is suitable when decision-makers require greater analytical rigour while maintaining interpretability. (Saaty, 1980; Saaty, 2008)
3. Analytic Network Process (ANP): Extends the AHP by allowing interdependencies between elements (criteria and/or LOs). This is particularly relevant because some LOs may depend on or influence others. However, experts highlighted that modelling such dependencies introduces significant complexity, both in terms of model construction and interpretation. This raises practical challenges, especially for non-expert users. (Saaty, 1996; Saaty, 2004)
4. SWING Method – Truncated Structure: Represents a simplified approach that focuses on direct evaluation and ranking of alternatives without explicitly modelling a full criteria hierarchy. It involves ranking LOs and assigning relative importance through normalisation. This method is less demanding in terms of input data and computational effort, making it suitable for quick assessments and participatory decision-making settings. (von Winterfeldt, 1986; Mustajoki, 2005)
5. Decision Making Trial and Evaluation Laboratory (DEMATEL) – Truncated Network Structure: Used to analyse relationships and influences between elements. In LO prioritisation, it can be applied to identify causal relationships between LOs. This method is particularly useful when dependencies between LOs are important, such as when certain LOs serve as prerequisites for others. However, interpreting these relationships in an educational context may not always be straightforward. (Gabus, 1972; Si et al., 2018)
6. Social Network Analysis Process (SNAP) – Combined Network Approach: It combines elements of network-based approaches (such as ANP and DEMATEL) with social network analysis. It aims to reduce complexity while preserving the ability to model relationships between elements. Experts identified SNAP as a promising method that balances analytical power and usability. Compared to ANP, it requires fewer inputs and is easier to interpret, which makes it more suitable for practical application in education. (Kadoić et al., 2017; Kadoić et al., 2019)

These methods were analysed by expert and focus groups' participants considering their complexity. The discussion also involved the suitability of the methods for LO prioritization for implementation in the BDP learning design tool. The expert and focus group discussions revealed several important insights. First, the choice of method depends on the trade-off between complexity and ease of use.

While advanced methods (e.g., ANP, DEMATEL, SNAP) provide more accurate modelling of relationships and dependencies, they are also more difficult to implement and understand, particularly for educators without a background in MCDM. Second, the role of dependencies between LOs remains an open question. Although network-based methods allow modelling such dependencies, experts questioned whether this added complexity is always necessary or beneficial in practice. Third, simpler structures were recognised as more suitable for practical implementation in educational settings, especially when the goal is to support teachers in curriculum design and assessment planning.

5 Discussion

The comparison of the identified methods highlights key differences in their applicability to LO prioritisation (Table 2). Simpler approaches, such as SAW and SWING, are characterised by low complexity, minimal data requirements, and high usability, making them particularly suitable for educators without expertise in decision-making methods. AHP represents a balanced approach, offering increased analytical rigour through a hierarchical structure while remaining relatively interpretable. In contrast, ANP, DEMATEL, and SNAP enable modelling of dependencies between LOs, which may be relevant in educational contexts where outcomes are interrelated and the stakes are high, e.g., at the study programme level. However, these methods introduce higher complexity and lower usability, which may limit their practical adoption.

The methods discussed in this paper primarily represent compensatory decision models, allowing lower performance on one criterion to be offset by stronger performance on another. Extending the framework to include non-compensatory decision models based on thresholds or veto conditions represents an interesting direction for future research.

The choice of method depends on the trade-off between analytical rigour and ease of use. While network-based approaches provide richer modelling

capabilities, simpler methods may be more appropriate in practice, especially when prioritisation is conducted by non-experts or a dedicated functionality is implemented in a learning design tool (e.g., BDP). There is potential for broader adoption of advanced prioritisation approaches, including MCDM methods, in future curriculum design and assessment research.

Table 2 compares selected LO prioritisation methods based on criteria relevant to the prioritisation context, including support for modelling interdependencies among LOs, complexity handling, data requirements, ease of use, and result interpretability. Based on these characteristics and the framework proposed by Divjak et al. (2024), each method is associated with the prioritisation approach for which it appears most suitable. Simpler methods such as SAW and SWING are better suited to low-complexity and low-stakes contexts, whereas methods such as AHP, ANP, and Conjoint Analysis are more appropriate for highly complex and high-stakes educational settings. DEMATEL and SNAP occupy an intermediate position by supporting the analysis of relationships among LOs while maintaining a moderate level of methodological complexity. Approaches based on Likert-scale evaluations, can be conceptually aligned with the SWING method due to their reliance on direct scoring and normalisation of importance.

Table 2. Comparison for LO prioritisation methods

Criterion	SAW	AHP	ANP	SWING	DEMATEL	SNAP	Conjoint Analysis
Structure of the MCDM model	Flat (additive)	Hierarchical	Network	Simplified ranking	Network (influence-based)	Network (hybrid)	Preference-based
Support for modeling LO interdependencies	Not supported	Not supported	Supported	Not supported	Supported	Supported	Partially supported
Support for modeling LO complexity	Low	Medium	High	Low	Medium–High	Medium–High	Medium–High
Data requirements	Low	Medium	High	Low	Medium	High	High
Ease of use for educators	Very high	Medium	Low	Very high	Low	Medium	Medium–Low
Interpretability of results	High	High	Medium–Low	High	Medium	Medium	Medium
Prioritisation Approach	Simple	Complex	Complex	Simple	Complicated	Complicated–Complex	Complex

The SLR findings provide substantial support for the conceptual framework proposed by Divjak et al. (2024). There is a clear progression from simple ranking and survey methods in lower-complexity contexts, through expert consensus approaches, to formal weighting and preference-modelling techniques in highly complex and high-stakes educational settings.

The majority of studies (75%) demonstrated strong alignment between the characteristics of the educational context and the prioritisation method used. Nevertheless, the presence of partially and weakly aligned studies suggests that LO prioritisation in some high-stakes domains, particularly medical education, continues to rely heavily on stakeholder perceptions rather than on formal decision-making methodologies.

Several limitations of the study should be acknowledged. First, the SLR was limited to English-language publications indexed in WoS and Scopus. Second, the expert analysis involved a relatively small number of experts. Third, the proposed methods were analysed conceptually and through expert evaluation rather than by applying them to a real educational case. Consequently, future work should empirically compare the proposed methods in authentic learning design settings.

Future research should therefore focus on empirical validation of the identified methods in authentic educational contexts, comparison of prioritisation results across stakeholder groups (e.g., teachers, students, employers), and development of practical software-supported prioritisation tools. Additional research should also investigate whether modelling dependencies between LOs provides sufficient added value to justify the increased complexity of network-based methods such as ANP, DEMATEL, and SNAP.

6 Conclusion

This paper addresses the problem of LO prioritisation from the perspective of multi-criteria decision-making (MCDM) by applying a qualitative approach (expert analysis and focus groups) and by analysing approaches in the literature. The study demonstrated that, despite the growing importance of LO-based education, the relatively low number of studies explicitly addresses the prioritisation of LOs.

The literature review identified several existing approaches to LO prioritisation, including survey- and rating-based approaches, expert consensus methods, conjoint analysis, and fuzzy analytic network process. The expert-based analysis demonstrated that LO prioritisation can be effectively conceptualised as a MCDM problem. Six candidate methods were identified as particularly suitable for this purpose: SAW, AHP, ANP, SWING, DEMATEL, and SNAP. The comparison of these methods revealed an important trade-off between analytical rigour and practical usability. Simpler methods, such as SAW and

SWING, appear more suitable for educational practitioners due to their interpretability and ease of use, while more advanced network-based methods provide richer modelling capabilities at the cost of increased complexity. Finally, the choice of MCDM method depends on LO complexity and influence of the decision on learners.

Acknowledgments

This study has been supported by Streamlining AI in Learning Reimagination (SAILeR) (Erasmus+) and Trustworthy Learning Analytics and Artificial Intelligence for Sound Learning Design (TRUELA) (Croatian Science Foundation, IP-2022-10-2854) projects.

References

References marked with an asterisk indicate studies included in the systematic literature review.

*Ali, Q. I. (2024). Towards more effective summative assessment in OBE: A new framework integrating direct measurements and technology. *Discover Education*, *3*, 107. https://doi.org/10.1007/s44217-024-00208-5

*Andayani, S., Hartati, S., Wardoyo, R., & Mardapi, D. (2016). Combining rating scale and FANP to determine the important weight of learning competency for students' assessment. AIP Conference Proceedings, 1755, 070009. https://doi.org/10.1063/1.4958504

*Baker, W. M., Holcomb, J. E., & Baker, D. B. (2017). An assessment of the relative importance of criminal justice learning objectives. Journal of Criminal Justice Education, 28(1), 129–148. https://doi.org/10.1080/10511253.2016.1172650

*Beaudoin, D. E., Richardson, S. J., Sheng, X., & Mitchell, J. A. (2013). Medical students' perspectives on biomedical informatics learning objectives. International Journal of Medical Education, 4, 1–8. https://doi.org/10.5116/ijme.50ce.316b

Biggs, J. (1996). Enhancing teaching through constructive alignment. Higher Education, 32(3), 347–364. https://doi.org/10.1007/BF00138871

*Boettcher, K., Kaufhold, N., Schade, M., Terkowsky, C., & Ortelt, T. R. (2024). Perceived importance of laboratory learning objectives by female and male engineering students. In M. E. Auer et al. (Eds.), Shaping the Future of Learning and Teaching Engineering (Lecture Notes in Networks and Systems, Vol. 1027, pp. 44–55). Springer. https://doi.org/10.1007/978-3-031-61891-8_4

Divjak, B., Kadoić, N., & Žugec, B. (2021). The use of decision-making methods to ensure assessment validity. Proceedings of the IEEE TEMSCON-EUR 2021. https://doi.org/10.1109/TEMSCON-EUR52034.2021.9488603

Divjak, B., Grabar, D., Svetec, B., & Vondra, P. (2022). Balanced Learning Design Planning: Concept and Tool. Journal of Information and Organizational Sciences, 46(2). https://doi.org/10.31341/jios.46.2.6

Divjak, B., Barthakur, A., Kovanović, V., & Svetec, B. (2025). The impact of learning design on the mastery of learning outcomes in higher education. Proceedings of the 15th International Learning Analytics and Knowledge Conference (LAK '25). https://doi.org/10.1145/3706468.3706565

Divjak, B., Svetec, B., Horvat, D., & Kadoić, N. (2023). Assessment validity and learning analytics as prerequisites for ensuring student-centred learning design. British Journal of Educational Technology, 54(1), 313–334. https://doi.org/10.1111/bjet.13290

Divjak, B., Svetec, B., Rienties, B., Astles, P., Matthews, C., Tillmann, A., Eichhorn, M., & Schmitt, J. (2024). Prioritizing learning outcomes in different learning design contexts. In Proceedings of the 35th International Scientific Conference on Information and Intelligent Systems (CECIIS 2024) (pp. 333–338). Faculty of Organization and Informatics, University of Zagreb.

*Duke, N. K., & Pearson, P. D. (2002). Effective practices for developing reading comprehension. In A. E. Farstrup & S. J. Samuels (Eds.), What research has to say about reading instruction (3rd ed., pp. 205–242). International Reading Association.

*Eckel, J., Schüttpelz-Brauns, K., Miethke, T., Rolletschek, A., & Fritz, H. M. (2017). The inventory as a core element in the further development of the science curriculum in the Mannheim Reformed Curriculum of Medicine. GMS Journal for Medical Education, 34(2), Article Doc23. https://doi.org/10.3205/zma001103

European Commission. (2015). ECTS Users' Guide. https://doi.org/10.2766/87192

European Commission/EACEA/Eurydice. (2015). The European Higher Education Area in 2015: Bologna Process implementation report. Publications Office of the European Union

Fishburn, P. C. (1967). Additive utilities with incomplete product sets: Applications to priorities

and assignments. Operations Research, 15(3), 537–542. https://doi.org/10.1287/opre.15.3.537

Gabus, A., & Fontela, E. (1972). World problems, an invitation to further thought within the framework of DEMATEL. Battelle Geneva Research Centre.

Hwang, C. L., & Yoon, K. (1981). Multiple attribute decision making: Methods and applications. Springer. https://doi.org/10.1007/978-3-642-48318-9

Kadoić, N., Begičević Ređep, N., & Divjak, B. (2017). A new method for strategic decision-making in higher education. Central European Journal of Operations Research, 25(3), 571–592. https://doi.org/10.1007/s10100-016-0453-8

Kadoić, N., Divjak, B., & Begičević Ređep, N. (2019). Integrating DEMATEL with the analytic network process for effective decision-making. Central European Journal of Operations Research, 27(3), 653–678. https://doi.org/10.1007/s10100-018-0539-5

Kumpas-Lenk, K., Eisenschmidt, E., & Veispak, A. (2018). Does the design of learning outcomes matter from students' perspective? *Studies in Educational Evaluation, 59*, 179–186. https://doi.org/10.1016/j.stueduc.2018.07.008

*Moore, G. P., Ferretti, E., & Daboval, T. (2017). Developing a knowledge test for a neonatal ethics teaching program. Cureus, 9(12), e1971. https://doi.org/10.7759/cureus.1971

Mulder, M. (2014). Conceptions of professional competence. In S. Billett et al. (Eds.), International handbook of research in professional and practice-based learning. https://doi.org/10.1007/978-94-017-8902-8_5

Mustajoki, J., Hämäläinen, R. P., & Salo, A. (2005). Decision support by interval SMART/SWING—Incorporating imprecision in the SMART and SWING methods. Decision Sciences, 36(2), 317–339. https://doi.org/10.1111/j.1540-5414.2005.00077.x

Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., MayoWilson, E., McDonald, S., … McKenzie, J. E. (2021). PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ, n160. doi: 10.1136/bmj.n160

*Persell, C. H. (2010). How sociological leaders rank learning goals for introductory sociology. Teaching Sociology, 38(4), 330–339. https://doi.org/10.1177/0092055X10378822

Saaty, T. L. (1980). The analytic hierarchy process. McGraw-Hill.

Saaty, T. L. (2008). Decision making with the analytic hierarchy process. International Journal of Services Sciences, 1(1), 83–98. https://doi.org/10.1504/IJSSCI.2008.017590

Saaty, T. L. (1996). Decision making with dependence and feedback: The analytic network process. RWS Publications.

Saaty, T. L. (2004). Fundamentals of the analytic network process—Dependence and feedback in decision-making with a single network. Journal of Systems Science and Systems Engineering, 13(2), 129–157. https://doi.org/10.1007/s11518-006-0158-y

Si, S. L., You, X. Y., Liu, H. C., & Zhang, P. (2018). DEMATEL technique: A systematic review of the state-of-the-art literature on methodologies and applications. Mathematical Problems in Engineering, 2018, 3696457. https://doi.org/10.1155/2018/3696457

*Sullivan, S. S., Wittenberg, E., West, L. J., Gill, A. S., Tomko, C., & Mann, C. (2025). Development of a communication curriculum for family care partners of persons living with dementia. Dementia. Advance online publication. https://doi.org/10.1177/14713012251342029

*Varga, Z., Pótó, Z., Czopf, L., & Füzesi, Z. (2020). Do we need special pedagogy in medical schools? Attitudes of teachers and students in Hungary: A cross-sectional study. BMC Medical Education, 20, Article 472. https://doi.org/10.1186/s12909-020-02385-x

von Winterfeldt, D., & Edwards, W. (1986). Decision analysis and behavioral research. Cambridge University Press.

*Yoon, S. W., Vedd, R., & Jones, C. G. (2013). IFRS knowledge, skills, and abilities: A follow-up study of employer expectations for undergraduate accounting majors. Journal of Education for Business, 88(6), 352–360. https://doi.org/10.1080/08832323.2012.727889

Students' Perceived Usefulness of Prompts in Blended Learning: Two Pilot Studies

Sabina Rako
University of Zagreb
Faculty of Organization and Informatics
Pavlinska 2, 42000 Varaždin, Croatia
sabina.rako@student.foi.hr

Abstract. *Prompts are often used in blended learning to support self-regulation. This paper investigates students' perceived usefulness of prompts across two pilot studies in higher education. The aim was to explore perceived usefulness of different types of prompts and how self-regulation modifies this perception. Perceived usefulness ratings were collected immediately after each prompt, while self-regulated learning was assessed prior to prompt exposure. Qualitative insights were obtained through think-aloud sessions. The results showed that students perceived all prompt types as highly useful. Higher self-regulated learning was associated with higher perceived usefulness ratings in Pilot Study 1, but not in Pilot Study 2.*

Keywords. prompts, self-regulated learning, perceived usefulness, higher education, blended learning

1 Introduction

Self-regulated learning (SRL) is considered an important factor for successful learning in higher education, particularly in digital and blended learning environments where students are expected to manage their learning processes with greater autonomy (Eggers et al., 2021; Michael et al., 2025). SRL refers to learners' active regulation of their cognition, motivation, and behavior in pursuit of learning goals (Zimmerman, 2002). However, students do not always engage in self-regulatory activities spontaneously and often require instructional support to effectively plan, monitor, and evaluate their learning.

One frequently used approach to supporting SRL is the use of prompts. Prompts are short hints, questions, or instructions designed to activate cognitive, metacognitive, motivational, or resource-management strategies that learners already possess but may not apply independently (Wirth, 2009). By directing students' attention to important aspects of the learning process, prompts can encourage reflection, self-monitoring, planning, and deeper engagement with learning tasks. Consequently, prompts have been incorporated into digital, online, and blended learning environments in higher education. Because learning in authentic educational settings differs from controlled laboratory conditions, it is important to examine how prompts function in real-world learning environments (Thomann & Deutscher, 2025).

Previous research has demonstrated that prompts can support self-regulated learning and improve learning processes in digital learning environments (Thomann & Deutscher, 2025). However, less is known about how students interpret and utilize different prompts during learning and how these perceptions relate to learner characteristics (Engelmann et al., 2021). In particular, level of self-regulation might influence how students perceive usefulness of prompts. To address this gap, this paper investigates the following research questions:

- RQ1. How do students perceive usefulness of different types of prompts in a blended learning environment?
- RQ2. What is the relationship between perceived usefulness and students' self-regulated learning score?

To answer these questions and inform the development of a larger-scale study, pilot studies were conducted at two higher education institutions. The findings provide initial insights into students' experiences with prompts and their perceived usefulness.

2 Related Work

2.1 Effectiveness of Prompts in Higher Education

Based on a meta-analysis of 68 studies published between 1999 and 2022, Thomann and Deutscher (2025) found that cognitive, metacognitive, and non-cognitive prompts have a positive effect on learning achievement (Cohen's $d = .394$; adjusted $d = .220$). The strongest effects were observed for action-based prompts, prompts tailored to specific learner groups, and combinations of generic and directed prompts. Although 44 included studies were conducted in higher education,

the findings were aggregated across educational levels, limiting conclusions about higher education context (Thomann & Deutscher, 2025).

Recent studies in higher education have reported mixed findings regarding the effectiveness of prompts. Although metacognitive prompts can support self-regulated learning processes such as planning, monitoring, and reflection (Bannert et al., 2015), their effects on learning achievement are less consistent (Juhaňák et al., 2025). Moreover, while prompt research has primarily focused on learning outcomes and learning behavior, less is known about how learners perceive different types of prompts and what factors influence these perceptions (Engelmann et al., 2021; Schumacher & Ifenthaler, 2021).

3 Methodology

3.1 Pilot Study 1

Pilot Study 1 was conducted during March and April 2023 in an undergraduate course on bibliographic organization. The course enrolled 37 second-year students. Ethical approval was obtained from the institutional ethics committee prior to the study. Participation was voluntary, and informed consent was obtained before data collection.

As shown in Fig. 1, students first completed an initial questionnaire that included demographic information and the Online Self-regulated Learning Questionnaire (OSLQ). During a face-to-face lecture session, students were presented with two prompts and rated their perceived usefulness using paper-based questionnaires. Subsequently, students were provided with anonymous access to a Learning Management System (LMS) using generic usernames and invited to complete an online practice test. Additional prompts were presented before and after the practice test, and their perceived usefulness was rated through the LMS plug-in. The intervention period lasted one week (2–9 March 2023). Following the intervention, students who had indicated their willingness to participate in follow-up activities were contacted regarding think-aloud sessions. Separate consent for audio recording was obtained prior to the sessions.

The following prompts, developed in collaboration with the course teachers and informed by previous course experiences and data, were used in Pilot Study 1:

- *mixed (face-to-face lecture 1)*: "Today, we will develop a mind map to reflect on the concepts acquired so far in the course "Introduction to Information Organization" and identify the additional knowledge and skills needed to become competent cataloguers capable of producing formal bibliographic descriptions."
- *motivational (face-to-face lecture 1)*: "You will receive an additional 2 points if you can list all the conditions for attributing authorship to a corporate body, both for institutions and for events."
- *resource-related (online, prePracticeTest)*: "Before starting the practice test, recall the mind map that we prepared at the beginning of the previous lesson. The mind map contains the information necessary for completing the practice test."
- *resource-related (online, postPracticeTest)*: "The results are presented below. For each incorrect answer, references are provided to the relevant sections of the learning materials where detailed explanations of the correct answers can be found. Please review the indicated materials before the next class session.".

3.2 Pilot Study 2

Pilot Study 2 was conducted between December 2023 and March 2024 in an undergraduate course on business analytics. The course enrolled 55 third-year students. Ethical approval was obtained from the institutional ethics committee prior to the study. Participation was voluntary, and informed consent was obtained before data collection.

As shown in Fig. 2, students first completed an initial questionnaire that included demographic information and the OSLQ. An online metacognitive prompt was provided before the first face-to-face practice session, followed by a resource-related prompt during that session. At the begining of the next face-to-face practice session, students received a metacognitive prompt, and after the session an online resource-related prompt was presented. Immediately following each prompt, students rated its perceived usefulness through the LMS plug-in. The prompts were implemented during January 2024. After the last prompt, students who had indicated their willingness to participate in follow-up activities were contacted regarding participation in think-aloud sessions. Among other activities, students participating in the think-aloud sessions were asked to rank four generic prompts and explain their perceived usefulness. Separate consent for audio recording was obtained prior to the sessions. Informed consent, the initial questionnaire, and OSLQ responses were collected using the LimeSurvey platform (LimeSurvey Project Team / C. Schmitz, 2024).

The following prompts, developed in collaboration with the course teachers and informed by previous course experiences and data, were used in Pilot Study 2:

- *metacognitive (online)*: "As preparation for the face-to-face practice session, it is useful to create notes containing the most important concepts covered during the lecture. Review the lecture and seminar materials and write down the key concepts along with their definitions."
- *resource-related (face-to-face practice session 1)*: "Before completing the exercises in this practice session, please review the solved examples available on

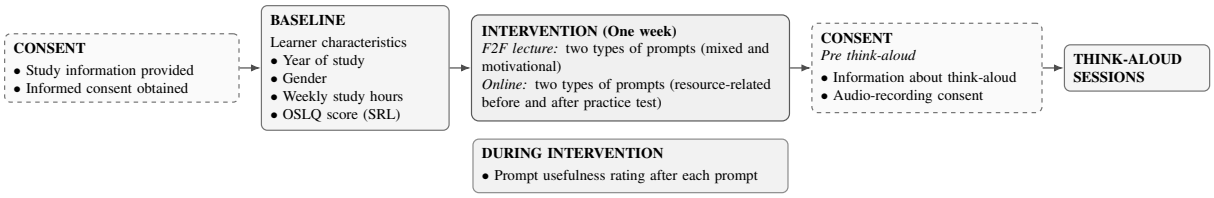


Figure 1. Design of Pilot Study 1.

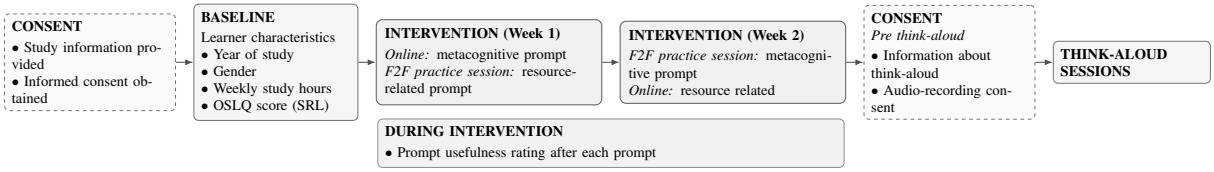


Figure 2. Design of Pilot Study 2.

[directions]."

- *metacognitive (face-to-face practice session 2)*: "Following this face-to-face practice session, you will be required to complete an online self-assessment test. Before starting the test, prepare notes summarizing the key concepts covered to date based on the available course materials. Please, avoid postponing the practice test until the last minute."
- *resource-related (online, prePracticeTest)*: "Materials for preparation for the online test are available within the Network Analysis topic. Please review these materials before attempting the test."

3.3 Measurement Instruments

The perceived usefulness construct was measured using the six-item Perceived Usefulness scale from the Technology Acceptance Model (TAM) (Davis, 1989). Responses were recorded on a visual analog rating scale ranging from 1 (strongly disagree) to 10 (strongly agree), and an overall perceived usefulness score was computed as the mean of the six items. The prompt usefulness rating interface implemented in the LMS is shown in Fig. 3.

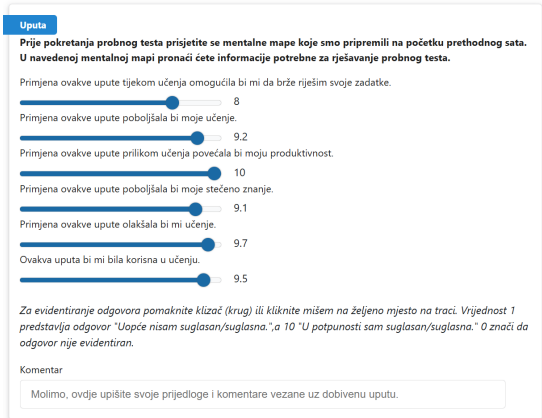


Figure 3. Prompt usefulness rating interface implemented in the LMS.

Within our context, standard Cronbach's alpha reflects how reliably perceived usefulness scale differentiates between students. However, our research question is related to differences in perceived usefulness among prompt types. Generalizability Theory (Brennan, 1992) offers an alternative approach enabling estimation of both between-student and within-student (i.e., among prompt types) reliabilities. To evaluate the reliability of the six-item perceived usefulness scale, variance components were estimated using a linear mixed-effects model with crossed random effects for students, prompt types, perceived usefulness items, and their two-way interactions. Following Generalizability Theory, these variance components were used to compute reliability coefficients for within-student differences across prompt types (R_C) and between-student differences (R_{1F}) (Bolger & Laurenceau, 2013). The perceived usefulness scale showed high reliability in both pilot studies for within-student differences across prompt types ($R_C = .88$ for Pilot 1 and $R_C = .87$ for Pilot 2) and between-student differences ($R_{1F} = .95$ for Pilot 1 and $R_{1F} = .85$ for Pilot 2).

Self-regulated learning was measured using the Online Self-regulated Learning Questionnaire (OSLQ) (Barnard et al., 2009), which consists of 24 items and has been validated in both online and blended learning environments. Responses were recorded on a 7-point Likert-type scale ranging from 1 (strongly disagree) to 7 (strongly agree). In Pilot 2, students could additionally select the option "I don't understand this statement.". An overall self-regulated learning score was calculated as the mean across all items, with higher scores indicating greater self-regulated learning.

The overall OSLQ score used to represent self-regulated learning showed high internal consistency (Cronbach's α = .90 for Pilot 1 and .88 for Pilot 2).

Both measurement instruments were reviewed through cognitive testing following the cross-cultural adaptation procedure recommended by Beaton et al. (2000).

3.4 Data Analysis

For RQ1, a linear mixed-effects model (LMM) was used to examine differences in perceived usefulness ratings across prompt types, with prompt type included as a fixed effect and student as a random intercept:

$$Rating \sim PromptType + (1 \mid StudentID)$$

Think-aloud sessions were audio-recorded, transcribed, and analyzed to identify students' explanations for the perceived usefulness of different prompts.

For RQ2, the relationship between self-regulated learning (OSLQ score) and perceived usefulness ratings was examined using a LMM. OSLQ score, cohorts (Pilot Study 1 vs. Pilot Study 2), and their interaction were included as fixed effects, while student was included as a random intercept to account for repeated ratings provided by the same student:

$$Rating \sim OSLQ_score * Cohort + (1 \mid StudentID)$$

Assumptions for all LMMs were assessed using graphical diagnostics (normality of residuals, homogeneity of residual variance, and normality of random effects), and no substantial violations were detected.

Data were analyzed in R (R Core Team, 2026) using the RStudio integrated development environment (Posit team, 2026) and packages `psych` (William Revelle, 2026), `lmerTest` (Kuznetsova et al., 2017), `lme4` (Bates et al., 2015), `tidyverse` (Wickham et al., 2019) and `ggplot2` (Wickham, 2016).

4 Results

Participant characteristics for both pilot studies are presented in Table 1.

Table 1. Participant characteristics and prompt rating response counts

Characteristic	Pilot Study 1	Pilot Study 2
OSLQ completed, n	30	40
Year of study	Second	Third
Female, n (%)	20 (66.7%)	20 (50.0%)
Weekly study hours, M (SD)	9.1 (6.7)	11.9 (8.0)
OSLQ score (1–7), M (SD)	5.0 (1.0)	4.6 (0.8)
Prompt 1 ratings, n	30	31
Prompt 2 ratings, n	30	31
Prompt 3 ratings, n	17	28
Prompt 4 ratings, n	17	26

The following sections report results for each research question.

4.1 RQ1: Perceived Usefulness of Different Types of Prompts

In both pilot studies, students rated all prompt types as highly useful (Fig. 4). Fig. 4 indicates that usefulness ratings were consistently high across prompt types in both pilot studies, with slightly higher median ratings in Pilot Study 2.

Linear mixed-effects models indicated no significant differences in perceived usefulness across prompt types in either Pilot Study 1, $F(3, 62.9) = 1.34$, $p = .268$, or Pilot Study 2, $F(3, 78.8) = 0.26$, $p = .855$.

Although prompt types were not directly comparable across pilot studies, the resource-related online prompt provided before the practice test received higher ratings in Pilot Study 2 ($M = 8.49$, $SD = 1.46$, $n = 26$) than the conceptually similar prompt in Pilot Study 1 ($M = 6.91$, $SD = 1.66$, $n = 17$). However, the prompts were not identical, which limits the direct comparability of the results.

In both pilot studies, six students participated in think-aloud sessions: two in Pilot Study 1 and four in Pilot Study 2.

In Pilot Study 1, students primarily commented on the technical aspects of the prompt implementation (e.g., the slider used to evaluate prompts). They generally perceived the prompts as helpful and appreciated the additional support for learning.

In Pilot Study 2, students commented prompt types. For motivational prompt, students most frequently highlighted bonus points as a key motivator for participation, while also recognizing learning activity for bonus points (the outcome) as relevant. One student remarked that the prompt "offers something concrete". Metacognitive prompt was generally perceived as beneficial. However, students often attributed their value to the associated learning activity rather than the metacognitive guidance itself. For resource-related prompt, students valued being directed to relevant resources, particularly because this reduced the effort required to locate correct information. As one student noted, "it is much easier when I have guidance on where to find the correct information". For cognitive prompt, students appreciated reminders of important concepts and their interconnections, but generally considered this prompt less useful than motivational, metacognitive, or resource-related prompts because the information was expected to be learned through regular studying. But, one student explained, "Simply reading it can create a picture of what has been learned so far and help make connections".

4.2 RQ2: Relationship between Perceived Usefulness and Self-regulated Learning

A linear mixed-effects model revealed a significant positive association between self-regulated learning (OSLQ) score and perceived usefulness ratings ($b = 1.32$, $SE = 0.27$, $p < .001$). A significant interaction between OSLQ score and cohort ($b = -1.41$, $SE = 0.38$, $p < .001$) indicated that this relationship differed between the two pilot studies. Specifically, higher OSLQ scores were associated with higher perceived usefulness ratings in Pilot Study 1, whereas

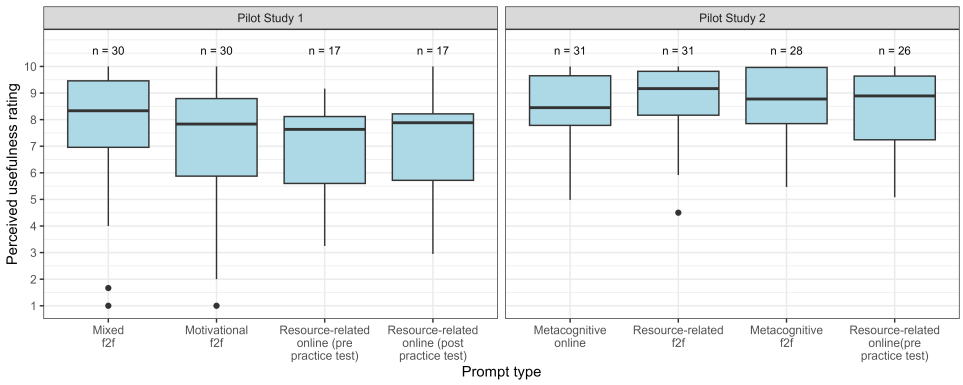


Figure 4. Perceived usefulness ratings across prompt types in Pilot Studies 1 and 2. Numbers above boxplots indicate the number of available ratings.

no meaningful association was observed in Pilot Study 2 (Fig. 5).

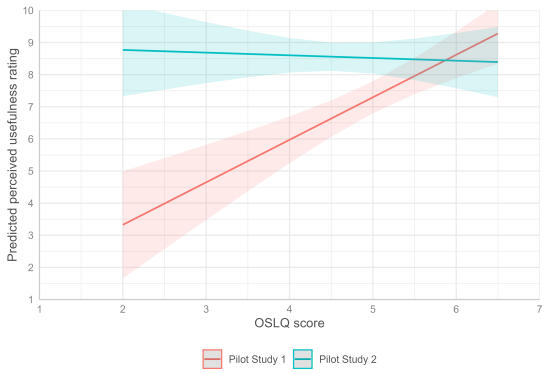


Figure 5. Model-predicted perceived usefulness ratings vs. self-regulated learning (OSLQ) scores in Pilot Studies 1 and 2 with 95% prediction band.

5 Conclusions, Limitations and Future Work

The findings suggest that students generally perceived motivational, metacognitive, resource-related, and mixed prompts as useful for supporting learning in blended learning environments. No significant differences in perceived usefulness were observed across prompt types in either pilot study. Higher self-regulated learning was associated with higher perceived usefulness ratings in Pilot Study 1, but not in Pilot Study 2. This could be a ceiling effect because perceived usefulness scores in Pilot Study 2 are higher and their variability is lower. However, further research is needed to better understand this relationship across different learning context.

This study has several limitations. The pilot studies involved relatively small samples and were conducted in different courses and institutional contexts. In addition, some prompt types were evaluated by fewer students than others, and the prompts were not identical across the two pilot studies, limiting direct comparisons. Another limitation is reduced variability and compression towards the upper limit of perceived usefulness scale in Pilot Study 2, creating potential for the ceiling effect.

Future research should examine perceived usefulness of prompts in larger and more diverse educational settings and investigate factors that may moderate the relationship between self-regulated learning and perceptions of prompt usefulness. Furthermore, it might be beneficial to develop a new measurement scale that could better differentiate students who perceive prompts as highly useful to reduce the ceiling effect. Further studies should also explore how perceived usefulness relates to actual self-regulated learning behaviours and learning outcomes.

Acknowledgments

The author would like to thank the students and teachers who participated in and supported the pilot studies.

References

Bannert, M., Sonnenberg, C., Mengelkamp, C., & Pieger, E. (2015). Short- and long-term effects of students' self-directed metacognitive prompts on navigation behavior and learning performance. *Computers in Human Behavior*, *52*, 293–306. https://doi.org/10.1016/j.chb.2015.05.038

Barnard, L., Lan, W. Y., To, Y. M., Paton, V. O., & Lai, S.-L. (2009). Measuring self-regulation in online and blended learning environments. *The Internet and Higher Education*, *12*(1), 1–6. https://doi.org/https://doi.org/10.1016/j.iheduc.2008.10.005

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Beaton, D. E., Bombardier, C., Guillemin, F., & Ferraz, M. B. (2000). Guidelines for the Process of Cross-Cultural Adaptation of Self-Report Measures: *Spine*, *25*(24), 3186–3191. https://doi.org/10.1097/00007632-200012150-00014

Bolger, N., & Laurenceau, J.-P. (2013). *Intensive longitudinal methods: An introduction to diary and experience sampling research*. Guilford Press.

Brennan, R. L. (1992). An NCME Instructional Module on Generalizability Theory [Module 14, Instructional Topics in Educational Measurement (ITEMS)].

Davis, F. D. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly*, *13*(3), 319–340. https://doi.org/10.2307/249008

Eggers, J. H., Oostdam, R., & Voogt, J. (2021). Self-regulation strategies in blended learning environments in higher education: A systematic review. *Australasian Journal of Educational Technology*, *37*(6), 175–192. https://doi.org/10.14742/ajet.6453

Engelmann, K., Bannert, M., & Melzner, N. (2021). Do self-created metacognitive prompts promote short- and long-term effects in computer-based learning environments? *Research and Practice in Technology Enhanced Learning*, *16*(1), 3. https://doi.org/10.1186/s41039-021-00148-w

Juhaňák, L., Juřík, V., Dostálová, N., & Juříková, Z. (2025). Exploring the effects of metacognitive prompts on learning outcomes: An experimental study in higher education. *Australasian Journal of Educational Technology*. https://doi.org/10.14742/ajet.9486

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

LimeSurvey Project Team / C. Schmitz. (2024). *Limesurvey: An open source survey tool*. LimeSurvey Project. Hamburg, Germany. https://www.limesurvey.org

Michael, J., B Santoso, H., Junus, K., & Sandiwarno, S. (2025). Systematic Review of Self-Regulated Learning Intervention and Measurement in Higher Education: Towards a Holistic, Integrated, and Technology-Assisted Model. *Journal of Information Technology Education: Research*, *24*, 033. https://doi.org/10.28945/5636

Posit team. (2026). *Rstudio: Integrated development environment for R*. Posit Software, PBC. Boston, MA. http://www.posit.co/

R Core Team. (2026). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://doi.org/10.32614/R.manuals

Schumacher, C., & Ifenthaler, D. (2021). Investigating prompts for supporting students' self-regulation – a remaining challenge for learning analytics approaches? *The Internet and Higher Education*, *49*, 100791. https://doi.org/https://doi.org/10.1016/j.iheduc.2020.100791

Thomann, H., & Deutscher, V. (2025). Scaffolding through prompts in digital learning: A systematic review and meta-analysis of effectiveness on learning achievement. *Educational Research Review*, *47*, 100686. https://doi.org/10.1016/j.edurev.2025.100686

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., ... Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

William Revelle. (2026). *Psych: Procedures for psychological, psychometric, and personality research* [R package version 2.6.4]. Northwestern University. Evanston, Illinois. https://CRAN.R-project.org/package=psych

Wirth, J. (2009). Promoting self-regulated learning through prompts. *Zeitschrift für Pädagogische Psychologie / German Journal of Educational Psychology*, *23*(2), 91–94. https://doi.org/10.1024/1010-0652.23.2.91

Zimmerman, B. J. (2002). Becoming a Self-Regulated Learner: An Overview. *Theory Into Practice*, *41*(2), 64–70. https://doi.org/10.1207/s15430421tip4102_2

Pedagogical Friction in ICT-Integrated K-12 Classrooms: Cognitive Offloading, Technostress, and Shallow Processing

Goran Hajdin
University of Zagreb
Faculty of organization and informatics
Pavlinska 2, 42000 Varaždin, Croatia
goran.hajdin@foi.unizg.hr

Abstract. *National K-12 digitalization initiatives assume that interactive digital tools inherently enhance learning outcomes. This convergent parallel mixed-methods study (N=90 teachers) investigates the socio-cognitive impacts and pedagogical friction of digital tools in K-12 classroom environment. The scope of this paper is delimited to a specific thematic fragment of the broader survey dataset, focusing specifically on the relationship between teacher digital competence, technostress, and observed student cognitive behaviors. Results reveal a significant discrepancy between general teacher satisfaction and actual student cross-curricular understanding. Ordinal logistic regression indicates that teachers' digital competence does not predict satisfaction or ease of use. Instead, persistent first-order barriers act as an absolute tax, triggering uniform technostress. According to teachers' observations, interactive digital scenarios are perceived to be associated with student behaviors such as skimming, scrolling, and prompt-dependency, which may reflect a tendency toward surface-level task engagement and cognitive offloading rather than sustained schema construction.*

Keywords. Teaching Scenarios, Cognitive Offloading, Technostress, Shallow Learning, First-order Barriers, Educational Technology

1 Introduction

Integrating Information and Communication Technology (ICT) in K-12 classrooms has historically operated under a technocentric, confirmatory paradigm. Driven by nationwide digitalization initiatives, such as Croatia's "e-Schools" initiative, policymakers frequently assume that replacing paper-based media with interactive digital scenarios inherently enhances motivation and curricular efficiency (Altamura et al., 2025; Caena & Redecker, 2019; Liman Kaban & Karadeniz, 2021). However, a critical counter-narrative has emerged. Rather than witnessing a seamless translation of digital visual interaction into deep learning process, K-12 teachers report significant pedagogical friction, fragmented classroom orchestration, and flattened learning outcomes (Ajani, 2024; Skulmowski, 2023).

This disconnect represents a clash between software design architectures and human cognitive architecture (Sweller et al., 2011). Digital learning scenarios often prioritize rapid, fragmented interactivity and gamification. While these elements temporarily boost surface-level engagement, they simultaneously trigger high extraneous cognitive load (Skulmowski, 2023; Sweller et al., 2011). Under this paradigm, students are conditioned to engage in shallow processing characterized by non-linear visual scanning, keyword spotting, and rapid scrolling (Altamura et al., 2025; Liu, 2005; Mason et al., 2024). Furthermore, immediate, frictionless access to information on screens facilitates cognitive offloading (Risko & Gilbert, 2016). Instead of exerting the germane cognitive effort required for schema construction, students outsource memory processes to digital devices, treating the interface as a prosthetic memory bank (Gilbert et al., 2023; Grinschgl et al., 2021; Sun et al., 2025).

This transition imposes parallel burdens on teachers. Teachers must navigate complex orchestration while managing persistent first-order infrastructural failures like unstable networks and rigid software boundaries (Alptekin & Taneri, 2025). Troubleshooting technology under compressed lesson times triggers severe technology-induced anxiety, or technostress (Tarafdar et al., 2007, 2024).

While policymakers assume that digital upskilling frameworks like DigCompEdu mitigate these challenges, empirical evidence suggests a paradox of competence (Antonietti et al., 2022; Caena & Redecker, 2019). Highly competent teachers do not experience greater satisfaction. Instead, their advanced skills afford them a critical diagnostic gaze that exposes systemic infrastructural deficits and pedagogical mismatches (Ajani, 2024; Suzer & Koc, 2024). This diagnostic tension is intertwined with teachers' self-efficacy, their integrated competence

profiles to promote self-regulated learning (SRL), and their capacity to scaffold metacognition (Dignath, 2021; Hu et al., 2025; Karlen et al., 2023). When automated software designs bypass didactical mediation, they decouple self-efficacy from competence deployment, exacerbating classroom orchestration friction and student cognitive dependency (Dignath, 2021; Karlen et al., 2023).

Using a convergent parallel mixed-methods design grounded in the pragmatist paradigm (Cohen et al., 2018), this study analyzes empirical survey data and qualitative reflections from 90 teachers. The primary goal of this study is to investigate the dual-sided pedagogical friction of ICT integration in classical K-12 classrooms by evaluating the mitigating effect of teacher digital competence on structural barriers and technostress, while simultaneously analyzing the observed impacts of interactive digital scenarios on student cognitive load, reading behaviors, and knowledge retention.

2 Previous Research

2.1 Screen-Mediated Reading and Multimodal Integration

Transitioning from print to digital mediums reshapes students' reading habits and comprehension strategies. Although digital mediums deliver interactive multimedia, empirical research indicates that screen reading alters cognitive engagement, fostering superficial processing (Altamura et al., 2025; Liu, 2005). Liman Kaban and Karadeniz (2021) observed that personalized, gamified digital reading increased motivation but failed to yield superior comprehension, encouraging rapid, less focused reading behaviors. Over time, frequent digital reading negatively impacts comprehension and long-term motivation, conditioning students to engage with highly fragmented texts (Altamura et al., 2025).

To counteract shallow processing, digital interventions like highlighting are used. However, Mason et al. (2024) found that manual digital highlighting can overload working memory, whereas pre-provided highlights often lead to passive processing, showing that digital tools must actively support generative cognitive engagement to be effective. This is particularly critical in complex environments integrating text with diagrams, which represents a major hurdle for young learners (Jian & Cheung, 2025; Sweller et al., 2011). Jian and Cheung (2025) identified stable, distinct student strategic profiles, while integrative readers engaged in recurrent cross-referencing between text and diagrams (predicting superior comprehension), non-integrative readers struggled to coordinate visual attention. Because these profiles remained stable from grades four to six, students do not naturally develop multimodal integration strategies through passive exposure to digital media. Instructional designs must actively guide visual attention (Jian & Cheung, 2025; Sweller et al., 2011).

2.2 Cognitive Offloading and Metacognitive Biases

The ease of retrieving information from digital devices has altered internal memory utilization, leading to cognitive offloading - the strategic use of external resources to reduce internal cognitive demands (Risko & Gilbert, 2016). Gilbert et al. (2023) note that intention offloading (outsourcing tasks to digital reminders) is highly effective and metacognitively guided. However, individuals frequently exhibit a positive reminder bias, over-allocating intentions to digital tools even when internal memory is fully capable (Gilbert et al., 2023).

While offloading delayed intentions preserves cognitive bandwidth, offloading factual information into digital storage can degrade long-term memory representations (Grinschgl et al., 2021). When the physical or cognitive cost of offloading is low, individuals readily externalize memory processes to boost immediate task execution. However, this immediate performance benefit comes at a direct cost to subsequent memory; in unexpected post-tests, individuals who offloaded demonstrated significantly poorer recall (Grinschgl et al., 2021). To integrate these findings, Skulmowski (2023) developed a model of digital externalization, positioning extraneous cognitive load in technology-enhanced learning as a triple barrier that prevents key information from entering working memory, inhibits schema construction, and encourages premature cognitive offloading.

This offloading bias has profound developmental implications. Children as young as four engage in offloading, but demonstrate significant biases, showing periods of non-selective over-offloading (Armitage & Gilbert, 2025). Sun et al. (2025) verified this developmental trajectory from ages seven to fifteen. Primary school students display an under-offloading bias due to overestimating their memory capacity, whereas secondary school students exhibit more calibrated, strategic choices. This presents a major challenge for K-12 digital scenarios; younger students may lack the metacognitive maturity to know when to rely on digital aids, while older students may over-rely on instant feedback, leading to detrimental cognitive offloading (Armitage & Gilbert, 2025; Sun et al., 2025).

2.3 Systemic Orchestration, Teacher Competence, and SRL Scaffolding

The successful operationalization of digital scenarios depends on the teacher's digital competence. Under the DigCompEdu framework, teacher proficiency profiles span from Explorer (basic) to Innovator (advanced)

(Caena & Redecker, 2019). However, average competence levels typically cluster around the intermediate (Integrator) tier, reflecting persistent environmental boundaries that prevent advanced innovation (Suzer & Koc, 2024). According to the Technology Acceptance Model (TAM), digital competence beliefs predict perceived usefulness and ease of use, determining adoption intentions (Antonietti et al., 2022). Yet, this linear relationship is frequently disrupted by classroom orchestration demands, where teachers must navigate physical space, student engagement levels, and constant instructional interruptions (Ajani, 2024; Alptekin & Taneri, 2025).

This orchestration challenge is directly tied to the promotion of SRL and metacognitive scaffolding. Digital competence and teacher self-efficacy are deeply reciprocal (Hu et al., 2025). However, a decoupling between self-efficacy and actual digital execution occurs when teachers are confronted with rigid, pre-packaged digital scenarios that bypass active didactical mediation (Hu et al., 2025; Karlen et al., 2023). Karlen et al. (2023) show that a teacher's capacity to promote student metacognition is fully mediated by their professional pedagogical-metacognitive knowledge, subjective self-efficacy, and intrinsic valuation of self-regulation. If digital scenarios restrict teacher mediation, they render professional knowledge useless, depriving students of the human scaffolding needed to build independent learning strategies.

Furthermore, teacher competence in promoting SRL is highly heterogeneous. Dignath (2021) identified distinct competence profiles, demonstrating a “Matthew effect” where high-competence teachers benefit exponentially more from professional training. Digital scenarios that ignore these diverse latent profiles fail to support teachers who need the most assistance, imposing uniform constraints that convert potential technological aids into sources of classroom friction and cognitive dependency (Dignath, 2021; Karlen et al., 2023).

3 Research Methodology

This study utilizes a convergent parallel mixed-methods design grounded in the pragmatist paradigm (Cohen et al., 2018). By combining the quantitative modeling of Likert-scale evaluations with the qualitative depth of reflexive thematic coding, this framework seeks to triangulate teacher-reported satisfaction against emergent socio-cognitive barriers in real-world K-12 classrooms.

The target population comprises active primary and secondary teachers exposed to a national digitalization initiative. The final sample consists of 90 teachers selected via convenience sampling. The sample is heavily female (88.9%). The sample reflects an experienced teaching profile: 22.2% possess 11–15 years of experience, 16.7% possess 16–20 years, 16.7% possess 21–25 years, and 13.3% have 31 or more years of teaching practice. Self-reported digital competence, measured via DigCompEdu, reveals a proficient workforce: 54.8% identify as intermediate (Integrator), 35.5% as advanced (Innovator), and 9.7% as beginner (Explorer).

To address the research goal, we defined two research questions:

- RQ1: To what extent do self-reported high levels of teacher digital competence mitigate the perception of first-order infrastructural barriers and the experience of technostress?
- RQ2: In what ways do interactive digital scenarios influence K-12 students' cognitive load, reading behaviors, and knowledge retention, as observed by their teachers?

Data collection was executed using a validated online survey instrument adapted from Framework for Digitally Mature Schools in the e-Schools initiative. The quantitative component includes structured Likert-scale questions (1–5 scale) evaluating instructional quality (Q10), student cross-curricular understanding (Q11), and perceived efficiency (Q12), alongside a 15-item evaluation grid (Q9) measuring interactive and organizational attributes. Scale reliability was confirmed via Cronbach's alpha, yielding a high reliability coefficient of 0.88. The qualitative component comprises open-ended questions (Q13, Q14, Q15) exploring advantages, disadvantages, and cognitive barriers.

The empirical baseline of this research is situated within the broader context of Croatia's national digitalization campaign, specifically the e-Schools initiative (2nd phase 2018 to 2023), which aimed to establish digital maturity across 1321 schools. While the broader, multi-dimensional dataset captured macro-level institutional indicators such as network performance, administrative e-services, and institutional platform usage, this study targeted the specific intersection of teacher digital competence, classroom orchestration, and perceived student cognitive dynamics. Isolating this micro-level socio-cognitive fragment is methodologically justified because it directly addresses the pedagogical friction and classroom orchestration realities of educators, preventing the confounding influence of macro-systemic administrative data on teachers' immediate pedagogical experiences.

Quantitative data were analyzed using descriptive statistics and ordinal logistic regression modeling in R Studio to evaluate whether digital competence (low, medium, high) and teaching experience predict satisfaction. Qualitative data were analyzed using Braun and Clarke’s (2006, 2021) six-phase reflexive thematic analysis, generating semantic and latent codes to capture the lived experiences of digital classroom integration.

3.1 Research Limitations

Several limitations of this study must be acknowledged. First, the empirical analysis is confined to a selective thematic fragment of a broader survey dataset. While isolating this micro-level socio-cognitive intersection was necessary to model pedagogical friction and classroom orchestration, the exclusion of macro-systemic variables means that positive technological adaptations, infrastructure successes, and broader system-level variables captured by the wider instrument are not represented here. Second, the sample size is relatively modest (N = 90 total, with n = 62 active users) and was drawn using convenience sampling, which severely limits the immediate generalizability of the results. Furthermore, the sample is heavily female-skewed (88.9%), which represents a potential confounding variable given that technology-related stress and coping mechanisms may display gendered variations. The choice to isolate this specific thematic fragment of the broader dataset, was guided by a desire to focus on micro-level socio-cognitive dynamics. However, the exclusion of macro-systemic variables means that broader institutional factors influencing technology acceptance are not represented in these findings. Finally, the evaluation of student cognitive load, reading patterns, and knowledge retention relies entirely on the subjective observations and perceptions of classroom teachers rather than direct objective student-side metrics such as eye-tracking or standardized comprehension tests.

4 Results and Discussion

To establish a statistical baseline, descriptive statistics were calculated for the primary quantitative evaluation variables within the active user sub-sample of 62 teachers (excluding non-users filtered out via Q5). These statistics are summarized in Table 1.

Table 1. Quantitative evaluative baselining (Positive evaluations represent scores of 4 and 5; Neutral represents 3; Negative represents 1 and 2)

	Q10 (Satisfaction with Instructional Quality)	Q11 (Student Cross-Curricular Understanding)	Q12 (Cross-Curricular Efficiency)
Mean	3.80	3.82	3.44
SD	0.72	0.84	0.81
Positive	85.5%	74.2%	62.9%
Neutral	11.3%	21.0%	27.4%
Negative	3.2%	4.8%	9.7%

The data reveal an evaluative discrepancy of educator perceptions. While 85.5% of teachers report high overall satisfaction with instructional quality when using digital scenarios (with a mean of 3.80 and a standard deviation of 0.72), their consensus on cognitive outcomes is more distributed. Specifically, 74.2% of respondents evaluate students' actual cross-curricular understanding and application positively; although the overall proportion of positive ratings is lower, the mean rating remains statistically comparable (mean = 3.82, SD = 0.84) due to a lower frequency of highly negative evaluations in this dimension. Perceived efficiency shows a more pronounced decline, with only 62.9% of respondents agreeing that digital scenarios are more efficient than traditional, paper-based approaches (mean = 3.44, SD = 0.81).This drop-off indicates that while teachers value interactive potential, they observe a decline in the efficiency of knowledge retention and deep conceptual processing.

4.1 Addressing RQ1

The first research question investigated the extent to which self-reported teacher digital competence mitigates the perception of first-order infrastructural barriers and the experience of technostress. To evaluate whether teacher digital competence and pedagogical experience predict satisfaction, an ordinal logistic regression model was constructed with Satisfaction with Instructional Quality (Q10) as the ordinal outcome variable. The model was statistically non-significant (Wald chi-square = 2.12, df = 3, p = 0.548), indicating that individual professional variables are poor predictors of instructional satisfaction. The complete parameter estimates are presented in Table 2.

Table 2. Ordinal logistic regression predicting Instructional Satisfaction (Q10)

Predictor	Coeff. (Beta)	SE	Wald X^2	p	Odds ratio
Compet. (med vs low)	0.25	0.35	0.51	0.475	1.28
Compet. (high vs low)	0.42	0.38	1.22	0.267	1.52
Teaching exp.	-0.05	0.08	0.39	0.532	0.95

The Brant test confirmed that the proportional odds assumption was satisfied, indicating no significant differences in coefficients across response thresholds (Chi-Square=1.88, df=3, p=0.598). The Cox and Snell (R-squared=0.021) and Nagelkerke (R-squared = 0.024) pseudo-coefficients confirm that personal professional variables do not explain variance in satisfaction. Teaching experience is also a non-significant predictor (Beta=-0.05, p=0.532), indicating that veteran and novice teachers experience the same integration hurdles.

Furthermore, a Spearman rank correlation between digital competence and structural complaints confirms this complete lack of mitigation (rs = 0.048, p = 0.707).

Rather than individual competence serving as a protective buffer, systemic first-order barriers operate as an absolute, indiscriminate environmental tax on classroom efficiency, disrupting workflows and inducing technostress uniformly across all teacher profiles. This statistical independence supports the paradox of competence narrative, challenging linear models that assume digital upskilling naturally resolves integration challenges (Caena & Redecker, 2019; Suzer & Koc, 2024).

In TAM, digital competence is theorized to predict perceived usefulness and ease of use (Antonietti et al., 2022). However, where severe first-order barriers (such as unreliable hardware, network dropouts, and rigid software boundaries) are absolute, they sever the expected link between competence and ease of use (Antonietti et al., 2022). When first-order deficits act as absolute environmental constraints, even highly digitally competent teachers (Innovators) find self-efficacy operationally useless (Caena & Redecker, 2019; Hu et al., 2025; Suzer & Koc, 2024). Advanced skills afford them a critical diagnostic gaze that exposes the mismatch between the theoretical didactics of standardized scenarios and classroom realities, shifting the locus of technostress from personal inadequacy to systemic deficits (Ajani, 2024; Alptekin & Taneri, 2025; Tarafdar et al., 2024). This is reinforced by reflexive thematic analysis. Teachers across all profiles reported experiencing severe first-order barriers that generated acute technostress, noting that executing digital scenarios requires "significantly more time for preparation" and is regularly hindered by "frequent technical problems".

4.2 Addressing RQ2

The second research question (RQ2) examined how standardized interactive digital scenarios influence K-12 students' cognitive load, reading behaviors, and knowledge retention as observed by teachers. To evaluate these cognitive and behavioral outcomes, paired-samples t-tests were conducted using the active user dataset (N=62, df=61). The results show a statistically significant negative divergence across four out of five core pedagogical parameters, demonstrating that the interactive design frequently failed to deliver its anticipated benefits. These metrics are summarized in Table 3. In the table positive t-values indicate areas where the transition to digital tools increased student dependency or gaming behaviors, while negative t-values indicate areas where actual classroom outcomes fell significantly short of initial pedagogical expectations.

To establish methodological transparency regarding the paired-samples t-tests, baseline programmatic pedagogical expectations (representing the intended curricular outcomes of the national digitalization project) were paired with active teachers' post-implementation quantitative assessments. The operationalization and pairing of these constructs are defined as follows:

1. Flexible Independent Study: Baseline expectations of independent student learning are paired with post-implementation ratings on the Q9 item "The instructional process is easy to follow."
2. Softer Skills Development: Baseline expectations of collaborative classroom synergy are paired with post-implementation ratings on the Q9 item "Allows adaptation of tasks to different social forms of work."
3. Data Interpretation & Reasoning: Baseline expectations of active scientific inquiry are paired with post-implementation ratings on the Q9 item "Encourages students to reflect on the topic."
4. Independent Learning Strategies: Baseline expectations of student self-regulated learning are paired with post-implementation ratings on the Q9 item "Encourages students to independently build upon knowledge."
5. Skills for the Modern Workplace: Baseline expectations of practical competency acquisition are paired with post-implementation ratings on the Q9 item "Makes the applicability of knowledge in the real world visible."

Table 3. Paired-samples t-test comparing pedagogical expectations vs. observed classroom realities

Pedagogical Dimension	t	p	Observed Didactical Discrepancy (Expectation vs. Reality)
Flexible Independent Study	-3.78	0.001	Teachers expected digital scenarios to foster structured, self-regulated learning, but observed unstructured pacing and erratic scrolling.
Softer Skills Development (Teamwork/ Communication)	-4.12	0.001	Collaborative gains were significantly lower than expected due to task division struggles and off-task behavior.
Data Interpretation & Reasoning	-1.71	0.092	Student gains in scientific reasoning were lower than anticipated, as students relied on immediate visual feedback rather than deep

			analytical evaluation.
Independent Learning Strategies	3.12	0.003	The transition to digital scenarios did not build student independence; instead, it increased reliance on external automated prompts.
Skills for the Modern Workplace	3.95	0.001	Standardized tools failed to build real-world competencies because students treated the interactive scenarios as simplistic games.

These findings reveal a complex, multi-dimensional gap between baseline pedagogical expectations and observed classroom realities. The divergence in the signs of the t-statistics highlights a methodological tension characteristic of convergent parallel mixed-methods designs, where structured quantitative ratings diverge from qualitative thematic reflections.

First, regarding screen-mediated reading behaviors, a negative divergence was observed in Flexible Independent Study ($t = -3.78$, $p = 0.001$). This indicates that the practical ease of following digital scenarios fell significantly short of initial expectations. Teachers' qualitative reports help explain this deficit, indicating that students struggled with non-linear reading patterns, frequently defaulting to erratic scrolling, rapid visual scanning, and skimming through text layouts, which disrupted the construction of cohesive mental schemas.

Second, in terms of cognitive load, the results show a highly significant negative divergence for Softer Skills Development ($t = -4.12$, $p = 0.001$) and a marginally negative, non-significant trend for Data Interpretation and Reasoning ($t = -1.71$, $p = 0.092$). While the capacity for deep scientific reasoning did not experience a statistically significant decline, the overall findings suggest that unguided visual interactivity often overburdens working memory. Teachers observe that students struggle to coordinate and integrate text with diagrammatic elements. This spatial split-attention effect leads to cognitive passivity, where students rely on immediate, interactive feedback to rapidly guess answers rather than engaging in sustained, active conceptual processing.

Third, the analysis reveals a striking methodological divergence regarding student self-regulation and long-term knowledge retention. For Independent Learning Strategies ($t = 3.12$, $p = 0.003$) and Skills for the Modern Workplace ($t = 3.95$, $p = 0.001$), the positive t-values indicate that teachers' quantitative evaluations of the scenarios' capacity to support independent knowledge construction and real-world applicability exceeded initial baseline expectations.

However, the qualitative findings reveal an underlying friction. While the standardized digital designs are quantitatively recognized as supportive, in practice, this structured support degenerates into prompt-dependency and gaming behaviors. Because the digital scenarios provide permanent, non-fading automated scaffolding, students bypass active metacognitive monitoring and treat the tasks as low-stakes, trial-and-error games.

However, to present a balanced discussion, the positive aspects of ICT integration must be acknowledged alongside these pedagogical challenges. The broader scientific trials associated with the digitalization project documented substantial positive affective and motivational outcomes. Observers noted marked increases in student engagement, curiosity, and excitement when digital scenarios were introduced, particularly in subjects such as biology and mathematics. Qualitative feedback from teachers confirmed that these scenarios successfully 'break the monotony of traditional teaching' and motivate disengaged or hyperactive students through visually rich, interactive media. Furthermore, the observed classroom friction may not stem from intrinsic flaws in digital mediums but rather from transitional curricular adjustments, as teachers and students adapt to screen-mediated paradigms without prior instruction in digital reading strategies. Finally, systemic deficits, such as inadequate teacher preparation time and rigid physical classroom layouts, partially explain this friction, shifting the focus of the challenge from digital technology itself to broader educational support infrastructures.

5 Conclusion

The thematic analysis of open-ended responses highlighted a cognitive vulnerability embedded within digital pedagogies, providing an answer to RQ2. Under a screen-mediated learning paradigm, the digital environment alters reading behaviors, promoting scanning and browsing that shifts cognitive engagement toward shallow processing (Liu, 2005). Because information is presented in fragmented, interactive segments, students are conditioned to navigate the interface rapidly, prioritizing task completion over the effort required for schema construction in working memory (Sweller et al., 2011).

Teachers' qualitative observations that students habitually "skim the lessons" are highly consistent with contemporary reading research showing that reading on screens habituates students to non-linear reading patterns, keyword spotting, and rapid visual scanning

(Altamura et al., 2025; Liu, 2005). Furthermore, digital environments lack the tactile and spatial anchors of paper, which are critical for cognitive mapping and comprehension (Mason et al., 2024). Without deliberate, design-level interventions to promote active, generative engagement, such as structured self-explanation prompts or intentional pacing, screen reading inevitably encourages speed over internalization, resulting in weak conceptual consolidation and rapid memory decay (Mason et al., 2024).

The immediate availability of information on screens facilitates intention offloading, leading to cognitive atrophy where students conserve mental energy by treating the digital scenario as a prosthetic memory bank. This is captured in qualitative data - teachers observed that "students feel no need to memorize data" because everything is accessible at a single click.

In cognitive psychology, the decision to offload is mediated by metacognitive monitoring and evaluation of retrieval costs (Gilbert et al., 2023; Grinschgl et al., 2021). When physical or cognitive costs of externalizing are low, individuals display a positive reminder bias, outsourcing memory storage even when working memory is fully capable (Gilbert et al., 2023). However, this immediate performance boost comes at a direct cost to subsequent memory. Because the brain does not expend the cognitive effort required to encode the information internally, recall accuracy is severely diminished (Grinschgl et al., 2021). Teachers are witnessing this pattern of detrimental cognitive offloading in real time. Students execute interactive steps, but because they have outsourced the factual parameters, "little remains in their heads" once the screen is turned off (Grinschgl et al., 2021; Skulmowski, 2023).

This cognitive offloading bias is problematic in K-12 education due to the developmental trajectory of student's metacognitive abilities. Younger students (primary school) frequently suffer from an under-offloading bias, overestimating prospective memory and failing to utilize helpful digital reminders (Sun et al., 2025). As students transition into secondary school, their metacognitive monitoring improves, but they can easily develop an over-offloading bias, using technology as a cognitive crutch to bypass memory effort entirely (Armitage & Gilbert, 2025; Sun et al., 2025). Digital scenarios that present information in highly supportive, low-cost interfaces fail to support this developmental transition. Instead of adaptive scaffolding that teaches students how to balance internal memory with external resources, current designs facilitate indiscriminate offloading, resulting in cognitive passivity across all age groups (Armitage & Gilbert, 2025; Skulmowski, 2023).

Furthermore, the analysis revealed a severe limitation regarding educational inclusivity. Teachers noted that the rigid design of current platforms disproportionately affects vulnerable demographics, stating that "students with difficulties very often struggle more". Standardized digital media presenting excessive visual noise and rapid interactive loops construct new, invisible barriers to learning for neurodivergent students and those with special educational needs (Burgstahler, 2015). Under Skulmowski's (2023) model of digital externalization, high extraneous cognitive load acts as a severe roadblock for students with lower working memory capacity or executive functioning challenges. When an interface is visually cluttered and lacks adaptive, structured scaffolding, it overwhelms the limited working memory capacity of vulnerable learners, preventing basic information from being processed or integrated (Skulmowski, 2023). This exacerbates existing educational inequities, highlighting the urgent need to move beyond standard, "one-size-fits-all" software designs toward proactive, inclusive digital architectures.

To shape the next wave of intelligent systems safely and effectively, developers and policymakers must move beyond metrics of mere interactive engagement. Systems must be engineered proactively with Universal Design for Learning frameworks to prevent the digital exclusion of neurodivergent learners and those with special educational needs (Burgstahler, 2015). Most crucially, future AI-driven educational platforms should introduce algorithmic pacing and deliberate pedagogical friction, such as mandatory reflection pauses, to disrupt skimming habits, counter cognitive offloading, and force the deep cognitive processing required for genuine, long-term knowledge retention.

References

Ajani, O. A. (2024). Teachers' competencies in digital integration of learning contents in dynamic classroom practices: A review of teacher professional development needs. Acta Educationis Generalis, 14(3), 91-111. https://doi.org/10.2478/atd-2024-0016

Alptekin, Z., & Taneri, A. (2025). Technology integration in pedagogical processes: digital competence and teaching practices of primary school teachers in Turkey. Discover Education, 4, 351. https://doi.org/10.1007/s44217-025-00646-9

Altamura, L., Vargas, C., Naumann, J., & Salmerón, L. (2025). How digital reading habits shape reading motivation and comprehension over time: longitudinal associations in primary school. Developmental Psychology, 61(2), 204-219. https://doi.org/10.1037/dev0002094

Antonietti, C., Cattaneo, A., & Amenduni, F. (2022). Can teachers' digital competence influence technology acceptance in vocational education?

Computers in Human Behavior, 132, 107266. https://doi.org/10.1016/j.chb.2022.107266

Armitage, K. L., & Gilbert, S. J. (2025). The nature and development of cognitive offloading in children. Child Development Perspectives, 19(2), 109-115. https://doi.org/10.1111/cdep.12532

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77-101. https://doi.org/10.1191/1478088706qp063oa

Braun, V., & Clarke, V. (2021). Thematic analysis: A practical guide. London: Sage.

Burgstahler, S. E. (2015). Universal Design in Higher Education: From Principles to Practice (2nd ed.). Cambridge: Harvard Education Press.

Caena, F., & Redecker, C. (2019). Aligning teacher competence frameworks to 21st century challenges: The case for the European Digital Competence Framework for Educators (DIGCOMPEDU). European Journal of Education, 54(3), 352-369. https://doi.org/10.1111/ejed.12345

Cohen, L., Manion, L., & Morrison, K. (2018). Research Methods in Education (8th ed.). London: Routledge. https://doi.org/10.4324/9781315456539

Dignath, C. (2021). For unto every one that hath shall be given: teachers' competence profiles regarding the promotion of self-regulated learning moderate the effectiveness of short-term teacher training. Metacognition and Learning, 16(3), 555-594. https://doi.org/10.1007/s11409-021-09271-x

Gilbert, S. J., Boldt, A., Sachdeva, C., Scarampi, C., & Tsai, P. C. (2023). Outsourcing memory to external tools: A review of 'intention offloading'. Psychonomic Bulletin & Review, 30(1), 60-76. https://doi.org/10.3758/s13423-022-02139-4

Grinschgl, S., Papenmeier, F., & Meyerhoff, H. S. (2021). Consequences of cognitive offloading: Boosting performance but diminishing memory. Quarterly Journal of Experimental Psychology, 74(9), 1477-1496. https://doi.org/10.1177/17470218211008060

Hu, X., Xu, W., Wan, Z., Liu, M., & Xu, X. (2025). Bridging self-efficacy and digital competence: A comprehensive scoping review of teachers' readiness for the digital age. SAGE Open, 15(3), 1-25. https://doi.org/10.1177/21582440251363716

Jian, Y. C., & Cheung, L. Y. T. (2025). A longitudinal eye-movement study of text-diagram integrative processing during multimedia reading among upper elementary children. Reading and Writing, 38(3), 557-578. https://doi.org/10.1007/s11145-023-10509-0

Karlen, Y., Hirt, C. N., Jud, J., Rosenthal, A., & Eberli, T. D. (2023). Teachers as learners and agents of self-regulated learning: The importance of different teachers competence aspects for promoting metacognition. Teaching and Teacher Education, 125, 104055. https://doi.org/10.1016/j.tate.2023.104055

Liman Kaban, A., & Karadeniz, S. (2021). Children's reading comprehension and motivation on screen versus on paper. SAGE Open, 11(1), 1-11. https://doi.org/10.1177/2158244020988849

Liu, Z. (2005). Reading behavior in the digital environment: Changes in reading behavior over the past ten years. Journal of Documentation, 61(6), 700-712. https://doi.org/10.1108/00220410510632040

Mason, L., Ronconi, A., Carretti, B., Nardin, S., & Tarchi, C. (2024). Highlighting and highlighted information in text comprehension and learning from digital reading. Journal of Computer Assisted Learning, 40(2), 639-653. https://doi.org/10.1111/jcal.12903

Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. Trends in Cognitive Sciences, 20(9), 676-688. https://doi.org/10.1016/j.tics.2016.07.002

Skulmowski, A. (2023). The cognitive architecture of digital externalization. Educational Psychology Review, 35(4), 101. https://doi.org/10.1007/s10648-023-09818-1

Sun, J., Wang, Y., Hu, Z., Liu, Y., & Sun, Y. (2025). Cognitive offloading bias in primary and secondary school students and its relationship with metacognitive monitoring. British Journal of Developmental Psychology, 43(1), 1-18. https://doi.org/10.1111/bjdp.12540

Suzer, E., & Koc, M. (2024). Teachers' digital competency level according to various variables: A study based on the European DigCompEdu framework in a large Turkish city. Education and Information Technologies, 29, 22057-22083. https://doi.org/10.1007/s10639-024-12711-1

Sweller, J., Ayres, P., & Kalyuga, S. (2011). Cognitive Load Theory. New York: Springer. https://doi.org/10.1007/978-1-4419-8126-4

Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. Educational Psychology Review, 31(2), 261-292. https://doi.org/10.1007/s10648-019-09465-5

Tarafdar, M., Tu, Q., Ragu-Nathan, B. S., & Ragu-Nathan, T. S. (2007). The impact of technostress on role stress and productivity. Journal of Management Information Systems, 24(1), 301-328. https://doi.org/10.2753/MIS0742-1222240109

Tarafdar, M., Tu, Q., & Wang, K. (2024). Technostress and professional well-being: A longitudinal study of technology-induced anxiety. Information & Management, 61(2), 103-121.

Evaluation of Student Assignments Using Local Large Language Models and Graph-Based Evidence

Neven Pintarić
University of Zadar
Department of Information Sciences and Technologies
Ul. dr. F. Tuđmana 24i, Zadar, Croatia
neven@unizd.hr

Ivan Peraić
University of Zadar
Department of Applied Communication Studies
Ul. dr. F. Tuđmana 24i, Zadar, Croatia
iperaic@unizd.hr

Abstract. *The paper presents an experimental system, Graph2LLM, for evidence-based graph evaluation of student term papers using a locally implemented LLM. The system first deterministically maps paper content into a graph with linked evidence and then uses the local LLM to validate the assignment requirement nodes. Evaluation of 25 student papers across two repeated runs shows that the deterministic mapper identifies 132 out of 150 AR4–AR9 instances. Through the local LLM, 129 instances were validated per run. The experimental results demonstrate 94.57% agreement of LLM decisions, 98.45% agreement on evidence sufficiency, and an average confidence difference of 0.0229. A human evaluation of all 150 AR instances was carried out by two evaluators. Differences were resolved by consensus. The comparison shows that the system is precise where it affirms full coverage of requirement, but stricter than human evaluators elsewhere.*

Keywords. automated essay scoring, evidence-based evaluation, graph-based evaluation, DigComp, local LLM

1 Introduction

Large language models (LLMs) have become relevant tools in the field of education, especially in the analysis and generation of textual content and in the automated evaluation of student papers. Research (Song et al., 2024) points to the potential of LLMs in the evaluation of student papers, while research (Narreddy, Joordens, & Prompiengchai, 2025) shows the possibility of integrating them into different educational contexts, including automatic assessments conducted by teachers.

The classic approach in which the student paper is sent directly to the language model reveals several methodological limitations documented in recent research: inconsistencies and variability of outputs (Pack, Barrett, & Escalante, 2024), reduced transparency and inability to revise evaluation decisions (Kumar & Boulanger, 2021), and inconsistent application of criteria due to sensitivity to query formulation (prompt sensitivity problem).

Taken together, these limitations suggest that direct LLM evaluation can reduce the repeatability, verifiability and consistency, which is problematic in educational contexts where evaluation must be methodologically grounded and equally applicable to all students.

A further limitation relates to transparency. In direct evaluation using an LLM, it is often unclear which parts of the student paper support a model decision, and it is not easy to verify the textual evidence on which assessment is based (Xu et al., 2026). This reduces the auditability and reliability of the system, which is one of the key issues in educational contexts.

Existing systems are mainly focused on generating a final assessment or certain aspects of feedback in isolation, while structured and integrated feedback that includes the coverage of task requirements and guidelines for improvement is often absent (Narreddy, Joordens, & Prompiengchai, 2025). In practice, students receive little information about which parts of their work are missing, which task requirements are met, and how the quality of their work can be improved (Li et al., 2026). The absence of these elements reduces the potential of LLMs as learning support systems.

These problems are further reinforced when evaluating a larger set of seminar papers in the same field. In such cases, it is expected that the system produces stable and comparable results for input that is identical or similar in content, without significant differences between individual evaluations, while maintaining a uniform criterion for comparing student papers.

Student work may contain personal data: information that can identify an individual, directly or indirectly. Sending such content to external LLM services therefore raises questions not only of compliance with data protection, but also of intellectual property. A locally implemented LLM addresses both of these challenges by ensuring that all data remains within the institutional infrastructure, without being transferred to third-party services (Zesch et al., 2024).

To address these limitations, this paper introduces Graph2LLM, an experimental system for evidence-based graph evaluation of student papers using a locally implemented LLM. The local LLM does not serve as a primary evaluator, but as a part of a structured process that includes the construction of a reference graph, deterministic content mapping, and evidence-based node validation. In its experimental phase, Graph2LLM is a teacher support and quality control instrument for verifying the coverage of assignment requirements (AR); it does not produce grades or scores for student evaluation.

The paper makes five contributions: a.) Graph2LLM architecture that separates deterministic mapping from LLM validation, b.) an evidence-based evaluation approach built on a graph representation, c.) empirical verification of reproducibility, d.) evaluation on 25 student term papers (STP) through two repeated runs with a locally implemented LLM and e.) a human evaluation of all 150 AR instances.

Building on these contributions, this work addresses the following research questions:

RQ1: Does the deterministic graph-based mapper produce consistent and rule-governed AR node identification across all student seminar papers when applied repeatedly under identical conditions, regardless of input document quality?

RQ2: Does a locally implemented large language model produce consistent validation decisions for assignment requirement nodes when the same evidence-based payload is sent repeatedly under identical conditions, and does the AR5 decision remain equally stable when it is additionally validated against the DigComp 3.0 reference context?

RQ3: Can an evidence-based approach built on a graph representation distinguish between fully covered, partially covered, missing, uncertain, and potentially wrongly mapped assignment requirement nodes?

RQ4: On which AR nodes, and for what reasons, do the decisions of the proposed architecture differ from the consensus decisions of human evaluators?

This study is presented as follows. Section 2 reviews related work on LLM-based paper evaluation, automated essay scoring, rubric-based approaches, and graph-based assessment. Section 3 describes the architecture and methodology. The experimental setup and results are presented in Section 4. The findings and limitations are discussed in Section 5, and the conclusions are given in Section 6.

2 Literature Review

2.1 LLM in Evaluation of Student Papers

The development of LLMs has opened new possibilities for applying artificial intelligence in educational assessment, in evaluating student essays and other written work. Unlike traditional systems that rely on predefined linguistic features, modern LLMs enable analysis of the full text and consider the semantic, argumentative and structural characteristics of written work (Pack, Barrett, & Escalante, 2024).

Existing research shows that LLMs can achieve a substantial level of agreement with the assessments of human evaluators. For example, Ishida et al. (2024) found that using clearly defined rubrics and pairwise comparison of papers results in high correlation between model and teacher assessments. Similarly, Liu, Ye, & Yan (2026) show that some modern models can achieve a high level of consistency and alignment with expert raters when evaluating student essays.

However, research also points to methodological limitations. Among the most frequently noted problems are variability of results across repeated runs, sensitivity to prompt formulation, and differences in the interpretation of evaluation criteria (Huang et al., 2025; Pack, Barrett, & Escalante, 2024). In addition, the evaluation decisions generated by LLMs are often characterized by limited transparency, since it is not always possible to determine which parts of the student paper form the basis for a particular assessment (Kostic et al., 2024). Chiang et al., (2024) also warn that models do not always consistently follow the given evaluation criteria and can be susceptible to manipulation through prompt design.

These findings indicate that, despite the substantial potential of LLMs in educational assessment, questions of reliability, transparency and verifiability of evaluation decisions remain open. For this reason, recent research focuses on structured and evidence-based approaches that link evaluation decisions to explicit textual evidence and enable their subsequent verification and revision.

2.2 Automated Essay Scoring

Automated Essay Scoring (AES) refers to the use of computational methods for the automatic evaluation of written work. Early AES systems were based on statistical models and surface linguistic features such as text length, word frequency and grammatical correctness, with systems such as PEG and e-rater laying the foundations for automated assessment in education (Elmassry et al., 2025).

Advances in machine learning and deep learning enabled a shift from manually defined features to models that autonomously learn text representations. The use of neural networks, transformers and pretrained models such as BERT improved the systems' ability to recognize semantic and discourse patterns and increased agreement with human evaluators (Elmassry et al., 2025; Mughal et al., 2026). At the same time, AES systems have shown substantial potential to reduce grading time and increase the consistency of assessment, in settings with a large number of student papers (García-Varela et al., 2025). Despite this progress, AES systems still face limitations. Models based on surface linguistic features

often fail to adequately assess the quality of argumentation, textual coherence, creativity and the relevance of content to the assigned task (Atkinson & Palma, 2025). In addition, numerous studies warn of problems with generalization across domains, sensitivity to scoring criteria, and challenges related to fairness and explainability of models (Elmassry et al., 2025; Mughal et al., 2026; García-Varela et al., 2025).

Although modern models achieve better results compared to human raters, questions of transparency and verifiability of evaluation decisions remain open research challenges.

The problem of explicitly linking the assigned score to concrete textual evidence that supports the evaluation decision is pronounced. For this reason, recent research turns toward approaches that enable transparent tracking of the evaluation process and the subsequent verification and revision of automatically generated assessments (Xu et al., 2026).

2.3 Rubric-Based Evaluation

For the structured evaluation of student work, rubrics have been most often used as tools lately. The rubrics are based on predefined criteria through which different aspects of the quality of work, such as content, organization, and language quality, are evaluated separately (Yavuz, Çelik & Yavaş, 2025).

The evaluation process with this method is more transparent and allows students to gain insight into the reasoning behind the grade.

Their further application can be seen in order to reduce the subjectivity of evaluation and increase consistency among evaluators. Furthermore, research shows that rubrics not only serve as an assessment tool, but also as a mechanism for providing structured feedback, thus expanding their contribution to the development of writing skills as well as self-regulated learning (Alghizzi and Alshahrani, 2024).

This approach is also applied in modern automated evaluation systems, which integrate rubrics into the evaluation process to improve the reliability and explainability of the results.

Research (Yavuz et al., 2025) shows that large language models can achieve a high level of agreement with human evaluators when evaluating work according to predefined analytical criteria. Similarly, Vangelova and Gancheva (2026) propose a framework that combines large language models, Retrieval-Augmented Generation (RAG), and analytic rubrics so that evaluation decisions are based on clearly defined criteria and contextual evidence.

While such systems show positive aspects, rubric-based systems often face challenges related to the interpretation of criteria and the justification of the grades awarded. Modern models achieve high accuracy but often do not provide a sufficient explanation of how individual criteria influenced the final estimate (Kumar and Boulanger, 2021).

2.4 Graph-Based Approaches and Evidence-Based Assessment

Graphs provide a framework for modelling relationships between entities, and knowledge graphs have been applied across various domains (Ji et al., 2022). In education, graph-based approaches model learning outcomes and educational structures, giving an organised representation of concepts and their interconnections that supports content recommendation, learning-outcome prediction, and educational evaluation (Qu et al., 2024). Entities are represented as nodes and their relationships as edges, which makes the underlying data interpretable (Li et al., 2024).

Progress in LLM has motivated research into combining them with graph structures, so as to join the semantic understanding of text with an explicit representation of structural relationships in the data. Such a combination links the implicit knowledge held in model parameters with explicitly represented facts, which improves interpretability and reasoning as well as the reliability of generated responses (Li et al., 2024; Pan et al., 2023). LLMs interpret textual content effectively, whereas graph structures expose relationships that often remain implicit in the text. GraphRAG illustrates this: it uses a graph-based knowledge representation to organise and summarise information from large text corpora and to analyse related concepts and entities (Edge et al., 2024).

In educational measurement, evidence-based assessment has become important. Evidence-Centered Design (ECD) is a prominent example, grounding assessment decisions in an explicit chain of reasoning that links student responses to identified evidence and to the constructs being assessed (Arieli-Attali et al., 2019). This contributes to the validity, transparency, and interpretability of the assessment process. ECD has recently been applied to LLM-based conversational assessment, where it structures prompts so that the model elicits and records evidence of student knowledge (Zapata-Rivera et al., 2024).

These principles also extend to graph-based approaches, where graphs do more than represent knowledge: they also guide reasoning and support factchecking and validation of LLM output (Ma et al., 2025). Rubric-based assessment, explainable artificial intelligence, and graph-based knowledge representation have all advanced, yet integrating AR, student content, and explicit textual evidence within a single evaluation framework remains an active research problem.

3 Methodology

3.1 System Overview

The proposed methodology evaluates STP through a graph-based, evidence-based validation process. The

system is implemented as a two-node distributed architecture. The orchestration layer runs on a computer in a combination of Windows, Python, Django with a PostgreSQL database. A local LLM validation layer was established on a MacBook Pro M4 computer with 48 GB RAM through LM Studio (0.4.12) with the locally implemented model *qwen3.5-35b-a3b-uncensored-hauhaucs-aggressive* (GGUF, Q5_K_M). This variant was chosen deliberately. An uncensored LLM is one whose safety training has been removed, and this removal is known to weaken the general capability of the model and to make its output less predictable (Qi et al., 2024; Young, 2026). It is therefore the hardest case for a system that must produce repeatable decisions. The design intention is that stability should be achieved through the architecture described in the following sections, and not through the LLM model itself. Communication between nodes and layers takes place over a local network through an OpenAI-compatible REST API endpoint.

Rather than directly evaluating the full text of a student term paper (STP) using a local LLM, Graph2LLM applies a structured four-layer processing pipeline, illustrated in Fig. 1.

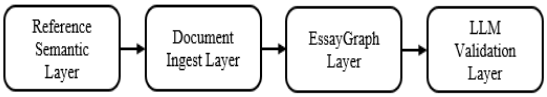


Figure 1. Graph2LLM processing pipeline

The reference semantic layer, comprising the DigComp Graph and the Assignment Graph, is constructed in advance and serves as the foundation for all subsequent processing. The system then converts each STP into a structural graph, deterministically maps it against predefined AR, and only then validates individual requirement nodes using the local LLM on the basis of evidence segments linked to each node. Each layer is described in detail in the following subsections.

3.2 Formal Graph Layers

Three formal graph layers were defined: DigComp Graph, Assignment Graph and Essay Graph.

Serving as reference model of digital competences, the DigComp Graph is modelled on the basis of the publicly available JSON-LD dataset for DigComp 3.0 framework (Directorate-General for Employment, Social Affairs and Inclusion, 2025). It contains competence areas, competences and learning outcomes with associated properties such as outcome types and proficiency levels. In this layer, the node is the primary structural unit of the graph, carrying learning outcomes as properties that serve as the evidential basis for validation. Learning outcomes represent a measurable and operationalizable foundation for automated assessment (Divjak, Svetec, & Horvat, 2024). It is used exclusively where the task requirement directly asks the student to identify and describe a DigComp area, competences and learning outcomes.

Capturing the pedagogical model of the STP; the Assignment Graph defines the expected requirements through ten requirement nodes (AR1–AR10). These requirements include formal document elements, title page (AR1), table of contents (AR2) and references (AR10), and substantive elements, including an introduction (AR3), a definition of digital competence (AR4), a selected DigComp area (AR5), a self-assessment (AR6), relevance to early childhood education practice (AR7), a personal development plan (AR8) and a conclusion (AR9).

The Essay Graph is created from the provided STP. Segment nodes preserve the order of appearance in the document, ensuring both structural integrity and traceability to the original text. Any subsequent validation decision must be linked to concrete evidence from the STP.

3.3 Document Ingest Layer

The evaluation process begins with uploading an STP document that is stored as a document object with a basic metadata. The ingest stage extracts document pages (*DocumentPage*), text segments (*DocumentSegment*), and evidence records (*Evidence*). This phase is deterministic and does not involve a local LLM. After the ingest phase, an initial structural *EssayGraph* is built in which each relevant document segment becomes an *EssayNode*, while the relationships within the document are represented as *EssayEdge* objects.

3.4 Deterministic Assignment Mapping

This step constructs the semantic assignment layer through a deterministic mapper that implements an architectural principle, considering prior work showing that evidence components must be explicitly identified and structured before LLM scoring (Wang et al., 2026). Unlike LLM-based extraction approaches, mapping in this work is implemented deterministically through configurable *AssignmentRequirementProfile* records. Each profile describes a single task requirement and contains parameters such as requirement code, name, requirement type, expected order, title aliases, negative patterns, expected evidence types, evidence policy, visible support segment limits, and acceptance thresholds.

The system groups document segments into sectional units and compares each sectional group with the active requirement profiles. Structural segmentation in unstructured documents according to section headings with variable formats represents an NLP problem (Tannous, Ramadan, Rajab, 2023).

The result of the mapping is calculated by combining multiple signals and heuristic value: section title similarity (0.34), structural order (0.20), section

number (0.14), evidence type compatibility (0.10), preface bonus (0.08), content keyword match (0.14), negative patterns, and TOC penalty. These signals are aggregated into a weighted mapping score with fixed weights calibrated on baseline documents. Each candidate mapping is then classified as accepted, rejected, or flagged for review. Only accepted candidates become semantic AR nodes in *EssayGraph*.

As an input control before local LLM validation, a baseline graph diagnostic is performed, which checks for the presence of expected AR nodes, missing nodes and suspicious mappings.

Deterministic mapping is reproducible under the same input conditions: the same document, the same version of the ingest procedure, the same configuration of the *AssignmentRequirementProfile* records, and the same version of the code produce the same result. Consistency in this context does not necessarily imply correctness, but rather the repeatability and auditability of mapping decisions. This is particularly important in the educational context, where transparency, verifiability and ex-post auditability of automated evaluation decisions are considered methodological requirements (Kim et al., 2025).

3.5 Assignment Requirement

Each AR node belongs to one of two methodological categories. Structural requirements describe formal document properties and include AR1, AR2 and AR10. Content requirements are mapped to actual textual sections of the paper and include AR3, AR4, AR5, AR6, AR7, AR8 and AR9. Although AR3 belongs to the content requirements, it is excluded from the initial local LLM validation set because introductions are generally of a general nature and do not contain a specific DigComp validation structure. Accordingly, local LLM validation in this paper is focused on AR4–AR9.

3.6 Evidence-Based Local LLM Validation

LLM-based validation has known limitations (Zheng et al., 2023). This can be mitigated by clearly defined multi-dimensional rubrics, in which the model does not produce a holistic assessment but instead responds to pre-defined evaluation questions (Hashemi et al., 2024).

After deterministic mapping, the local LLM validation layer validates identified individual AR nodes with explicitly linked evidence segments. The model does not receive the entire STP, but rather a structured payload for a single AR node containing document metadata, graph identifier, assignment node code and name, requirement description, section headings, deterministic mapper confidence, and evidence texts linked to that AR node. In this way, validation is limited to the verification of a clearly defined requirement and the associated evidence.

This logic is operationalised in the local LLM validator prompt. The prompt defines the model as a JSON-only validation API and explicitly requires the response to be a single valid RFC8259 JSON object with no additional text, comments, markdown or visible reasoning. Furthermore, the prompt instructs the model to use only *the submitted evidence_texts* and the reference context provided, without using outside knowledge, fabricating facts, or assuming that the missing content exists elsewhere in the document.

Every response from the local LLM must return a structured JSON response validated against a pre-defined schema. It includes the assignment node code, decision, evidence sufficiency, detected requirement, confidence, problems, reasoning summary and an object for DigComp reference validation. Permitted decisions are *fully_covered*, *partially_covered*, *missing*, *wrong_mapping* and *uncertain*, interpreted under a minimum evidence-sufficiency threshold: a node with insufficient semantic content is *marked* missing with *evidence_sufficient=false* rather than *partially_covered*.

For content requirements AR6, AR7 and AR9, a stricter interpretation rule is applied: if the evidence is too short, generic, poorly reasoned or does not contain requirement-specific semantic content, the decision must be *missing* rather than *partially_covered*.

A further part of the prompt distinguishes between assignment coverage validation and DigComp reference alignment validation. If *digcomp_reference_context* is not provided, the model must not perform exact alignment with DigComp areas, competences or learning outcomes, but validates only the coverage of the task requirement. If *digcomp_reference_context* is provided, the model may use it exclusively as the supplied reference context and must not fabricate DigComp codes, areas, competences, learning outcomes or proficiency levels.

A *wrong_mapping* decision marks cases where deterministic mapper accepted candidate mapping but the local LLM judges, from the linked evidence, that it does not correspond to the assigned AR node, separating the formal presence of section from the semantic correctness of its evidence.

3.7 Assignment Coverage

Assignment coverage validation is applied to all semantic nodes included in the initial validation set, i.e. AR4–AR9. The coverage criteria differ by requirement: a defined digital competence (AR4); an explicit self-assessment with meaningful reasoning about the proficiency level and a relevant example (AR6); a concrete pedagogical application connecting the selected area or competence to the work of the early childhood educator (AR7); a development plan with a goal, a current or desired proficiency level, and a means of achievement, with partial coverage permitted only when at least two mandatory plan elements are present (AR8); and a conclusion synthesising the topic

of the paper, the selected area or competence, and the main findings (AR9).

Node AR5 is a special case, as it directly links the AR to the DigComp framework. Therefore, AR5 is validated both against the Assignment Graph and against the DigComp Graph through DigComp reference alignment validation. This allows the model to verify whether the student's description is semantically aligned with the detected DigComp area, competences and learning outcome dimensions. In the current phase, AR5 is the only node with dual validation. The results are not interpreted as a complete DigComp validation of the entire student paper.

3.8 Architecture Evaluation Metrics

The evaluation is based on node-level graph and validation metrics aligned with the research questions RQ1 to RQ3. For the deterministic mapper, we measure repeated-run agreement, AR coverage rate, missing node count and wrong mapping count. For local LLM validation, we measure decision distribution, evidence sufficiency rate, confidence score distribution and decision stability across repeated runs. Node AR5 is reported separately because it is the only node with additional DigComp Graph alignment, while AR4, AR6, AR7, AR8 and AR9 are validated only against the Assignment Graph. For each run, the system stores the AR node, evidence count, decision, confidence, prompt version, model metadata, run identifier and DigComp context usage. This makes it possible to verify the factual basis for each validation decision at a later stage.

3.9 Human Evaluation Protocol

Once the architecture had shown stable and repeatable decisions across the two runs, the next step was to test whether those decisions were also correct. Human evaluation was carried out in order to establish a reference against which the correctness of the system's decisions, and not only their reproducibility, could be assessed. Two evaluators independently coded all 150 AR instances (25 STPs × AR4–AR9). The evaluators are the authors of this research. To limit the resulting bias, the following procedure was applied: coding was blind, so that neither evaluator saw the system output or the decisions of the other, and the papers were presented in a randomised but identical order. The evaluators followed a written set of coding rules that mirrored those applied by the system, namely the minimum-sufficiency threshold, the stricter interpretation of AR6, AR7 and AR9, and the two-of-three rule for AR8. Decisions were based exclusively on the text of the paper, mirroring the constraint imposed on the local LLM by the prompt. Instances on which the two evaluators initially disagreed were resolved by consensus, in which the relevant text segment was re-read, the written rules re-examined, and the final decision reached through reasoned argument. A measurement instrument with two parts was prepared, one for each layer of the architecture. Part A records whether content addressing the requirement is present at all, irrespective of quality, and serves as the reference for the deterministic mapper. Part B assigns one of the five decision categories used by the system and serves as the reference for local LLM validation. Inter-rater agreement was computed on the independent codings; the disagreements were then resolved by consensus between the two evaluators; the reported kappa values are computed on the independent codings and are therefore not inflated by the consensus step. For the deterministic mapper, Part A provides precision, recall and F1 against the consensus presence decisions. For the local LLM validation layer, agreement with the consensus decisions in Part B is reported as a confusion matrix of system against human decisions, percentage agreement, Cohen's κ, and, for the *fully_covered* category, precision, recall and F1.

3.10 Ethics

This experimental research was conducted in accordance with the policy on the use of artificial intelligence at University of Zadar (University of Zadar, 2026), which recommends locally implemented solutions for sensitive data and prohibits automated decision making without human oversight. Approval was obtained from the Ethics Committee of the Department of Information Sciences and Technologies for the implementation of the research (decision of 10 July 2026, ref. 602-03/26-04, no. 2198-1-79-33-26-01). All processing was carried out locally on institutional infrastructure, and only aggregated results are reported, so that no individual STP can be identified. The experiment was run after the winter semester and all examination has been completed, and it had no influence on scoring or the final grades of the students enrolled in the course. Students were informed that their term papers could be used for research purposes.

4 Results

4.1 Experimental Set-up

Experimental evaluation is carried out on a set of 40 STPs from the winter course "Child and Computer", taught in the winter semester of the 2025/2026 academic year in the undergraduate university study programme in Early and Preschool Education at University of Zadar. The course includes the preparation of a seminar term paper on digital competencies according to the DigComp framework. We have three subsets of STP papers. For the purpose of development, 5 papers were used. System calibration and baseline testing were performed on 10 papers, and 25 papers were used for final testing.

Each STP goes through two independent validation runs (*run_01* and *run_02*) under identical conditions: the same LLM, temperature (0.1), prompt version (*assignment_node_validation_v1*), context window 16,384 tokens and *Assignment-RequirementProfile* configuration.

4.2 Baseline Testing and Calibration

Baseline testing was conducted on 10 STPs using the locked *AssignmentRequirementProfile* configurations for AR4–AR9. The deterministic mapper identified 49 of 60 expected AR nodes (81.67% coverage). Repeated runs produced identical mapper output, confirming deterministic reproducibility.

Local LLM validation revealed that table content in AR8 (personal development plan) was not consistently extracted as separate evidence segments. In these cases, the local LLM validated only the chapter heading rather than the table content. This resulted in systematic misclassification.

To address this, the ingest layer was extended to extract table rows as individual evidence segments. Prompt calibration was subsequently performed on a purpose-built control document containing formally correct but semantically thin chapters alongside a properly structured AR8 table. This control document was designed to verify that the local LLM distinguishes formal chapter presence from semantically sufficient coverage. After calibration, the local LLM correctly classified generic single-sentence evidence as *missing* and complete development plan tables as *fully_covered*.

All mapping weights, acceptance thresholds, negative patterns, *AssignmentRequirementProfile* configurations, prompt settings and LLM inference parameters were derived exclusively from the development and calibration subsets and were locked before the final test set was processed. The 25 final-test STPs were not used in any tuning step, which limits the risk of overfitting.

4.3 Testing Results for Run 1 and Run 2

Results for Run 1 and Run 2 are presented in Table 1. The deterministic mapper processed 150 expected AR4–AR9 instances across 25 documents, identifying 132 instances (88.00%) with identical output in both runs. Of the 21 *Not_validated* instances, 18 correspond to AR nodes not detected by the deterministic mapper, while 3 reflect local LLM processing failures in partial-status validation runs. The 7 recorded decision changes produced a net distributional shift of four instances out of *fully_covered* (38 to 34), appearing as three additional *partially_covered* and one additional *missing*, all between adjacent categories. No transitions between non-adjacent categories were observed.

Table 1. Final testing results Run 1 and Run 2

Metric	Run 1	Run 2
Expected AR4–AR9 instances	150	150
Mapper: found	132	132
Mapper: missing	18	18
Local LLM validated instances	129	129
Decision: fully_covered	38	34
Decision: partially_covered	64	67
Decision: missing	27	28
Not_validated	21	21
Evidence sufficient: True	96	96
Evidence sufficient: False	33	33

Cross-run stability metrics are shown in Table 2. Decision agreement on 129 comparable pairs reached 94.57%, with evidence sufficiency agreement of 98.45%. Seven decision changes and two evidence sufficiency changes were recorded across both runs. The average absolute confidence difference was 0.0229, indicating low numerical variability of the local LLM.

Table 2. Cross-run stability metrics (Run 1 vs Run 2)

Stability Metric	Value
Mapper agreement	150/150 (100%)
Decision agreement (comparable pairs)	122/129 (94.57%)
Evidence sufficiency agreement	127/129 (98.45%)
Decision changes	7
Evidence sufficiency changes	2
Average confidence difference	0.0229

The results per AR node are shown in Tables 3 and 4. Node AR6 and AR9 achieved complete agreement in the decision (100%).

Table 3. AR node decision stability (Run 1 vs Run 2)

AR Node	Comparable Pairs	Decision Agreement
AR4	19/25	18/19 (94.74%)
AR5	22/25	21/22 (95.45%)
AR6	19/25	19/19 (100%)
AR7	24/25	21/24 (87.50%)
AR8	23/25	21/23 (91.30%)
AR9	22/25	22/22 (100%)

The highest variability occurs in AR7 (87.50%) with 3 changes, with problems related to insufficient pedagogical application and incomplete reference context of DigComp. Changes in the sufficiency of evidence occurred only in AR8, in 2 cases, suggesting structural variation in the content of the personal development plan. Node AR5, validated against both

the Assignment Graph and the DigComp 3.0 reference context, showed high stability with a single *fully_covered* to *partially_covered* transition.

All recorded decision changes occurred between adjacent decision categories. This pattern confirms that local LLM variability is concentrated in borderline semantic cases rather than in stable instances of clear coverage or clear absence of evidence.

Table 4. AR node evidence sufficiency and confidence (Run 1 vs Run 2)

AR Node	Evid. Suff. Agreement	Avg. Confidence	Decision Changes
AR4	19/19	0.7737	1
AR5	22/22	0.7695	1
AR6	19/19	0.7563	0
AR7	24/24	0.7458	3
AR8	21/23	0.8443	2
AR9	22/22	0.6773	0

4.4 Human Evaluation Results

This section compares the system output with the human evaluation: the deterministic mapping is examined through Part A of the human coding, and the validation of the AR instances through Part B. Inter-rater agreement, computed on the independent codings before the consensus step, was near-perfect for presence (Po = 0.993, Pe = 0.980, Cohen's κ = 0.664; the low κ reflects the extreme prevalence of present, as only one instance was coded differently) and moderate for coverage (Po = 0.720, Pe = 0.488, κ = 0.453): 108 of 150 instances matched, and 32 of the 42 disagreements were between the adjacent categories *fully_covered* and *partially_covered*. The consensus coding produced 107 *fully_covered*, 26 *partially_covered*, 16 *missing* and 1 *wrong_mapping*. Part A yielded 148 present and 2 absent.

Recorded in Part A of the instrument, the deterministic mapping was measured against the consensus presence decisions. It reached a precision of 1.000, a recall of 0.892 and an F1 of 0.943 (132 true positives, 0 false positives, 16 false negatives, 2 true negatives), identically in Run 1 and Run 2. As there are no false positives, every error is a requirement the mapper did not find: the 16 false negatives are concentrated in AR6 (6), AR4 (5) and AR9 (3), with one each in AR7 and AR8.

Analysis of the AR instances, recorded in Part B, is limited to those the system actually validated. Of the 150 AR instances, the mapper identified 132 as present and these were passed to the local LLM; the remaining 18 were not, and 16 of them had been recorded as present by the evaluators. Of the 132 instances sent to the model, three AR5 calls returned no decision in either run, which leaves 129 validated instances. Assessment of the validation layer therefore rests on these 129 instances alone. Table 5 shows the confusion matrix for Run 1, with the human decision in rows and the system decision in columns. Measured end to end, with the 21 unvalidated instances counted as failures, agreement with the human evaluation is 36.7% in Run 1 and 34.0% in Run 2.

Table 5. Confusion matrix of system decision against human evaluation (Run 1, 129 validated instances)

Human	missing	partially	fully	Σ
missing	3	9	2	14
partially	4	17	1	22
fully	20	38	35	93
Σ	27	64	38	129

The system and the evaluators agree on 55 of the 129 decisions (42.6%, Cohen's κ = 0.157). Almost all errors lie entirely below the diagonal: the system chose a lower category in 62 cases and a higher one in only 12, which makes it on average 0.53 of a category stricter. Most disagreement is minor – 107 of the 129 decisions (82.9%) stay within one category of the human decision. This strictness is visible in the totals: the evaluators used *fully_covered* 93 times and the system only 38 times, lowering 20 of these instances as far as *missing*. Its positive decisions, however, are reliable. Of its 38 *fully_covered* decisions the evaluators confirmed 35, a precision of 0.921 against a recall of 0.376 and an F1 of 0.534. Run 2 gives the same picture: 51 decisions on the diagonal (39.5%, κ = 0.134), 66 stricter against 12 more lenient, and a precision of 0.912, a recall of 0.333 and an F1 of 0.488 for *fully_covered*.

Agreement with the human evaluation is unevenly distributed across AR nodes, as shown in Table 6.

Table 6. Agreement with human evaluation per AR node (Run 1)

AR Node	Validated Instances	Agreement	Stricter : More Lenient
AR4	19	53%	8 : 1
AR5	22	45%	10 : 2
AR6	19	21%	11 : 4
AR7	24	21%	16 : 3
AR8	23	74%	4 : 2
AR9	22	41%	13 : 0

It ranges from 74% on AR8, whose criterion is enumerable, down to 21% on AR6 and AR7, and the direction of divergence is consistent throughout: on every node the system is stricter more often than it is more lenient, most clearly on AR9, where all 13 disagreements go in the direction of strictness.

5 Discussion and Limitations

The deterministic mapper produced consistent and rule-governed AR node identification across all 25 STPs and both repeated runs. In this experiment, the behaviour of the mappers is deterministic, since mismatched instances appeared in the same position across runs. Incomplete coverage reflects the quality of the input document, e.g. short chapters, heading-only evidence, and implicitly stated content, rather than changes due to execution. These observations correspond affirmatively to RQ1.

In the context of answering RQ2, the local LLM validation layer demonstrated cross-run stability when the same evidence-based payload was submitted under identical conditions, whit decision agreement reached 94.57% and evidence sufficiency agreement 98.45% on 129 comparable pairs. Under dual validation, AR5 showed comparable stability, indicating that the additional alignment layer does not introduce systematic instability. Therefore, the cross-run consistency, including the stability of the AR5 decision under dual validation, is supported for RQ2.

Research question RQ3 is only partially answered. The evidence-based approach produced consistent and auditable distinctions between *fully_covered*, *partially_covered*, and *missing* across both runs, with all decision changes occurring between adjacent categories. The categories *wrong_mapping* and *uncertain* are operationally defined but were not triggered in the final test set. Their behaviour requires targeted evaluation on deliberately ambiguous inputs. The three core coverage categories are distinguished consistently and auditable, although the comparison with the human evaluation in Section 4.4 shows that the sufficiency threshold applied to them is miscalibrated relative to human judgement.

The answer to the last research question, RQ4, shows that the divergence of the system from the human decisions is not random but systematic. It is systematic in direction. Of the 74 disagreements in Run 1, the system assigned a lower category than the evaluators in 62 cases and a higher one in only 12; across all 129 decisions this makes the system on average 0.53 of a category stricter. Run 2 gives the same ratio, 66 to 12.

It is systematic in kind as well. When the system declares a requirement fully covered it is almost always right: of its 38 such decisions the evaluators confirmed 35 (precision 0.921). Yet of the 93 instances the evaluators considered fully covered, the system recognised only 35 (recall 0.376). The system is therefore not unreliable when it affirms, but too strict when it judges. It does not err coarsely either, as 107 of the 129 decisions stay within one category of the human decision.

The divergence is not evenly spread across requirements. Agreement is highest on AR8 (74%) and lowest on AR6 and AR7 (21%), with AR4 at 53%, AR5 at 45% and AR9 at 41%. On AR9 all 13 disagreements go in the direction of strictness and none in the direction of leniency.

Two findings show where the cause does not lie. First, in 37 of the 62 stricter disagreements the system itself recorded the evidence as sufficient and lowered the category nonetheless, so it was not short of evidence. Second, the system agrees with itself across the two runs almost perfectly ($\kappa = 0.912$) while agreeing with the evaluators only weakly ($\kappa = 0.157$), so these are not erratic lapses but a consistent shift.

Both findings point to the same conclusion: the problem lies in the threshold at which the system declares a requirement covered, not in the reliability of its inference. One possible source of the difference is that the system judges each AR node solely on the evidence segments assigned to it, whereas the evaluator reads the paper as a whole. This cannot be confirmed on the basis of the present data and requires further research.

The evaluation was conducted on 25 STPs, sufficient for controlled evaluation but not for broad generalization, and the results apply to a single local LLM configuration and prompt version. Because the comparison with the human evaluation covers the 129 validated instances and the mapper did not identify 16 requirements that the evaluators recorded as present, the quality and the structure of the input document directly affect what reaches the validation layer. Moreover, the human reference itself has limits: it was coded by the authors and settled by consensus rather than by an independent adjudicator, so it is not fully independent of the coders, and the pre-consensus coverage agreement ($\kappa = 0.453$) marks the ceiling against which the system's $\kappa = 0.157$ should be read rather than against perfect agreement. Between runs, the few decision changes are concentrated in AR7 (three) and AR8 (two). Mapping weights and acceptance thresholds were set by hand on the development and calibration subsets, and it is not known how well they transfer to other assignments or courses. The categories wrong_mapping and uncertain were not triggered by the system in the final test set, although the human evaluation produced one *wrong_mapping* decision, so their behaviour remains empirically untested. Each of these limitations calls for additional research.

6 Conclusion

This paper presented Graph2LLM, an experimental system for evaluating STP that combines deterministic graph-based mapping, evidence grounding, and a locally implemented LLM validation layer. Each validation decision is traceable to specific evidence, confidence scores, and run metadata, enabling post-hoc audit of the full evaluation process.

Experimental results demonstrate that the deterministic mapper produces stable and reproducible output, and that local LLM validation achieves cross-run stability, with variability concentrated in the most interpretive and structurally complex node. This

reproducibility was obtained with an intentionally unstable, uncensored local LLM, which indicates that cross-run stability stems from the architecture as a whole rather than from the model alone, although isolating the contribution of each component requires further research.

The results confirm that the approach distinguishes fully covered, partially covered and missing requirements reproducibly and auditably. Reproducibility and auditability are necessary but not sufficient conditions for educational assessment. A system can be highly stable and still be miscalibrated with respect to human judgement, as the comparison with the human evaluation shows. Responsibility for grading, and for its consequences, remains with the teacher and the institution; the role of the system is to support that work, not to replace it.

Future work follows from the comparison with the human evaluation. The sufficiency threshold of the validation layer will be recalibrated against human-coded examples from the development subset, since the divergence reported here is systematic rather than random. An ablation study, running the same model without the evidence payload and without the JSON schema, will show how much of the observed stability comes from the graph layer itself. The mapper will be extended with a semantic fallback for requirements that no rule maps, recorded as a probabilistic decision, so that determinism, the absence of false positives and auditability are preserved on the rule-based path. Beyond this, the same evidence-based payloads will be validated across several locally implemented models, and LLM-assisted construction of the semantic assignment layer will be explored.

References

Alghizzi, T. M., & Alshahrani, T. M. (2024). Effects of grading rubrics on EFL learners' writing in an EMI setting. *Heliyon*, *10*(18), e36394. https://doi.org/10.1016/j.heliyon.2024.e36394

Arieli-Attali, M., Ward, S., Thomas, J., Deonovic, B., & Von Davier, A. A. (2019). The Expanded Evidence-Centered Design (e-ECD) for Learning and Assessment Systems: A Framework for Incorporating Learning Goals and Processes Within Assessment Design. *Frontiers in Psychology*, *10*, 853. https://doi.org/10.3389/fpsyg.2019.00853

Atkinson, J., & Palma, D. (2025). An LLM-based hybrid approach for enhanced automated essay scoring. *Scientific Reports*, *15*(1), 14551. https://doi.org/10.1038/s41598-025-87862-3

Chiang, C.-H., Chen, W.-C., Kuan, C.-Y., Yang, C., & Lee, H. (2024). Large Language Model as an Assignment Evaluator: Insights, Feedback, and Challenges in a 1000+ Student Course. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2489–2513. https://doi.org/10.18653/v1/2024.emnlp-main.146

Directorate-General for Employment, Social Affairs and Inclusion. (2025). *DigComp 3.0 framework data supplement* [Dataset]. European Commission, Joint Research Centre. https://doi.org/10.2905/JRC.FR75K8R

Divjak, B., Svetec, B., & Horvat, D. (2024). How can valid and reliable automatic formative assessment predict the acquisition of learning outcomes? *Journal of Computer Assisted Learning*, *40*(6), 2616–2632. https://doi.org/10.1111/jcal.12953

Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., Metropolitansky, D., Ness, R. O., & Larson, J. (2024). *From Local to Global: A Graph RAG Approach to Query-Focused Summarization*. (arXiv preprint arXiv:2404.16130). https://doi.org/10.48550/ARXIV.2404.16130

Elmassry, A. M., Zaki, N., Alsheikh, N., & Mediani, M. (2025). A Systematic Review of Pretrained Models in Automated Essay Scoring. *IEEE Access*, *13*, 121902–121917. https://doi.org/10.1109/ACCESS.2025.3584784

García-Varela, F., Nussbaum, M., Mendoza, M., Martínez-Troncoso, C., & Bekerman, Z. (2025). ChatGPT as a Stable and Fair Tool for Automated Essay Scoring. *Education Sciences*, *15*(8), 946. https://doi.org/10.3390/educsci15080946

Hashemi, H., Eisner, J., Rosset, C., Van Durme, B., & Kedzie, C. (2024). LLM-Rubric: A Multidimensional, Calibrated Approach to Automated Evaluation of Natural Language Texts. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 13806–13834. https://doi.org/10.18653/v1/2024.acl-long.745

Huang, Y., Wilson, J., Ormerod, C., & Beiting Parrish, M. (2025). Evaluating LLM-Based Automated Essay Scoring: Accuracy, Fairness, and Validity. *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Works in Progress*, 71–83.

Ishida, T., Liu, T., Wang, H., & Cheung, W. K. (2024). *Large Language Models as Partners in Student Essay Evaluation*. (arXiv preprint 2405.18632). https://doi.org/10.48550/ARXIV.2405.18632

Ji, S., Pan, S., Cambria, E., Marttinen, P., & Yu, P. S. (2022). A Survey on Knowledge Graphs: Representation, Acquisition, and Applications. *IEEE Transactions on Neural Networks and Learning Systems*, *33*(2), 494–514. https://doi.org/10.1109/TNNLS.2021.3070843

Kim, Y., Hardy, M., Tey, J., Thille, C., & Piech, C. (2025). *Interpretability from the Ground Up: Stakeholder-Centric Design of Automated Scoring*

in Educational Assessments (Version 3). (arXiv preprint arXiv:2511.17069). https://doi.org/10.48550/ARXIV.2511.17069

Kostic, M., Witschel, H. F., Hinkelmann, K., & Spahic-Bogdanovic, M. (2024). LLMs in Automated Essay Evaluation: A Case Study. *Proceedings of the AAAI Symposium Series*, *3*(1), 143–147. https://doi.org/10.1609/aaaiss.v3i1.31193

Kumar, V. S., & Boulanger, D. (2021). Automated Essay Scoring and the Deep Learning Black Box: How Are Rubric Scores Determined? *International Journal of Artificial Intelligence in Education*, *31*(3), 538–584. https://doi.org/10.1007/s40593-020-00211-5

Li, Y., Shen, X., Miao, Y., Yao, X., Ding, X., Krishnan, R., & Padman, R. (2026). *Beyond Single-Turn: A Survey on Multi-Turn Interactions with Large Language Models*. (arXiv preprint arXiv:2504.04717). https://doi.org/10.48550/arXiv.2504.04717

Li, Y., Li, Z., Wang, P., Li, J., Sun, X., Cheng, H., & Yu, J. Xu. (2024). A Survey of Graph Meets Large Language Model: Progress and Future Directions. *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 8123–8131. https://doi.org/10.24963/ijcai.2024/898

Liu, T., Ye, L., & Yan, W. (2026). A framework for evaluation of large language models in essay assessment: Reliability, alignment, and causal reasoning. *Computers and Education: Artificial Intelligence*, *10*, 100565. https://doi.org/10.1016/j.caeai.2026.100565

Ma, C., Chen, Y., Wu, T., Khan, A., & Wang, H. (2025). Large Language Models Meet Knowledge Graphs for Question Answering: Synthesis and Opportunities. *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 24589–24608. https://doi.org/10.18653/v1/2025.emnlp-main.1249

Mughal, N., Imran, A. S., Daudpota, S. M., Kastrati, Z., & Noor, W. (2026). Exploring potential of large language models for automated essay scoring in education. *Discover Artificial Intelligence*, *6*(1), 166. https://doi.org/10.1007/s44163-026-01002-y

Narreddy, C., Joordens, S., & Prompiengchai, S. (2025). Harnessing Large Language Models for Scalable and Effective Formative Assessment in Higher Education: A Review. *Trends in Higher Education*, *4*(4), 65. https://doi.org/10.3390/higheredu4040065

Pack, A., Barrett, A., & Escalante, J. (2024). Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability. *Computers and Education: Artificial Intelligence*, *6*, 100234. https://doi.org/10.1016/j.caeai.2024.100234

Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., & Wu, X. (2023). *Unifying Large Language Models and Knowledge Graphs: A Roadmap*. (arXiv preprint arXiv:2306.08302) https://doi.org/10.48550/ARXIV.2306.08302

Qi, X., Zeng, Y., Xie, T., Chen, P.-Y., Jia, R., Mittal, P., & Henderson, P. (2024). Fine-tuning aligned language models compromises safety, even when users do not intend to! In *Proceedings of the Twelfth International Conference on Learning Representations* (ICLR 2024). 30988–31043. https://proceedings.iclr.cc/paper_files/paper/2024/file/83b7da3ed13f06c13ce82235c8eedf35-Paper-Conference.pdf

Qu, K., Li, K. C., Wong, B. T. M., Wu, M. M. F., & Liu, M. (2024). A Survey of Knowledge Graph Approaches and Applications in Education. *Electronics*, *13*(13), 2537. https://doi.org/10.3390/electronics13132537

Song, Y., Zhu, Q., Wang, H., & Zheng, Q. (2024). Automated Essay Scoring and Revising Based on Open-Source Large Language Models. *IEEE Transactions on Learning Technologies*, *17*, 1880–1890. https://doi.org/10.1109/TLT.2024.3396873

Tannous, M. E., Ramadan, W. H., & Rajab, M. A. (2023). TSHD: Topic Segmentation Based on Headings Detection (Case Study: Resumes). *Advances in Human-Computer Interaction*, *2023*, 1–12. https://doi.org/10.1155/2023/6044007

University of Zadar. (2026). Policy on the Use of Artificial Intelligence at the University of Zadar. https://www.unizd.hr/Portals/0/doc/doc_pdf_dokumenti/pravilnici/politika_koristenja_umjetne_inteligencije_na_sveucilistu_u_zadru_20260225.pdf?ver=iaHevmwh7XPEjZAGSF0lOw%3D%3D

Vangelova, A., & Gancheva, V. (2026). AI-Based Automated Scoring Layer Using Large Language Models and Semantic Analysis. *Applied Sciences*, *16*(7), 3537. https://doi.org/10.3390/app16073537

Wang, Y., Ding, Z., Wu, X., Sun, S., Liu, N., & Zhai, X. (2026). AutoSCORE: Enhancing Automated Scoring with Multi-Agent Large Language Models via Structured Component Recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, *40*(48), 40898–40906. https://doi.org/10.1609/aaai.v40i48.42123

Xu, W., Kassim, M. S. S., Hoo, W. L., Yang, W., & Xu, T. (2026). Explainable AI for education: Enhancing essay scoring via rubric-aligned chain-of-thought prompting. *International Journal of Modern Physics C*, *37*(06), 2542013. https://doi.org/10.1142/S0129183125420136

Yavuz, F., Çelik, Ö., & Yavaş Çelik, G. (2025). Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments. *British Journal of*

Educational Technology, *56*(1), 150–166. https://doi.org/10.1111/bjet.13494

Young, R. J. (2026). Comparative analysis of LLM abliteration methods: A cross-architecture evaluation (arXiv preprint arXiv:2512.13655). https://doi.org/10.48550/arXiv.2512.13655

Zapata-Rivera, D., Forsyth, C., Graf, E., & Jiang, Y. (2024). *Designing and Evaluating Evidence-Centered Design based Conversations for Assessment with LLMs*. https://par.nsf.gov/biblio/10591735-designing-evaluating-evidence-centered-design-based-conversations-assessment-llms

Zesch, T., Hanses, M., Seidel, N., Aggarwal, P., Veiel, D., & de Witt, C. (2024). *FernUni LLM Experimental Infrastructure (FLEXI)—Enabling Experimentation and Innovation in Higher Education Through Access to Open Large Language Models* (Version 1). (arXiv preprint arXiv:2407.13013). https://doi.org/10.48550/ARXIV.2407.13013

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). *Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena* (Version 4). (arXiv preprint arXiv:2306.05685). https://doi.org/10.48550/ARXIV.2306.05685

Schema Before Output: Metacognitive Schema-First Scaffolding in Computing Higher Education in the GenAI Era

Vjeran Bušelić
University of Zagreb
Faculty of Organization and Informatics
Pavlinska 2, 42000 Varaždin, Croatia
vbuselic@foi.hr

Dinko Horvat
Zagreb University of Applied Sciences
Vrbik 8, 10000 Zagreb, Croatia
dinko.horvat@tvz.hr

Abstract*[1]*. *Generative AI can produce the visible outputs of computing competence — architectural designs, database schemas, technical documentation, working code — before the schema those outputs are assumed to represent has formed. This paper reframes that as pedagogical design rather than detection or integrity. As a theoretical contribution, it proposes metacognitive schema-first scaffolding rather than testing it: schema formation should precede output as evidence of competence; AI should be a temporary, fading scaffold; and metacognition should be taught explicitly through the protocol hold–rethink–continue. Five teaching and assessment practices follow. Empirical validation remains future work, beginning with a Croatian institutional study.*

Keywords. computing education, generative AI, metacognition, schema-first learning, scaffolding, scaffold fading, hold–rethink–continue, competence formation, AI-assisted learning

1 Introduction

Generative AI did not enter education through a planning cycle, a procurement decision, or a curriculum committee. It entered through the everyday infrastructure already used by students and educators: phones, browsers, and natural-language interfaces that were operational before institutional policy could respond. By the time universities began discussing frameworks, students were already using the tools.

This is not simply an institutional failure. Education systems change - curriculum cycles, accreditation processes, and assessment frameworks are slow by design. Their mission is to produce reliable and comparable competence credentials for the large interconnected system, not one particular domain. The real problem is that the surrounding environment has changed faster than those structures can adapt. UNESCO acknowledged in 2023 that publicly available generative AI tools were emerging faster than regulatory frameworks could adjust (UNESCO, 2023). The OECD Digital Education Outlook 2026 confirms that even knowledge about integrating generative AI into teacher-designed learning scenarios remains insufficient (OECD, 2026).

The practical consequence is a pedagogical paradigm shift. The visible output of learning — architectural designs, database schemas, technical documentation and working code — can now be generated by AI before the learner has necessarily performed the cognitive work through which competence forms. The standard assessment principle — this student produced this output, therefore; this student has this competence — has become unreliable. This creates a formation gap: a hidden difference between the output a student can produce with AI support and the schema-based competence that output is assumed to represent.

The problem is not primarily that students may be dishonest, but that AI tools can now produce exactly the kind of output that assessment was designed to certify.

Most students in higher education encounter GenAI as a disruption to their learning practice. A medical student using GenAI as a new tool in a field where core professional acts — diagnosis, clinical judgement, and patient interaction — remain formally human-led and institutionally regulated. The work is changing, and many AI-assisted tools must be learned, but the core professional role has not been redefined. An IT student faces a different situation: the tool that is restructuring their learning environment is also restructuring the professional field they are entering at an unprecedented speed.

[1] This paper is one of a pair developed in parallel for this conference. It addresses the institutional and pedagogical response to formation-stage AI adoption in IT education. A companion paper addresses the risk as experienced by the IT student. Each paper stands independently.

Their target is moving while they are being trained to hit it. This double exposure — GenAI reshaping both how they learn and what they will be expected to do (adapting O'Brien and Leichenko's [2000] double-exposure framework to a single technology) — is becoming the norm in IT education and has few parallels in other disciplines.

Professional pressure is already evident. Stack Overflow's Developer Survey reports widespread adoption of AI tools among developers, particularly for code generation, debugging, documentation, and testing – professional skills current IT students are developing right now (Stack Overflow, 2025). The World Economic Forum's Future of Jobs Report 2025 identifies information and technology services as sectors facing significant digital transformation and increased pressure on routine programming roles (World Economic Forum, 2025). This trend is not just a forecast for IT students; it reflects the current reality of the professional environment they are about to enter. IT students are joining the field during a period of rapid transition.

The lack of transitional memory makes this problem significant. While IT educators have experienced platform changes and developed new schemas gradually over time, IT students lack this experience. They have no practical way to distinguish when AI assistance supports a competency and when it replaces it. For these students, the logical choice is to use the available tools, improve results, and meet requirements. The cognitive cost is not immediate, but delayed, accumulating unnoticed during education and emerging later when independent judgment is required, even if AI remains available. This paper defines this moment as the point of delayed recognition: when students realize that although a tool can help with performance, it cannot replace a schema they never developed.

For computing educators, this clarifies both the urgency and the opportunity. Their role is not to discourage students from using essential tools, but to provide timely instruction that integrates AI use during schema formation. This paper supports educators navigating the gap between institutional delays and student adoption. While it does not present a comprehensive institutional solution, it introduces a metacognitive schema-first scaffolding framework for immediate action. The core principle is hold–rethink–continue, a metacognitive control loop that positions GenAI as a scaffold for learning rather than a shortcut.

2 From Productivity to Scaffolded Learning

AI offers real productivity benefits in education - the presented evidence highlights the associated formation costs.

2.1 When better output does not mean better learning

Bastani et al. (2025) present direct empirical evidence of formation risk. In a preregistered field trial in high school mathematics with nearly 1,000 students, unrestricted ChatGPT-4 access improved AI-assisted practice performance but reduced subsequent independent assessment performance. The study also finds that students were often unaware of this decline because their confidence was shaped by their AI-supported practice. Fan et al. (2025) describe the related habit as metacognitive laziness: when AI handles monitoring, evaluation, and checking, learners may underuse those capacities themselves. Shen and Tamkin (2026) provide convergent evidence in a developer-learning context, where AI-supported task completion did not translate into stronger post-task mastery in conceptual understanding, code reading, and debugging. Bassner et al. (2025) add a programming-specific finding: AI support reduced stress and improved exercise scores without automatically improving code understanding.

These studies highlight a structural risk: AI may enhance the appearance of competence without ensuring the underlying schema is developed. A schema is a mental model representing a student's internal understanding of why something works, not just how to perform it. Its value is in transferability. Once formed, students can explain, adapt, and apply knowledge in new contexts. If AI generates results before this structure is established, students may produce outputs without genuine understanding.

Even when students recognize this risk, the underlying incentives often prevail. Caosun and Aral (2026) model skill dynamics under AI adoption and demonstrate that fully informed students may still rationally choose to use AI heavily. Immediate benefits, such as higher assignment scores and project deliverables, can outweigh the future cost of skill erosion. This is not a result of naivety, but a rational response to conflicting performance and capability signals. When incentives drive behaviour, awareness-based guidance is insufficient.

Caosun and Aral's distinction between “performance-extracting” deployments, in which AI operates independently of student judgment, and “skill-preserving” deployments, in which students must configure, direct, or verify AI contributions, offers a useful framework for the practices proposed in Section 5.

2.2 Metacognition: the mechanism that is missing

The question this raises is not whether AI should be used, but how a student is supposed to notice when it has crossed from scaffold to substitute. Scaffold means temporary learning support that helps the student progress while responsibility gradually returns to the

learner. GenAI can play this role only if the student remains cognitively engaged with the underlying schema rather than accepting the output as a finished product. Metacognition, defined as the capacity to monitor, evaluate, and regulate one's own thinking process (Flavell, 1979), is a well-established concept. It refers to "thinking about thinking," an internal dialogue in which educators and students actively plan, monitor, and assess their learning, regardless of whether they explicitly use the term. In most learning environments, metacognition develops gradually through practice and challenge. For instance, it is practical students' strategy for passing the test, by understanding their own knowledge and abilities and the nature of the test (written, oral, …).

In GenAI environments, metacognition is no longer only a habit that forms in the background — it becomes a prerequisite for recognizing when AI has replaced independent thinking rather than supported it. The metacognitive question — did I learn this, or did the tool do it for me? — does not arise reliably on its own. It must be deliberately trained.

The theoretical framework for this habit is already established. Zimmerman (2002) identified three cyclical phases of effective self-regulation: (1) forethought, which includes goal-setting, planning, and anticipating outcomes; (2) performance, involving monitoring, self-instruction, and attention management; and (3) self-reflection, which covers evaluation, attribution, and adaptive adjustment. Panadero (2017), in his widely cited review of six Self-Regulated Learning models, confirms these phases are consistent across theoretical traditions.

GenAI disrupts each phase: forethought is bypassed when answers are immediately available, performance monitoring is replaced by accepting outputs, and self-reflection is skipped when learners accept AI evaluations instead of their own.

The **hold–rethink–continue** protocol introduced in Section 4.3 puts this cycle into practice. Educators are responsible for designing conditions that restore this process.

2.3 The educator's design obligation

By design, schooling rewards answers much more than the inspection of how answers are produced. Even in math, where the process of arriving at the answers is often rewarding, students won't get credit unless the results are right, too.

GenAI intensifies this pattern: the tool delivers fluent answers immediately, with no hesitation, definitely without the invitation to examine the procedure. The metacognitive habit of checking the procedure, not just the answer, is what would alert a student to the gap between a solution and the reasoning behind it. It does not develop on its own, which is precisely why educators must introduce it explicitly. Bassner et al. (2025) introduce what they call the comfort trap: students do not naturally choose the most effective learning interaction. In their research, students rated the unrestricted AI tool as more helpful in every area, even though it did not improve learning and a guided option was available. Students tend to prefer interactions that require less mental effort, even if those interactions are less beneficial. As a result, educators cannot expect students to develop strong metacognitive habits on their own when using GenAI. These habits need to be built into the way tasks and assessments are designed, since the students most likely to use AI passively are also the least likely to notice the risks involved.

The underlying challenge is not new. Over a decade ago, OECD (2015) tracked educational technology investments in eighteen countries and found no improvement in reading or mathematics; outcomes stagnated or declined in high-investment nations. The issue was not the technology itself, but the order of implementation: tools were introduced before pedagogical practices were established. Kirschner, Sweller, and Clark (2006) explain that minimally guided instruction consistently underperforms compared to direct instruction for novices, as they lack the ability to distinguish relevant from irrelevant information. Uncontrolled AI access in computing education creates a similar situation: the environment is highly responsive, guidance is not sequenced, and learners without prior schemas cannot differentiate productive engagement from plausible-sounding output. Educators must redesign tasks and assessments to address this recurring gap that each new wave of technology reopens.

Educators are in a different position. The OECD Teaching and Learning International Survey (TALIS) 2024 provides representative, comparative information on how lower secondary education teachers use AI. Educators are already using AI primarily for preparation and productivity tasks: on average, 68% of those who use AI, use it to efficiently learn about and summarise topics they teach, and 64% use it to generate lesson plan. By contrast, non-use is largely explained by insufficient skills, lack of confidence, limited training opportunities, or uncertainty about pedagogical value. (OECD, 2026, pp. 18-19).

That productivity is not automatically a pedagogical gain. In a pre-registered field experiment in K–12 schools, Sungu et al. (2026) found that giving teachers a generative-AI support tool — used mainly for the preparation tasks above — reduced students' intrinsic motivation and, among lower-performing teachers, their measured performance: a principal–agent gap in which the teacher's time saving did not translate into a student benefit. That study is in schools rather than higher education, so what carries over is the mechanism, not the effect size — but the substitution logic this paper traces for students applies to educators' own use just as well: AI adopted for output rather than engagement can undercut the outcome it appears to serve.

This highlights the need to position educators as key mediators. Their proximity to classroom practice enables them to redesign tasks, clarify requirements for explanations, and update assessment rubrics. Their direct experience introducing AI in practice provides a personal reference for understanding how to build a schema amid uncertainty. As a result, they are well-placed to introduce metacognitive habits before IT students reach the point of delayed recognition. This paper proposes the hold–rethink–continue cycle as a teachable metacognitive control loop, detailed in Section 4.3. Its aim is to re-engage students with AI through understanding rather than dependency.

3 Methodology: Theoretical Research Design

This paper is theoretical research in computing education. It does not present new empirical data and does not assert that metacognitive schema-first scaffolding has been validated as a comprehensive instructional intervention. The schema-first ordering is not merely hypothetical: in direct response to generative AI, one of the authors redesigned a first-year informatics Information Literacy and Critical Thinking course — a prerequisite aimed at independent reasoning — so that students build their own mental model of a problem before using AI for prompt formulation, and defend it in a live debate rather than a written essay, keeping the assessed output resistant to substitution (Bušelić, 2023). This is a design precedent, not evidence of effect: it shows the ordering is teachable, and it motivates the empirical study proposed in Section 6.

The objective is to develop an original conceptual framework for computing education in the GenAI era by synthesizing established theories of competence formation, scaffolding, tacit knowledge, metacognition, and relational understanding with recent empirical findings on AI-assisted learning. The complementary student-facing perspective on this risk is developed in a companion paper presented in parallel.

The theoretical framework of this paper is built upon six complementary foundations. First, the CC2020 Knowledge–Skills–Dispositions model defines the competencies that computing educational programs should develop and certify (ACM/IEEE-CS/AAAI CC2020 Task Force, 2020). Second, Polanyi's (1966) tacit knowledge theory shows that professional competence involves embodied understanding developed through practice, which cannot be fully conveyed by explicit demonstration. Third, scaffolding theory provides the pedagogical basis for schema-before-output instruction. Vygotsky (1978) establishes the zone of proximal development, in which support must fade as learner capacity grows. Wood et al. (1976) define scaffolding as temporary support for tasks beyond unaided capability, and Reiser (2004) explains how scaffolds structure the task while preserving its thinking challenge. Vendrell and Johnston (2026) extend this to LLM-supported critical thinking in higher education. Fourth, Flavell's (1979) metacognition framework and Schraw and Dennison's (1994) metacognitive awareness model provide the basis for treating metacognition as a teachable control mechanism in GenAI-rich learning environments. Fifth, Skemp's (1976) distinction between instrumental and relational understanding clarifies why correct performance may conceal the absence of transferable understanding. Sixth, established learning science provides the normative substrate for the teaching practices in Section 5: deliberate practice (Ericsson, Krampe & Tesch-Römer, 1993), self-regulated learning (Zimmerman, 2002), retrieval practice (Dunlosky et al., 2013), and feedback (Hattie & Timperley, 2007). Together these identify the conditions competence formation requires — conditions the framework proposes to reinstate.

The paper's original theoretical contribution is the framework of metacognitive schema-first scaffolding (MSFS), defined and developed in Section 4. In brief: schema formation must precede output as evidence of competence; AI must function as a fading scaffold rather than a substitute; and the two are held together by the taught metacognitive loop hold–rethink–continue. The framework draws on metacognitive monitoring and regulation (Flavell, 1979; Schraw & Dennison, 1994; Zimmerman, 2002), scaffolding as temporary support (Wood et al., 1976), and the distinction between surface performance and relational understanding (Skemp, 1976).

The sources for this paper are intentionally divided into two distinct categories, and the selection does not represent a comprehensive literature review. It serves a different purpose.

The first category includes established theoretical foundations: widely cited, empirically supported models whose core claims are uncontested in their fields. Examples include scaffolding and the zone of proximal development (Wood et al., 1976; Vygotsky, 1978), deliberate practice (Ericsson et al., 1993), self-regulated learning (Zimmerman, 2002; Panadero, 2017), retrieval practice (Dunlosky et al., 2013), and feedback (Hattie & Timperley, 2007). These works establish the normative basis for competence formation.

The second category includes recent empirical studies from GenAI-era educational research, such as Bastani et al. (2025), Fan et al. (2025), Shen and Tamkin (2026), Bassner et al. (2025), and Yan et al. (2025). These studies do not serve as proof of the proposed framework. Instead, they provide supporting evidence that the framework is relevant to real, observable situations in current practice. Those two categories are complementary: established learning science defines the requirements for competence formation, while recent GenAI studies demonstrate

that current practice is systematically disrupting these conditions.

As this paper is theoretical, its contribution is supported through analytical rather than experimental evidence. Three forms of confirmation are provided: First, theoretical coherence: the framework is based on the established theoretical foundations outlined above. Second, triangulation conceptual framework: different theoretical traditions identify the same educational requirement, that learners must develop transferable structures of understanding rather than simply produce correct outputs. Third, empirical plausibility: the studies reviewed in Section 2.1 indicate that output performance and independent competence formation may diverge, and that scaffolded or constrained AI assistance may better support learning than unrestricted answer production.

4 The Framework: Metacognitive Schema-First Scaffolding

The main idea is simple: in computing education, a student should build understanding before their work is taken as proof of skill. More precisely, metacognitive schema-first scaffolding (MSFS) — the paper's central contribution — is a conceptual framework: a lens for diagnosing where AI-mediated formation fails and for designing against it, not a validated instructional method or assessment model. It integrates three established principles into one response to AI-mediated formation risk, and each principle names one part of the term:

- Schema-first — schema formation, not output, is the primary instructional goal; a task is valued for the mental model it builds, not the artefact it yields. Once AI can generate the output, that output no longer evidences the understanding behind it, so that understanding must become the explicit target (Section 4.1).
- Scaffolding — AI is positioned as temporary, calibrated support that fades as the schema forms, never as a substitute for the cognitive work through which it forms (Section 4.2).
- Metacognitive — the framework is made teachable through hold–rethink–continue, the loop by which a student monitors their understanding, evaluates what the tool produced, and directs AI toward schema-building rather than answer-production (Section 4.3).

The contribution is the integration, not any single principle: none works alone. Schema-first without scaffolding is an instruction without a mechanism; scaffolding without a schema goal optimises better output rather than understanding; metacognition without either activates anxiety rather than learning. MSFS uses scaffolding to create the conditions under which metacognition can activate, and metacognition to keep the scaffolded work aimed at schema formation rather than production — which is why assessment must then certify the schema, not the output AI can now supply (Section 4.4). Section 5 offers concrete teaching and assessment practices that illustrate how the framework can be applied in a course. They are examples, not validated methods — and not the framework itself.

4.1 Schema before output

The clearest example comes from mathematics. A child learning addition may start with apples, blocks, or fingers, which make the operation visible. However, the educational goal extends beyond the apple calculation. It is the schema of addition: a transferable mental structure that later applies to any countable object, then to symbolic algebra, then to abstract structures the student has never seen. Once the schema is formed, any representation works. Without it, the student can only handle the specific cases they have been shown. Skemp (1976) named this distinction instrumental understanding — knowing the rule and producing the correct result — versus relational understanding — knowing why the rule works and being able to adapt it. The difference is invisible in the output. A student with relational understanding and a student with only instrumental understanding can produce identical answers. The gap appears the moment the problem changes.

GenAI makes this distinction urgent in computing education. A student can ask AI to generate working code, a correct database schema, or a well-structured requirements document. The output is often indistinguishable from what a competent student might produce independently. But if the student accepts the output without forming the underlying schema — without understanding why this structure, why this decomposition, why this trade-off — then the output is not yet evidence of independent competence. The next task, the changed constraint, and the novel problem cannot be addressed independently by accepting the previous output again.

They require the schema. If it was never formed, the tools are present but the understanding is not; this is where the formation gap becomes visible.

The educational question is therefore not whether students can produce required outputs with AI. The question is whether they can explain the reasoning behind what was produced, transfer it to a new case, diagnose its failure, and adapt it when the context changes.

Schemas are not built through successful performance. They are sharpened by error — by the moment a solution breaks, a constraint changes, or an approach that seemed correct reveals its limits. That friction is not a failure of instruction. It is the mechanism that exposes the untrained parts of a schema and forces them to develop. A student who never encounters this friction does not experience smooth learning; they experience a hidden ceiling: it

stops exactly where the difficulty was removed. When AI removes this friction before the schema is complete, it may remove the sharpening mechanism along with the obstacle.

4.2 Scaffolding rather than substitution

AI can support schema formation or bypass it. The difference is not in the tool — it is in how the tool is used. When a student asks AI to explain why a recursive function terminates, compare two sorting algorithms and justify the choice for a specific case, or generate test cases after predicting edge cases themselves — AI is functioning as a scaffold. It extends what the student can do while keeping the student cognitively accountable for the schema. When a student asks AI for a finished solution and submits it without engaging the underlying logic — AI is functioning as a substitute. The schema that the task was designed to form does not form.

There is a simple test for students: after using AI, can you explain what was created, why it was organized that way, and what would change if the situation were different? If you can, the AI helped guide your learning. If you cannot, the tool just replaced your own work. This is not about right or wrong. It is about growth. If a student cannot answer these questions, they have not yet learned what the task was meant to teach.

Reiser (2004) explains that effective scaffolds operate through two mechanisms: structuring, which reduces unnecessary complexity, allowing learners to focus on essential elements, and problematizing, which maintains the visibility and challenge of intellectually demanding components. A scaffold that eliminates all difficulty ceases to function as one and instead becomes a bypass. Artificial intelligence occupies this boundary, and its role depends on both the educator's task design and the student's engagement with the tool.

Wood, Bruner, and Ross's foundational study points out a key requirement that clarifies this boundary. Effective scaffolding needs two models: a theory of the task and a theory of the learner (Wood et al., 1976, pp. 97–98). The first is about knowing what a good solution looks like and how to get there, which general-purpose AI can often simulate. The second is about understanding where the learner is, what they already know, where their understanding is incomplete, and what kind of support will help them grow without skipping over their current abilities. General-purpose prompt-based AI tools do not reliably offer this unless they are part of a carefully designed educational approach.

A general-purpose AI system responds uniformly to prompts and cannot calibrate its support to the learner's developmental state without teacher-designed diagnostic structure. Without a theory of the learner, there is no reliable basis for calibration; without calibration, neither structuring nor problematizing can be targeted precisely to the developmental gap. This distinction also aligns with recent GenAI-specific scaffolding literature. Vendrell and Johnston (2026) argue that LLMs should be integrated into higher education through design principles that preserve critical thinking rather than outsource it to the tool.

AI does not bypass scaffolding by design. However, when it is used primarily as an answer-production system rather than as part of a scaffolded learning design, it may bypass the developmental function that scaffolding is meant to serve. A scaffold is also, by definition, temporary. The goal is not permanent support but gradual withdrawal — fading the scaffold as the learner's own schema grows strong enough to carry the task independently. In computing education, this means that AI support should change form as competence increases: from hints and guided comparisons, to explanation and transfer, to independent performance checks.

The aim is to ensure that answer-substitution fades as schema ownership grows.

The fading requirement comes from Vygotsky's (1978) idea of the zone of proximal development. Support should decrease as the learner becomes more capable; otherwise, the learner never takes over control. AI cannot meet this need because it always responds to every question with the same level of detail, no matter what the learner already knows. Fading must be built into the task itself by deciding when AI can be used, when it should be limited, and when learners must work independently to show they can complete the task on their own.

4.3 Hold–rethink–continue as a student protocol

Hold–rethink–continue was introduced earlier as a metacognitive control loop. In this section, it is presented as a student-facing protocol that can be explicitly taught and deliberately practiced.

Hold is a short pause before accepting AI output — enough to ask: do I really understand what this is, and why should I accept it?

Rethink is the reconstruction of the schema behind the output. Ask yourself: Why does this solution work? What assumptions are behind it? How would things change if the input was different, the constraints were stricter, or the scale was bigger? What could go wrong, and how would you notice? This is not just a checklist. It is about making a habit of stepping back from the result and looking at your own understanding. This is metacognition in its simplest, most practical form.

Continue refers to re-engaging with AI using a more specific question. Rather than requesting an answer, the student might ask for a hint, a contrasting case, or an explanation of why an approach fails under a particular constraint. After reflecting, the student returns to the tool with new questions and a better ability to assess the response.

Consider as an illustration, presented in Figure 1., a student asked to design a schema recording which students take which courses and the grades they earn.

A first prompt to a generative AI tool — "*give me the SQL tables for this*" — returns three tables that create without error, including an enrolment table the student did not anticipate.

```
CREATE TABLE student (
    student_id    INT PRIMARY KEY,
    full_name     VARCHAR(100) NOT NULL,
    enrolled_year INT
);
CREATE TABLE course (
    course_id  VARCHAR(10) PRIMARY KEY,
    title      VARCHAR(100) NOT NULL,
    credits    INT NOT NULL
);
CREATE TABLE enrollment (                          -- the junction table
    student_id INT         NOT NULL REFERENCES student(student_id),
    course_id  VARCHAR(10) NOT NULL REFERENCES course(course_id),
    semester   VARCHAR(10) NOT NULL,
    grade      CHAR(1),
    PRIMARY KEY (student_id, course_id, semester)
);
```

Figure 1. Schema returned by a generative AI tool for the enrollment assignment: three tables, including the enrollment junction table

Hold. The student pauses before accepting. The code runs; that is not the question. The question is: I described students and courses — why are there three tables, and why is the grade in the third one?

Rethink. The student reconstructs the schema behind the output. Why can a grade not sit in student (a student earns many) or in course (a course awards many)? What does the key (`student_id, course_id, semester`) actually enforce — and what happens to a retaken course if semester is removed from it? What would break if all three tables were merged into one? Working these through surfaces the many-to-many relationship, the grade as a property of the pairing rather than either entity, and the update anomalies a single flat table would create.

Continue. The student returns to the tool with a sharper question — not "is this right?" but "show me a concrete case where putting the grade in the student table fails." The tool demonstrates the anomaly; the student now evaluates the answer from a formed position rather than accepting it from an empty one.

The protocol encourages students to use AI in ways that keep their own thinking and understanding active. Educators can help by assigning reflection annotations, asking for oral explanations, or including structured phases in tasks. The protocol is not idiosyncratic to this framework. Reviewing thirty-nine studies of generative AI and critical thinking, Nasr et al. (2026) independently derive a structured human-in-the-loop model in which the learner plans, assesses, explores, evaluates, refines and applies — arriving, from a different literature and without reference to this work, at the same conclusion: unstructured use degrades the capability, and structure is what preserves it. Hold–rethink–continue operates one layer above that sequence, as the decision to engage at all; hold fires before planning begins.

The goal for students is to internalise this metacognitive habit before the delayed recognition point arrives.

4.4 Assessment as evidence of schema formation

If we agree that the mere output is no longer sufficient evidence of competence, assessment must make the process of building it (schema formation) visible. The difficulty this creates is not new, and it did not begin with generative AI. In 1990, the APA Delphi panel of forty-six critical-thinking experts recorded that the readily assessable skill dimension should not be allowed to distort the conceptualization of what matters, and that self-regulation — the capacity to monitor and correct one's own reasoning — is "extremely difficult to assess using standard paper and pencil instruments" (Facione, 1990). The gap between what is assessable and what is important is therefore a standing condition of education, identified by the field's own consensus body more than three decades ago. Generative AI did not create it. It industrialised it: what was a known limitation of assessment has become the ordinary case. This does not mean abandoning submitted outputs. Instead, we should add explanations, examples of transfer, error analysis, and independent performance checks.

This also completes a distinction the paper's own foundation draws but has not yet used. CC2020 separates Knowledge and Skills from Dispositions, and is explicit that the third is not a softer version of the first two: a graduate who holds knowledge and skills without the dispositions has codified competence without professional judgment (ACM/IEEE-CS/AAAI CC2020 Task Force, 2020). Dispositions form through repeated practice and **cannot be demonstrated in an output** — no artefact shows that its author would have checked had no one been watching. Where AI substitutes for the cognitive work, it can supply the Knowledge and reproduce the Skills while leaving the Dispositions unformed, and no inspection of the product will reveal the difference.

This heuristic shortcut shows how schema-first scaffolding can be used in assessment design. The framework is based on ideas from the literature; it is not an explicit taxonomy from any single source.

Table 1. Assessment targets ordered by resistance to AI substitution

Assessment target	What it asks the student to demonstrate	AI risk if assessed only as product
Local production	Produce syntactically valid code, diagram, or document	**High** — AI generates reliably

Comprehension	Explain what an output does and trace its behavior	**Moderate** — oral or constrained explanation reveals gaps
Design rationale	Justify structure, decomposition, and trade-offs	**Moderate** — contextual justification requires student ownership
Error diagnosis	Find, explain, and correct failures in a flawed output	**Lower** — requires schema activation
Transfer	Apply the same schema to a new problem or changed constraint	**Lower** — exposes whether schema is independent of the original output
Metacognitive regulation	Describe what was attempted, where AI was used, and how the result was checked	**Lower** — when students must explain their process, AI use, and verification steps

Note. The framework described and compared through Table 1. is the authors' synthesis.

Risk levels serve as heuristics to show how easily students might meet assessment targets using AI, without demonstrating independent schema formation. **High risk** means AI can reliably generate responses, allowing students to meet the target without engaging their own understanding. **Moderate risk** indicates AI can assist, but contextual specificity limits direct substitution. **Lower risk** requires schema activation, so AI output cannot replace context-specific performance.

This pattern keeps showing up. When assessments focus on the lower rows, they are much harder for AI to replace. If a student cannot explain the reasoning behind their output structure, it shows a gap in their understanding that is not evident from the output alone.

5 Teaching and Assessment Practices

The proposed framework offers teaching practices that educators can use right away in their current courses, without waiting for a full curriculum redesign. These are not just extra improvements added to existing assessments. Instead, they challenge three main assumptions: that output alone shows competence, that difficulty should always be removed, and that AI use is simply a matter of allowing or banning it. Each practice directly addresses one of these ideas. These approaches also align with new GenAI design research, which views LLMs as tools to support critical thinking rather than as replacements for student judgment (Vendrell & Johnston, 2026).

Each of the proposed practices restores a condition that learning science, before the rise of GenAI, regarded as essential for building competence. Ericsson, Krampe and Tesch-Römer (1993) showed that expertise comes from deliberate practice, which means working on tasks that challenge current abilities, getting feedback on mistakes, and staying engaged with difficult work. When students have unlimited access to AI, these conditions often disappear. When educators see these practices as a way to restore important learning conditions, they shift from policing AI use to helping students build real-life skills.

5.1 Scaffold AI access through task phases

One practical improvement is to decide not just if AI is available for a task, but when it is available. Breaking a task into phases like independent attempt, AI-supported exploration, explanation, and transfer helps students build their own understanding while still letting AI help later on. In the first phase, students try the task themselves, sketch a design, or predict what a correct answer would look like. This gets them thinking about the problem and shows what they already know before AI steps in. In the next phase, students use AI for hints, comparisons, or explanations instead of asking for full solutions. Asking questions like "why does this work?" helps them learn more than just asking AI to do the work. Finally, students explain what was produced and apply the same logic to a new situation. These steps make it clear whether AI was used as a learning tool or merely as a shortcut for output.

The pre-GenAI phase has a specific name. Dunlosky et al. (2013) identified retrieval practice as one of the most robustly supported learning techniques. Its effectiveness comes not from reviewing correct information, but from attempting recall before the answer is available, which serves as the consolidation mechanism. AI inverts this sequence structurally: the answer appears before retrieval has been attempted. Requiring an independent attempt, design sketch, or prediction before AI access operationalises retrieval practice in a computing context.

5.2 Use explanation gates

An explanation gate is a point in the task where progress depends on the student's ability to explain the schema behind the output — not only what it does, but why it is structured as it is. It may be oral, written, or interactive. Its function is not to detect AI use. It is to make schema formation a precondition for completion. Ask yourself why this structure was chosen instead of another. What would happen if you removed one part? Where could this approach fail, and how might you notice problems early? You can’t answer these

questions just by repeating what AI says. You need to think through the ideas on your own.

5.3 Design error-analysis tasks

Error-analysis tasks are especially useful for computing educators in the GenAI era. In these tasks, students receive a flawed output, such as code with a logic error, a database schema with an integrity problem, or a system design with a structural bottleneck. They are then asked to find the error, explain why it is a problem, and suggest a fix. This approach brings back important elements that AI-driven coursework can sometimes remove too soon: real uncertainty, the challenge of diagnosing issues, and the moment when students realize the difference between what they thought they knew and what they actually understand. In an error-analysis task, the schema cannot go untested. It must be put to use, confront the flaw, and either stand up to the test or show where it falls short.

5.4 Use deliberate non-use as scaffold fading

Scaffold fading — the deliberate withdrawal of support to verify that the schema can carry the task independently — is the natural counterpart to scaffolded AI use.

Periodic tasks that require students to work without AI are not meant to punish or restrict AI use. Instead, they help confirm whether the scaffold's support has been effective. For students' integrity, it is important not to position these tasks as tests of their honesty. Rather, they serve as checkpoints where support is removed, not because the tool is flawed, but so the learner can see if they can complete the task on their own. Being able to perform independently under these conditions shows that a student has truly learned the material, rather than just producing the right answers. This is the best way for both students and educators to see if real learning has taken place. Yan et al. (2025) studied this directly. In their three-group randomized controlled trial, both a proactive and a passive GenAI agent used the same AI technology; the only difference was how students interacted with them. While support was available, all three groups performed similarly. When support was removed, only the proactive group kept improving, while the passive group returned to the same level as the group without AI. This shows that scaffold-fading is a real and measurable effect: giving a task with AI support, followed by one that requires explanation or transfer without AI, helps reveal whether students have truly learned or just relied on support.

When we look at this through the idea of deliberate practice, the discomfort of working without AI—the effort, uncertainty, and mistakes—is not just a side effect of learning. Instead, it is the main way learning happens. The conditions that Ericsson et al. (1993) say are needed for developing competence are exactly what working without AI brings back.

5.5 Track schema formation longitudinally

A single assessment provides only a necessary snapshot. A formation trajectory requires data across time. The signal that matters is not whether students use AI, but whether their independent performance converges with their assisted performance as instruction progresses. A student whose assisted and independent performance diverge across a semester is on a formation-risk trajectory. A student whose independent performance rises toward assisted performance provides evidence that the intended schema is forming. Low-stakes diagnostic assessments — not high-stakes credentialing instruments, but formation-tracking checkpoints — give both students and educators the information needed to identify the formation gap before the delayed recognition point becomes the first evidence that independent competence has not formed.

Longitudinal tracking requires feedback that informs the learner's next step, not only feedback that evaluates the current output. Hattie and Timperley (2007) distinguish four feedback levels: task (is the product correct?), process (what strategy produced it?), self-regulation (can the student monitor and adjust?), and self (what does this mean for the learner's development?). Task-level feedback — the default in product-based assessment — is the least effective for long-term competence formation; the trajectory tracking proposed here requires process and self-regulation level feedback. The self-regulation level has a practitioner precedent: in the course introduced in Section 3, students assessed their own competence before and after as a metacognitive exercise — building the self-monitoring habit, not measuring whether competence had formed (Bušelić & Rajković, 2024). The convergence signal itself still requires the independent, unassisted assessment described here. The structural implication from Caosun and Aral (2026) also applies here: under high-AI-use conditions, assisted performance is no longer a reliable baseline. Periodic unassisted assessment is therefore not punitive — it is the only measurement context in which the convergence or divergence signal becomes visible.

5.6 A worked example: one assignment across the practices

The following worked example is illustrative only. It makes the framework's application concrete and constitutes no empirical claim; consistent with Section 3, this paper remains theoretical.

The practices above are not five separate assignments but five ways of designing one. The enrolment schema of Section 4.3 illustrates the whole sequence presented in Figure 2.

Independent attempt (5.1). Before any AI, the student sketches an entity–relationship diagram by hand and states the relationship in words: a student takes many courses; a course is taken by many students. This is retrieval practice — it reveals whether the student recognises a many-to-many relationship at all. A student who stops at two boxes and a connecting line has exposed the exact gap that AI would otherwise fill invisibly.

AI-supported exploration (5.1). The student now uses AI not to produce the schema but to compare designs: contrast recording the grade on the enrollment row with holding it in a separate assessment table — when is each better? The student remains the one choosing.

Explanation gate (5.2). Progress depends on the student explaining why the relationship needs a third table, why the grade belongs to the pairing, and what the composite key enforces. Pasted output cannot answer this.

Error analysis (5.3). The student receives a flawed variant that still executes, and must find, explain, and correct it:

```
CREATE TABLE enrollment (
    student_id      INT         NOT NULL REFERENCES student(student_id),
    course_id       VARCHAR(10) NOT NULL REFERENCES course(course_id),
    semester        VARCHAR(10) NOT NULL,
    grade           CHAR(1),
    instructor_name VARCHAR(100),            -- (1)
    PRIMARY KEY (student_id, course_id)      -- (2)
);
```

Figure 2. A flawed variant of the enrollment table, used for the error-analysis task

Both faults are silent at creation time. (1) `instructor_name` depends on the `course`, not on the `student-course-semester` pairing — a transitive dependency that will duplicate and eventually contradict itself. (2) The key omits semester, so a retaken course collides with its own earlier attempt: the schema quietly encodes the rule *no student may ever retake a course*. Neither fault is visible from the fact that the code runs; only an activated schema catches them.

Transfer. The constraint changes: a student may retake a course, and both attempts must be kept. The correct schema survives; the flawed one fails. Whether the student can adapt it shows whether the schema is theirs or the tool's.

The same assignment scored against Table 1.:

Table 2. The enrollment assignment mapped onto the Table 1. assessment targets

Table 1. target	What the student is asked to do here	AI substitution risk
Local production	Write the CREATE TABLE statements	**High** — AI generates reliably
Comprehension	Explain what the enrollment table does	Moderate
Design rationale	Justify the composite key and the third table	Moderate
Error diagnosis	Find and fix the two faults above	Lower
Transfer	Adapt the schema to allow retakes	Lower
Metacognitive regulation	State where AI was used and how it was checked	Lower

As figure 2. indicates, the assignment does not change across these rows — only what is asked of it does, and that is what separates an assessment AI can satisfy from one it cannot.

6 Research Agenda

The framework proposed in this paper is conceptual: its value lies in providing a disciplined basis for designing and testing an intervention. The central empirical question is direct: does metacognitive schema-first scaffolding produce better independent competence than unrestricted AI use, AI prohibition, or product-only assessment? No existing study answers this for computing education as an integrated framework. Testing it would proceed in two phases.

The first would be a baseline. Before the framework can be introduced or measured, it would help to know where practice currently stands: how computing educators who already work with generative AI design their tasks and assessment, and how far they recognise the formation gap described here. A survey of such educators — for instance at the Faculty of Organization and Informatics, University of Zagreb (FOI), and Zagreb University of Applied Sciences (TVZ), subject to institutional support — would provide early data from Croatian institutions and add a Central European perspective to a literature still largely Anglo-American. Its aim would be modest: not to test the framework, but to map current practice and to prompt educators to examine their own assessment habits. It is intended as an impetus, not a precondition.

The second phase would test the framework itself. Several questions follow from the analysis:

- Does scaffolded AI access — phased tasks, explanation gates, and error-analysis requirements

— improve independent transfer performance compared with unrestricted AI access?

- Which scaffold types are most effective for schema formation in computing education: hints, worked contrasts, explanation requirements, delayed access, or error diagnosis?
- Does explicit metacognitive training — teaching hold–rethink–continue as a student protocol — reduce passive answer acceptance and improve independent performance?
- Does the divergence between AI-assisted and independent performance narrow across a semester when schema-first scaffolding is applied?
- Can educators reliably assess schema formation through explanation, transfer, and error-analysis tasks, and does this assessment predict independent professional performance better than product-only assessment?
- Can hold–rethink–continue be taught to students who do not yet possess the metacognitive capacity it presupposes — and if not, what educator-supplied structure substitutes for that capacity until it forms?

The last question is the framework's own hardest, and it is not rhetorical. Nasr et al. (2026) find the benefit of generative AI for critical thinking conditional on the metacognitive skills needed to interrogate it — precisely the skills the formation-stage student has not yet built. Every protocol of this kind, including the one proposed here, therefore presupposes some measure of the capability it is meant to form. Whether that circle can be broken by educator-supplied structure, or whether it marks a genuine limit, is an empirical question this framework cannot settle by argument.

Such a study would compare scaffolded and unscaffolded computing cohorts, using independent — not AI-assisted — performance as the outcome measure, and would track whether the assisted-to-independent gap narrows over a semester.

Whether and when this becomes possible would depend on institutional support. Our own institutions would be one place to begin, but the framework and the two-phase path outlined here are offered to any computing programme to take up — and to adapt to its own context, constraints, and questions. It is put forward now precisely so that such work, in this form or a modified one, can be designed against it.

7 Conclusion

Generative AI does not eliminate the necessity of computing education; rather, it clarifies its objectives. If AI can generate outputs that traditional assessments were intended to validate, educational frameworks must specify the knowledge, understanding, self-regulation, and transfer skills students require before such outputs can serve as evidence of competence. This need for clarification should not be viewed as a crisis but as an opportunity to establish a more robust model than the previous output-focused approach.

This paper makes three central arguments.

First, information technology students encounter a formation risk that is structurally distinct from the broader AI-in-education challenge. It is characterised by double exposure, the absence of transition memory, and a professional field already being transformed by the very adoption logic these students will soon face.

Second, the mechanism underlying both the emergence and mitigation of this risk is metacognition, defined as the ability to reflect on the learning process and assess whether genuine understanding has been achieved.

Third, metacognitive schema-first scaffolding offers educators a practical framework for immediate implementation.

This includes sequencing AI integration through schema development, maintaining productive friction, instructing students in the **hold–rethink–continue** protocol, and confirming schema independence through scaffold fading.

Educators need not wait for institutions to change or for new policies to be implemented. The integrated framework awaits empirical validation, but the principles it draws together — schema before output, scaffolding that fades, and explicit metacognition — are well established in learning science and can be applied now, with the tools educators already have, treating the development of students' understanding as no less important than the results they produce.

Computing educators occupy a central role in facilitating **schema formation**. The hold–rethink–continue principle is **concrete enough** to introduce in the first week of a course and **consequential enough to shape** students' future professional practices.

- **Hold** before accepting the assumption that output proves competence.
- **Rethink** what each assignment is designed to form.
- **Continue** — forward, with AI as a scaffold for learning rather than a shortcut around it.

References

ACM/IEEE-CS/AAAI CC2020 Task Force. (2020). Computing curricula 2020: Paradigms for global computing education. *ACM*. https://doi.org/10.1145/3467967

Bassner, P., Lenk-Ostendorf, B., Beinstingel, R., Wasner, T., & Krusche, S. (2025). Less stress, better scores, same learning: The dissociation of performance and learning in AI-supported programming education. *Computers and Education: Artificial Intelligence*, 100537.

Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high

school mathematics. *Proceedings of the National Academy of Sciences*, 122(26), e2422633122.

Bušelić, V. (2023). Teaching information literacy and critical thinking skills in ChatGPT time. 2023 *International Conference on Computing, Networking, Telecommunications & Engineering Sciences Applications (CoNTESA),* Zagreb, Croatia.

Bušelić, V., & Rajković, I. (2024). Teaching generic skills with/or against ChatGPT. *2024 47th MIPRO ICT and Electronics Convention (MIPRO),* Opatija, Croatia.

Caosun, M., & Aral, S. (2026). The augmentation trap: AI productivity and the cost of cognitive offloading [arXiv preprint]. https://arxiv.org/abs/2604.03501

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques. *Psychological Science in the Public Interest,* 14 (1), 4–58.

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review,* 100 (3), 363–406.

Facione, P. A. (1990). Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction (The Delphi Report). *American Philosophical Association*.

Fan, Y., Tang, L., Le, H., Shen, K., Tan, S., Zhao, Y., Shen, Y., Li, X., & Gašević, D. (2025). Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. *British Journal of Educational Technology,* 56(2), 489–530.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist,* 34(10), 906–911.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research,* 77 (1), 81–112.

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work. *Educational Psychologist,* 41 (2), 75–86.

Nasr, N. R., Tu, C.-H., Sujo-Montes, L., Haniya, S., Yen, C.-J., & Hossain, S. (2026). Generative artificial intelligence impact on critical thinking in higher education: A comprehensive bibliometric analysis and systematic review. *Journal of Computers in Education*.

OECD. (2015). Students, computers and learning: Making the connection. PISA, *OECD Publishing*.

OECD. (2026). OECD Digital Education Outlook 2026: Exploring effective uses of generative AI in education. *OECD Publishing*. https://doi.org/10.1787/062a7394-en

O'Brien, K. L., & Leichenko, R. M. (2000). Double exposure: assessing the impacts of climate change within the context of economic globalization. *Global Environmental Change*, 10(3), 221–232.

Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology*, 8, 422.

Polanyi, M. (1966). The tacit dimension. *Doubleday*.

Reiser, B. J. (2004). Scaffolding complex learning: The mechanisms of structuring and problematizing student work. *The Journal of the Learning Sciences*, 13(3), 273–304.

Schraw, G., & Dennison, R. S. (1994). Assessing metacognitive awareness. *Contemporary Educational Psychology,* 19(4), 460–475.

Shen, J. H., & Tamkin, A. (2026). How AI impacts skill formation *[arXiv preprint]*. https://arxiv.org/abs/2601.20245

Skemp, R. R. (1976). Relational understanding and instrumental understanding. *Mathematics Teaching,* 77, 20–26.

Stack Overflow. (2025). 2025 Stack Overflow Developer Survey. https://survey.stackoverflow.co/2025/

Sungu, A., Lira, B., & Duckworth, A. L. (2026). Generative AI can harm teaching *[SSRN working paper]*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=7007339

UNESCO. (2023). Guidance for generative AI in education and research. https://www.unesco.org/en/articles/guidance-generative-ai-education-and-research

Vendrell, M., & Johnston, S.-K. (2026). Scaffolding critical thinking with generative AI: Design principles for integrating large language models in higher education. *Computers and Education: Artificial Intelligence,* 10, 100572.

Vygotsky, L. S. (1978). Mind in society: The development of higher psychological processes. *Harvard University Press.*

Yan, L., Martinez-Maldonado, R., Jin, Y., Echeverria, V., Milesi, M., Fan, J., Zhao, L., Alfredo, R., Li, X., & Gašević, D. (2025). The effects of generative AI agents and scaffolding on enhancing students' comprehension of visual learning analytics. *Computers & Education,* 234, 105322.

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory into Practice*, 41 (2), 64–70.

Skills Never Formed: AI Adoption and Career Formation Risk in IT Higher Education

Dinko Horvat
Zagreb University of Applied Sciences
Vrbik 8, 10000 Zagreb, Croatia
dinko.horvat@tvz.hr

Vjeran Bušelić
University of Zagreb
Faculty of Organization and Informatics
Pavlinska 2, 42000 Varaždin, Croatia
vbuselic@foi.hr

***Abstract*[1]. *Generative AI arrived beneath educational institutions rather than through them, ambient in everyday tools. Computing students face double exposure: the same technology restructures both their learning environment and the profession they are entering. This conceptual synthesis integrates cognitive-science learning theory with three convergent randomized trials to trace one mechanism — how AI use at the formation stage, without metacognitive discipline, produces never-skilling: the failure to acquire professional capabilities because AI performs the very work through which they would form. From it the paper derives hold–rethink–continue, a teachable metacognitive posture. A companion paper develops the educator response.*

Keywords. generative AI, IT education, never-skilling, double exposure, hold–rethink–continue, schema formation, self-regulated learning, verification economy

1 Introduction

IT students enter higher education at the start of a dynamic and demanding career — in software development, data analysis, infrastructure, cybersecurity, and related fields. AI is already part of that world. It just appeared as an ambient layer already in the editor and the browser, before anyone decided how it should be used in business, society, and education. And it is changing two things at once: how they learn, and how it is rapidly changing the industry they are educated to join — a double exposure (adapting O'Brien and Leichenko's [2000] framework, originally two distinct global forces, to a single technology on two fronts) that no other field of study faces in the same form.

This paper is not against AI. Students already use it, and they need to — working with AI is now part of the profession, and learning to do it well is part of their education. The real question is how they adapt: whether AI becomes a tool they learn with, or one that does the learning for them.

The risk is specific. During the formation stage — the early period when professional capability is built — AI can complete the very tasks through which that capability would form, so a student can produce credible work without developing the independent skill behind it. The credential — the diploma — forms; the capability it implies stays in question. This paper calls that outcome never-skilling — a term now emerging across professional-education research (e.g., Ke et al., 2026, in medical education) for the failure Macnamara et al. (2024) describe as AI-induced skill development hindrance. It is not a sharper form of deskilling but a totally different risk.

How never-skilling differs from deskilling — in when it strikes, its mechanism, whether it can be detected, and what remedy it admits — is set out in Section 2, where the paper's scope is defined (Table 1.).

The concern is neither speculative nor settled. A cluster of 2025–2026 studies has begun to measure how AI use during learning affects the competence that forms beneath the visible output (Shen & Tamkin, 2026; Bastani et al., 2025; Fan et al., 2025). What that literature has barely examined is the population most exposed to the risk — computing students at the formation stage, for whom AI restructures the learning environment and the profession at once. This paper addresses that gap from the student's standpoint; Section 2 sets out its approach and Section 4 the evidence it rests on.

The remainder of this section sets out why this risk is acute now: the gap between computing curricula and industry (Section 1.1), the labour-market pressure that transmits it (Section 1.2), and why generative AI differs from earlier tools (Section 1.3). Section 2 sets out the paper's approach and scope; Section 3 traces the cognitive risk cascade that produces never-skilling; Section 4 presents the empirical evidence; Section 5 proposes student-facing practices; Section 6 discusses

[1] This paper is one of a pair developed in parallel for this conference. It addresses formation-stage AI adoption from the IT student's perspective. A companion paper Schema Before Output: Metacognitive Schema-First Scaffolding in Computing Higher Education in the GenAI Era addresses the institutional and pedagogical response. Each paper stands independently.

the findings and their limitations; and Section 7 concludes.

1.1 The Curricula–Industry Gap

For years, computing degrees were built for a job market that is now disappearing — one where new graduates spent their first two to four years writing functions, debugging failures, and building components from specifications (CC2020 Task Force, 2020). It was during their junior phase, not in the classroom, that developers built their core skills. These are the patterns and habits experienced developers use almost automatically. Now, generative AI speeds up that stage, since boilerplate and scaffolding can be created in seconds. As a result, the process that once helped new developers build good judgment is getting shorter, if it survives at all. The World Economic Forum (2025) records the pace: information and technology services report the highest expected AI transformation of any sector (99%), alongside declining demand for routine programming. Curricula cannot keep up: revision cycles run three to eight years against an adoption curve that accelerated between 2022 and 2025. The result is a structural mismatch. Graduates are expected to step straight into orchestration work — problem framing, system design, quality assurance of AI output — that presupposes the very judgment the compressed junior years used to build (Gartner, 2025).

1.2 Labour-Market Pressure

That pressure is sharpest in software development, where adoption has moved fastest: 84% of surveyed developers use or plan to use AI tools, with 51% of professional developers now using them daily (Stack Overflow, 2025), and Gartner (2025) projects 90% enterprise adoption by 2028. It reaches students through job postings that demand AI proficiency and cohort dynamics that penalise whoever abstains — and it arrives at the formation stage, before students have the schemas that would make AI assistance adaptive rather than substitutive (Prather et al., 2023). The result is an adoption dilemma: adopt early and gain a hiring edge at the cost of formation, or resist and protect formation at a real competitive cost. Neither response is irrational, which is precisely why the tension is structural — it calls for an institutional response, not individual resolve.

1.3 Why AI Is Different from Prior Productivity Tools

Offloading cognitive work to tools is neither new nor harmful in itself — calculators offload arithmetic, compilers offload syntax. What matters is **what** gets offloaded, and **when**. Generative AI offloads a different class of task: problem decomposition, pattern recognition, error diagnosis, solution generation. Every earlier tool changed what a developer worked **on**; this one offers to absorb what a developer thinks **with**. In professional work with established schemas that is a legitimate efficiency gain; in learning, those same tasks are the mechanism through which the schema forms — so offloading them removes the very difficulty that builds competence. The logic is documented most precisely in software development; we argue it extends to any IT specialisation where AI can perform the domain task before the practitioner can do it independently — though that extension is a reasoned claim, not yet measured beyond programming.

A second difference compounds the first: GenAI has no access barrier. Every earlier tool required the user to adapt — to learn a syntax or interface; GenAI runs on natural language, the one every student already has. Because one can simply talk to it in natural language, Harari (2023) warns that this is a technology that has „hacked the operating system of human civilisation". Needing no learning curve, it can be deployed from day one — and because it responds like a knowledgeable peer, the interaction **feels like reasoning** even though the mechanism beneath it has no understanding, judgment, or accountability. Earlier tools were clearly just tools, but GenAI feels more like a collaborator. This makes its mental impact hard to notice at first, until you actually put it to the test.

Wang et al. (2025) show why a learner's schema completeness — how complete their mental template for a task is — affects the outcome. Practitioners with complete schemas use AI output as raw material, copying selectively and rewriting; those with incomplete schemas accept and reformat it without independent resolution. The same tool produces opposite developmental outcomes — and at the formation stage, the schema is incomplete by definition. Fatigue worsens it, pushing tired students toward delegation at exactly the late-night, end-of-term moments of highest formation value (Kahneman, 2011). Together — schema-building tasks offloaded, no barrier to use, and fatigue-driven delegation — these are what make the cognitive risk cascade possible, which Section 3 traces in detail.

2 Approach and Theoretical Background

This paper is a conceptual synthesis, not a new empirical study: it presents no original data. Its objective is twofold. The first is to integrate established cognitive-science accounts of skill formation with recent empirical findings on AI-assisted learning, tracing a single mechanism: how AI use at the formation stage, without metacognitive discipline, produces never-skilling. The second is to derive from that mechanism a response — hold–rethink–continue, a teachable metacognitive posture operationalised through four student-facing practices.

The synthesis draws on three bodies of evidence. The first is cognitive-science learning theory — cognitive load (Sweller, 1988), deliberate practice (Ericsson et al., 1993), retrieval practice (Dunlosky et al., 2013), self-regulated learning (Zimmerman, 2002; Panadero, 2017), and scaffolding (Vygotsky, 1978; Wood et al., 1976) — which establishes the conditions under which technical competence forms. The second is convergent empirical evidence: three independent randomised controlled trials across distinct domains — coding (Shen & Tamkin, 2026), mathematics (Bastani et al., 2025), and writing (Fan et al., 2025) — that isolate the same mechanism of cognitive bypass. The third is labour-market data on AI adoption in computing (World Economic Forum, 2025; Gartner, 2025; Stack Overflow, 2025). The argument's strength rests on triangulation: established learning science defines what competence formation requires, and the convergent trials demonstrate that current AI use systematically removes those conditions.

These sources were chosen purposively; this is not a systematic review, and the choice is deliberate. The three trials are not the only studies of AI in higher education. Each was chosen because it isolates the same mechanism — a measured gap between assisted output and independent competence — under a randomised design, and because together they span three different domains: coding, mathematics, and writing. That spread is the point. A single mechanism that appears in settings this far apart is hard to dismiss as a quirk of one task. Studies that only describe how much AI is adopted, or how students feel about it, were not drawn on for the mechanism: they cannot speak to whether competence forms.

A note on the status of this evidence, since the field is young and moving quickly. Where a study has reached its journal version, we cite that: the mathematics trial is published in the Proceedings of the National Academy of Sciences (Bastani et al., 2025) and the writing trial in the British Journal of Educational Technology (Fan et al., 2025). The coding trial (Shen & Tamkin, 2026) is still a preprint; we draw on it because it is the closest evidence to IT practice and because the case rests on the agreement of all three trials, not on any single one. The divergence result (Caosun & Aral, 2026) is a formal model, also a preprint, and we treat its prediction as a modelled projection rather than a measured outcome.

The scope is bounded to the IT-education formation stage, where professional schemas and judgment are established; it does not address practising professionals, for whom the dynamics differ; their concern is deskilling (Table 1.).

Table 1. Deskilling vs. never-skilling

Dimension	Deskilling	Never-skilling
Timing	After capability has formed	Before capability has formed
Who	A practising professional	A student at the formation stage
Mechanism	A formed capability erodes through disuse or automation	A capability never forms because AI performs the formative task
Detection	Loss shows against prior performance	No baseline to compare against — invisible until tested independently
Reversibility	Recoverable — retraining restores a known state	Built from scratch; no prior competence to restore
Credential	Reduced mastery behind a still-valid record	A credential without the capability it implies

The two differ in what is at stake, and in the remedy. Deskilling can end a career and is answered by re-skilling. Never-skilling can prevent one from beginning a career; because there are no pre-educational IT skills, it is answered not by recovery but by formation — the hold–rethink–continue posture and the four student practices developed in Section 5.

The paper takes the perspective of the IT student — the individual bearing the career-formation risk — and the student-facing practices that mitigate it. The complementary institutional response — assessment redesign, curriculum, and scaffolding configuration — is developed in a companion paper presented in parallel. Each paper stands independently; together they address the same phenomenon from the learner's and the educator's standpoints.

Because this risk could affect so many students, we think it is better to act on the evidence we have than to wait for proof of every detail. The evidence is indirect and early; we set out its specific limitations, and the research agenda they imply, in the Discussion (Section 6).

3 The Cognitive Risk Cascade

Never-skilling is not a single failure but the possible endpoint of a cognitive risk cascade. Each stage may appear ordinary in isolation — a reasonable shortcut, a missed check, a fluent answer, a passing grade — but together they can carry a student from assisted performance to unformed independent capability without any clear alarm point. Five stages unfold within the individual learner; a sixth then shifts the risk to cohort level.

These six stages build on one another; they are not separate signs to tick off. Each stage is what makes the next one possible: cognitive offloading is what lets

self-monitoring lapse; the monitoring lapse is what lets the illusion of capability form; and that illusion is what lets never-skilling pass unnoticed. Remove an earlier stage and the later ones never arrive — which is why the sequence is a cascade, not a set of independent snapshots.

Stage 1: **Cognitive offloading**

The cascade begins with cognitive offloading: the transfer of cognitive work to an external aid in order to reduce immediate mental demand (Risko & Gilbert, 2016). At the formation stage, generative AI can offload problem decomposition, pattern recognition, error diagnosis, and solution generation before the student has formed a hypothesis or attempted a solution path. Because schemas form through the effortful processing of working a problem through (Sweller, 1988), what is offloaded is precisely what would have built the competence. The tendency runs highest in exactly this age group (Gerlich, 2025), and the choice is not naïve but structurally rational: near-term gains can outrun a distant, discounted skill cost (Caosun & Aral, 2026).

Stage 2: **Self-monitoring lapse**

Offloading turns into risk when self-monitoring lapses — as a habit, not a decision. AI inverts the retrieval sequence that consolidates learning, presenting the answer before the student has tried to produce it (Dunlosky et al., 2013). Fan et al. (2025) measured the result: AI assistance raised output scores while producing no gain in knowledge or transfer — a closed generate-submit loop in which “do I understand this?” is never asked.

Stage 3: **Illusion of capability**

Once monitoring fades, the student loses the internal signal that would flag the gap between performing and understanding. The code compiles, the grade holds, and the self-assessment stays intact — while the capability beneath it has not formed. Fluency in the moment is mistaken for durable mastery (Koriat & Bjork, 2005), and Prather et al. (2023) name the computing form of this the illusion of capability. It is self-concealing by construction: the schema needed to notice the gap is the one the gap has prevented from forming.

Stage 4: **Never-skilling**

As Table 1. shows never-skilling is structurally different from deskilling. Deskilling assumes a skill fades over time, such as when a doctor relies on AI for diagnoses and their own judgment weakens, or when a pilot's flying skills decline after years of using autopilot. But a student cannot lose a skill they never developed in the first place. They simply fail to build it because AI is handling the very work through which it would have formed (Macnamara et al., 2024). The missing ingredient is productive struggle: difficulty, error, and independent resolution are not obstacles to skill but its mechanism (Ericsson et al., 1993). When AI removes all three too early, it can remove the work through which durable competence forms.

Stage 5: **Substitutive AI use**

In the end state, the student stops forming independent judgment and transfers the function to the tool. Scaffolding is meant to fade as the learner grows (Vygotsky, 1978; Wood et al., 1976). A support that never withdraws becomes a dependency, and AI is structurally incapable of withdrawing on its own: it answers every query with the same completeness regardless of who is asking. The output is then accepted rather than evaluated, indistinguishable from competence until the tool is taken away (Shaw & Nave, 2025).

Stage 6: **Cohort divergence**

The first five stages describe what happens inside one student. The sixth changes scale — and this is where the stakes become structural. Caosun and Aral (2026) show that when AI output no longer depends on the user’s judgment, a single cohort using a single tool divides in two along the line of prior skill. Table 2. sets out the two trajectories side by side.

Table 2. Divergent AI-use trajectories by prior foundational capability

Formation principle	**Student with foundational capability**	**Student without foundational capability**
Prior schema condition	Arrives with a usable foundational schema	Arrives before the relevant schema has formed
Evaluation of AI output	Filters AI output through existing judgment	Accepts AI output without reliable evaluation
Cognitive engagemen t	Uses AI output as material for refinement	Treats AI output as an endpoint
Role of AI	AI use becomes complementary	AI use becomes substitutive
Formation mechanism	Capabilities remain in active use	Capability formation is bypassed
Developm ental trajectory	Capability accumulates	Never-skilling trajectory begins
Likely outcome	Continued development of independent competence	Risk of a widening capability gap

The same tool, in the same classroom, produces both trajectories at once. For IT students the dividing line falls early: the foundational capabilities for much of the computing curriculum are still forming in the first year or two, so AI adoption at that point is not neutral. It becomes risky not because the tool is harmful in itself, but because the schema that would make its use complementary has not yet formed.

Read forward, the cascade is a warning. Read backward, it is a map: every stage is a point at which engagement, if preserved, can interrupt the cascade. The same model that predicts the widening gap also identifies who escapes it — the students whose use of AI stays complementary rather than substitutive. Whether that is teachable, and what it looks like in practice, is what the rest of this paper sets out to show.

4 Convergent Evidence

No study yet tests this on IT students directly — as Section 2 noted, the evidence is indirect. What it has instead is convergence: three independent randomised controlled trials, in three different settings, isolating the same effect. The closest in domain is a coding-learning task, though its participants were experienced working programmers rather than students. The furthest — school mathematics — matters as much as the closest: when the same effect appears that far from computing, and a single design change switches it off, it is hard to read as an artefact of one setting. Writing sits in between. None is a perfect match for the IT student; taken together they are difficult to dismiss.

4.1 Coding — Shen and Tamkin (2026)

The study used a randomised controlled design: 52 professional and freelance programmers learned a new Python library either with or without AI coding assistance. Both groups produced working code during the session, and the AI-assisted group produced higher-quality immediate output. Afterwards, both sat an independent knowledge assessment with no AI.

On that assessment the AI-assisted group scored about 17 percentage points lower than the control group, with no offsetting productivity gain during the session. The gap was largest in debugging — the skill that most depends on forming hypotheses about failure states, which develops only by meeting and resolving real failures. The unaided group encountered more errors while learning, and higher error-encounter went with better mastery: the productive struggle was not the obstacle but the instruction.

The most useful finding was that the outcome depended on how participants used AI, not whether. Six engagement patterns emerged; the three that kept the participant thinking after the AI answered scored markedly higher than the three that delegated the work. The subgroups are small, so the typology is qualitative rather than precise — but the direction is the point: the performance gap is not between AI users and non-users, but between those who engaged with the output and those who accepted it. That distinction is teachable.

4.2 Mathematics — Bastani et al. (2025)

The effect travels far from coding. In a pre-registered field trial with nearly a thousand school students (grades 9–11), unrestricted GPT-4 access improved performance on AI-assisted practice but left students about 17% worse on unassisted end-of-session exams than controls. The cause was cognitive bypass, not AI error: most requests asked directly for answers rather than method, and the tool's own mistakes were not what predicted exam failure.

The decisive detail is the guardrail. A second condition — a tutor mode giving hints and scaffolding instead of answers — removed the learning penalty entirely while keeping the practice gain. The same tool, redirected from answering to prompting, stopped doing harm: a design that preserves engagement preserves learning.

4.3 Writing — Fan et al. (2025)

The same split appears in writing. In a study of 117 learners on a multi-stage writing-and-revision task, AI assistance raised the quality of the written output but produced no gain in knowledge and none in transfer to a new task: the product improved while the learning did not. Fan and colleagues name the mechanism metacognitive laziness — when the tool handles the monitoring, evaluating, and checking, the learner stops doing them. It is the clearest statement of the pattern running through all three studies: better output is not better learning.

Beyond the three trials, the same shape recurs. Schema-incomplete students engage only at the surface, with no sign of the structure forming (Wang et al., 2025); those who cannot judge AI output accumulate errors through steps that each inherit the last (Reihanian et al., 2024). The signal is already visible in the profession these students are training for: 20% of experienced developers report declining confidence in their own problem-solving, and roughly one in six regularly meet AI-generated code they cannot understand (Stack Overflow, 2025).

5 Student-Facing Practices

By now the case runs in both directions. The cascade is real and measurable: AI use at the formation stage, left unstructured, runs from offloading to never-skilling and leaves a capability gap the student cannot see. While a significant performance gap exists between students who use AI and students who avoid it, a critical secondary gap divides AI users themselves: it lies between those who keep thinking after the tool answers and those who passively accept its output and move on. That difference decides which side of the divergent-trajectory line a student lands on, and it is teachable.

This paper's response is a single posture that makes that engagement deliberate, and it is the paper's central contribution: hold, rethink, continue.

Hold is a short pause before accepting AI output — long enough to ask: "do I actually understand what this is, and why should I accept it?". **Rethink** is reconstructing the reasoning behind the output: why does this solution work, what assumptions sit beneath it, how would it change if the input, the constraints, or the scale were different — and what could go wrong without your noticing? This is not a checklist; it is the habit of stepping back from the result to inspect your own understanding — metacognition in its most practical form. **Continue** is re-engaging the tool with a sharper question: rather than asking for the answer, ask for a hint, a contrasting case, or why an approach fails under a particular constraint, then return better able to judge the response.

Hold–rethink–continue is not a new theory but a way of putting an established one to work. Zimmerman's (2002) model of self-regulated learning runs in three cyclical phases — forethought (goal-setting and planning), performance (monitoring your own work as you go), and self-reflection (evaluating the result and adjusting) — phases Panadero (2017) finds consistent across the major learning-science traditions. Generative AI quietly switches each one off: forethought is skipped when the answer is already there, performance monitoring collapses into accepting the output, and self-reflection never happens when you take the tool's verdict instead of forming your own. Hold–rethink–continue is the posture that switches them back on. The label, and its application to the single decision of when to lean on AI and when to withhold, are this paper's; the underlying cycle is Zimmerman's (2002; Panadero, 2017).

This posture is not idiosyncratic. Reviewing thirty-nine studies of AI and critical thinking, Nasr et al. (2026) independently arrive at a similar structured, human-in-the-loop sequence for keeping the learner in control — convergent derivation from a different literature, not external validation of this one.

The point is not to use AI less, but to use it as the starting point of thinking rather than the end. Four practices put the posture into action, each grounded in the evidence above.

5.1 The Explanation Test

Before submitting any AI-generated output, close the tool and explain the solution's logic in three sentences — not what it does, but why it is built this way. If the explanation comes, the schema is intact and the AI was an accelerant; if it does not, you have accepted a solution you do not understand. It costs under two minutes, and it maps directly onto Shen and Tamkin's (2026) high-scoring patterns, which all required engaging with **why** the output works, not only whether it does.

5.2 Deliberate Non-Use

Pick specific task types — debugging, first-pass decomposition, algorithm design from scratch — and practise them without AI, as a regular discipline rather than a moral stance. Shen and Tamkin's (2026) control group, working unassisted, scored higher on the independent test; the errors they hit were the formation mechanism. Deliberate non-use recreates that condition on purpose, early, before delegation becomes the default — and Bastani et al.'s (2025) guardrail (hints, not answers) shows the same at n≈1,000: the design that preserved the struggle preserved the learning.

5.3 Interaction Pattern Awareness

What separates the high-scoring patterns from delegation is not effort but the direction of attention **after** the AI responds. Conceptual inquiry means asking why — why this algorithm, why this structure, what would make it fail, what alternative was passed over and why. It turns the output from an endpoint into a starting point — and it has to be decided before a session begins, not improvised within it.

5.4 Error-Awareness Practice

Being told you are responsible for AI output does not make you check it: Kupfer et al. (2023) found that participants reminded of their legal accountability changed no verification behaviour, while those told concretely how the tool fails scrutinised far more (d=0.55–0.67). What activates scrutiny is specific knowledge of failure modes. In computing, four recur: hallucinated API calls that do not exist until runtime; the right algorithm solving the wrong problem because the task was under-decomposed; logic-broken code that still passes style and security checks; and confident wrong output on unfamiliar libraries. Those are the conditions to verify against — generic reminders will not produce the check; knowing how the tool breaks will.

6 Discussion

This paper set out to integrate established cognitive-science accounts of skill formation with recent empirical findings on AI-assisted learning into a single mechanism — how generative AI use at the formation stage, absent metacognitive discipline, produces never-skilling — and, from that mechanism, to derive a teachable response. Its contribution is therefore neither a new empirical result nor a literature survey but a synthesis. That synthesis resolves to a single principle: hold–rethink–continue, a metacognitive posture that operationalises self-regulated learning (Zimmerman, 2002; Panadero, 2017) as a concrete decision procedure for when to lean on AI and when to withhold. The principle is earned by the cascade that

precedes it: the cognitive risk cascade (Section 3) reframes a set of separately documented findings as one connected chain. Its evidential weight rests on triangulation rather than on any single study. Established learning science specifies the conditions competence formation requires; three independent randomised controlled trials — across coding (Shen and Tamkin, 2026), mathematics (Bastani et al., 2025), and writing (Fan et al., 2025) — demonstrate that current AI use systematically removes them. Two implications follow.

Theoretically, never-skilling is structurally distinct from deskilling — **a difference of kind**, not degree (Table 1.). That distinction is what makes the double exposure of IT students — a learning environment and a profession restructured by the same technology at once — a categorically different condition from the single disruption other disciplines face.

Practically, the locus of the response is divided: the four student-facing practices in Section 5 that enact hold–rethink–continue are individually low-cost and immediately actionable, while the structural conditions that make them necessary fall to institutions — the educational response developed in the companion paper presented in parallel. What that synthesis means for the individual student — the party who bears the formation risk but did not create it — deserves to be stated plainly, because it is the hardest part of the double exposure to see from inside.

Part of what makes this condition hard to address is that it is hard to perceive from inside. The student arrives to build a career, not to monitor their own cognition, and by then the tool that will shape both is already present — an ambient layer in the editor and the browser that asks nothing to begin. A student can watch the industry change; they cannot as readily watch the same tool change what they think with, not only what they work on. Earlier technologies entered through the curriculum, slowly enough for teachers to update and students to adjust; this one arrived underneath it, and neither side was prepared. When AI helps a student pass, ship, and earn the credential, it looks like nothing but help, and never exposes the fork it conceals: that the same output can be the start of understanding or its end. The individual student can hardly be held culpable for this: a layer one cannot see is not one that was agreed to, and a fork one was never shown is not a choice that was declined.

That blindness is structural, not personal, which is why the response cannot rest on the student alone. The heavier obligation falls on institutions — to redesign assessment, sequence curriculum, and configure AI so that productive struggle survives. It is a prospective duty, not a charge of negligence: the layer arrived faster than any curriculum cycle was built to answer. How institutions should meet it is the subject of the companion paper presented in parallel. But a structural cause does not dissolve the individual stake — it sharpens it.

The professional bottleneck is shifting from generation to verification. AI produces code from natural language with rapidly improving fidelity; what stays scarce — and is intensifying as a requirement — is the judgment to tell whether that code is correct, secure, efficient, and appropriate for its context. That judgment runs on exactly the schemas that delegation at the formation stage prevents from forming. And the gap is easy to miss: "code runs" and „I understand why this code is correct" look identical until the tool is taken away. Stack Overflow (2025) supplies the professional signal: 66% of experienced developers name "AI solutions that are almost right, but not quite" as their primary frustration. The schema needed to catch „almost right" is the one that forms through the struggle delegation removes. Under the divergent trajectories modelled by Caosun and Aral (2026), the students who reach the formation stage without it do not catch up.

Three limitations bound these claims. First, none of the evidence is longitudinal. The trials measure performance immediately after a single session, not months later, so the widening gap between assisted and unassisted students is a model's projection (Caosun & Aral, 2026), not an outcome followed across a career; whether the deficit persists, or can later be recovered, remains open. Second, almost no study observes IT students directly. Even the closest case (Shen & Tamkin, 2026) used working programmers — professional and freelance, of varying experience — learning an unfamiliar library, not students; the argument therefore rests on the three nearest domains available — coding, school mathematics, and writing (Section 4) — which point the same way but leave the effect size for IT students unmeasured. Third, the practices in Section 5 are calibrated to how current AI fails; as the tools improve, those specific failure modes will shift, and the practices will need periodic revision — though the underlying need to verify AI output will not.

These limitations double as a research agenda. The first calls for a longitudinal study that tracks a single cohort across its formation, testing whether assisted and unassisted performance diverge or converge. The second calls for studies situated in IT education itself, across specialisations beyond programming. The third calls for the practices to be revisited as the tools evolve. None is a reason to defer action: too many students are forming their habits now. The direction is consistent with emerging GenAI-specific design work that treats large language models as scaffolds for critical thinking rather than substitutes for learner judgment (Vendrell & Johnston, 2026).

That **hold–rethink–continue** operationalises established self-regulated learning (Zimmerman, 2002; Panadero, 2017), rather than an ad hoc rule, is what makes it a testable proposition — and what the agenda above is designed to test.

7 Conclusion

The window is open now, and it does not stay open. The four practices in Section 5 are not a stand against AI; they are how you make the invisible fork visible and choose the side that builds the schema. Each one keeps you thinking after the tool answers — which is the whole game.

- **Hold** before accepting output you cannot yet explain.
- **Rethink** the reasoning until you could rebuild it without the tool.
- **Continue** — back to AI with a sharper question, now able to judge what it gives you.

Good equipment helps a trained rider win the race. It cannot make you a rider. AI is the equipment; the formation stage is where you become the professional who can use it to win — instead of the one it leaves behind.

References

Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences*, 122(26), e2422633122.

Caosun, M., & Aral, S. (2026). The augmentation trap: AI productivity and the cost of cognitive offloading *[arXiv preprint]*. https://arxiv.org/abs/2604.03501

CC2020 Task Force. (2020). Computing curricula 2020: Paradigms for global computing education. *ACM/IEEE-CS/AAAI*.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, 14 (1), 4–58.

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, 100 (3), 363–406.

Fan, Y., Tang, L., Le, H., Shen, K., Tan, S., Zhao, Y., Shen, Y., Li, X., & Gašević, D. (2025). Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. *British Journal of Educational Technology*, 56 (2), 489–530.

Gartner. (2025). Gartner top strategic technology trends: Software engineering intelligence platforms and AI code assistants. *Gartner, Inc*.

Gerlich, M. (2025). AI tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies*, 15 (1), 6.

Harari, Y. N. (2023, April 28). Yuval Noah Harari argues that AI has hacked the operating system of human civilisation. *The Economist.* https://www.economist.com/by-invitation/2023/04/28/yuval-noah-harari-argues-that-ai-has-hacked-the-operating-system-of-human-civilisation

Kahneman, D. (2011). Thinking, fast and slow. Farrar, *Straus and Giroux*.

Ke, Y., Jin, L., Ong, J. C. L., Thirunavukarasu, A. J., Car, J., Cheung, C. Y., Tham, Y. C., Ting, D. S. W., Ong, M. E. H., Compton, S., Narayan, A., Keane, P. A., Wong, T. Y., Bates, D. W., Tan, P., & Liu, N. (2026). AI-induced never-skilling in medical education. *Nature Medicine*, 32, 1997–2006.

Koriat, A., & Bjork, R. A. (2005). Illusions of competence in monitoring one's knowledge during study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31 (2), 187–194.

Kupfer, C., Prassl, R., Fleiß, J., Malin, C., Thalmann, S., & Kubicek, B. (2023). Check the box! How to deal with automation bias in AI-based personnel selection. *Frontiers in Psychology*, 14, 1118723.

Macnamara, B. N., Berber, I., Çavuşoğlu, M. C., Krupinski, E. A., Nallapareddy, N., Nelson, N. E., Smith, P. J., Wilson-Delfosse, A. L., & Ray, S. (2024). Does using artificial intelligence assistance accelerate skill decay and hinder skill development without performers' awareness? *Cognitive Research: Principles and Implications*, 9, 46.

Nasr, N. R., Tu, C.-H., Sujo-Montes, L., Haniya, S., Yen, C.-J., & Hossain, S. (2026). Generative artificial intelligence impact on critical thinking in higher education: A comprehensive bibliometric analysis and systematic review. *Journal of Computers in Education*.

O'Brien, K. L., & Leichenko, R. M. (2000). Double exposure: assessing the impacts of climate change within the context of economic globalization. *Global Environmental Change*, 10(3), 221–232.

Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology*, 8, 422.

Prather, J., Denny, P., Leinonen, J., Becker, B. A., Albluwi, I., Craig, M., Keuning, H., Kiesler, N., Kohn, T., Luxton-Reilly, A., MacNeil, S., Petersen, A., Pettit, R., Reeves, B. N., & Šavelka, J. (2023). The robots are here: Navigating the generative AI revolution in computing education. *In Proceedings of the 2023 Working Group Reports on Innovation*

and Technology in Computer Science Education (pp. 108–159). ACM.

Reihanian, I., Hou, Y., Chen, Y., & Zheng, Y. (2024). A review of generative AI in computer science education: Challenges and opportunities in accuracy, authenticity, and assessment. *In International Conference on Computational Science and Computational Intelligence* (pp. 144-158). Springer Nature Switzerland.

Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences*, 20(9), 676–688.

Shaw, J., & Nave, C. S. (2025). Thinking — fast, slow, and artificial: Navigating cognitive biases in the age of AI *[SSRN Working Paper]*. https://ssrn.com/abstract=5126527

Shen, Z., & Tamkin, A. (2026). How AI assistance impacts skill formation *[arXiv preprint]*. https://arxiv.org/abs/2603.04825

Stack Overflow. (2025). Stack Overflow 2025 developer survey. https://survey.stackoverflow.co/2025

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12 (2), 257–285.

Vendrell, M., & Johnston, S.-K. (2026). Scaffolding critical thinking with generative AI: Design principles for integrating large language models in higher education. *Computers and Education: Artificial Intelligence*, 10, 100572.

Vygotsky, L. S. (1978). Mind in society: The development of higher psychological processes. *Harvard University Press*.

Wang, X., Tang, H., & Yu, Q. (2025). Prior domain knowledge shapes cognitive outsourcing patterns in generative AI use: Evidence from a mixed-methods study. *Acta Psychologica Sinica*, 57 (3), 441–459.

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry*, 17 (2), 89–100.

World Economic Forum. (2025). Future of jobs report 2025. *World Economic Forum*. https://www.weforum.org/reports/the-future-of-jobs-report-2025/

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory into Practice*, 41 (2), 64–70.

Exploring Authentic Learning Scenarios in Online and Blended Environments Across European Universities

Michael Eichhorn, Julia Schmitt, Alexander Tillmann
Goethe University Frankfurt, studiumdigitale, Germany
{eichhorn,schmitt,tillmann}@sd.uni-frankfurt.de

Josmario Albuquerque, Paul Astles, Catriona Matthews, Bart Rienties
The Open University, Institute of Educational Technology, UK
{albuquerque,astles,matthews,rienties}@open.ac.uk

Blaženka Divjak, Tanja Šestanj-Perić
University of Zagreb Faculty of Organization and Informatics, Croatia
{bdivjak,tanja.peric}@foi.hr

Egle Gedrimiene, Hanni Muukkonen
University of Oulu, Finland
{egle.gedrimiene,hanni.muukkonen}@oulu.fi

Mirza Žižak
University of Zagreb, Faculty of Medicine
mirza.zizak@mef.hr

Abstract. *This paper explores authentic learning scenarios for online and blended higher education across four European universities. Grounded in situated cognition theory and Herrington et al.'s authentic learning design framework, we report a two-phase empirical study. In Study 1, 13 learning design experts used consensus workshop methodology to identify 11 applicable scenarios. In Study 2, 95 practitioners in 13 focus groups examined benefits, challenges, and enablers of implementation. Two structural clusters emerged: method-focused scenarios primarily demand pedagogical expertise and facilitation skills; technology-focused scenarios require technical infrastructure and digital competence. Professional development, institutional support, and learning design examples prove crucial cross-cutting enablers.*

Keywords. authentic learning; blended learning; learning design; higher education; focus groups; hybrid teaching; professional development; innovative pedagogy

1 Introduction

The rapid expansion of online and blended learning in higher education has renewed sustained interest in authentic learning scenarios; pedagogical approaches that engage students with real-world problems, collaborative inquiry, and meaningful knowledge construction (Donovan et al., 1999; Herrington et al., 2014). Over the past decade, more than one hundred such scenarios have been catalogued in influential reports such as the Innovating Pedagogy series (Kukulska-Hulme et al., 2022; Sharples et al., 2012), and a recent analysis confirmed strong alignment between emerging pedagogical proposals and empirical research trajectories (Liu et al., 2024).

Despite this proliferation, substantial evidence shows that many higher education practitioners struggle to implement innovative learning scenarios effectively, particularly in online and blended environments (Nelson et al., 2019; Rienties et al., 2022; Tondeur et al., 2012). Key barriers include limited pedagogical and digital competencies, insufficient institutional support, and inadequate professional development (Bragg et al., 2021; Koehler & Mishra, 2006). While individual studies have examined specific scenarios, e.g. flipped classrooms (Divjak, Rienties et al., 2022), problem-based learning (Dochy et al., 2003), and AI in education (Holmes & Tuomi, 2022), cross-institutional, multi-country studies systematically combining expert consensus with practitioner experience remain scarce.

This paper addresses that gap through a two-phase study involving four European universities in Croatia, Finland, Germany, and the United Kingdom, conducted as part of the Erasmus+ project 'Innovating Learning Design in Higher Education' (iLed). We investigate the following research questions:

RQ1: Which authentic learning scenarios do higher education experts consider most applicable across diverse European institutional contexts?

RQ2: What benefits, challenges, and enabling conditions do practitioners identify when implementing these scenarios in online and blended environments?

The study's contribution is threefold: it provides, to our knowledge, the first systematic cross-institutional expert consensus on applicable authentic learning scenarios across four European higher education contexts; it combines expert and practitioner perspectives in a single study design; and it generates an empirically grounded typology (method-focused versus technology-focused scenarios) that offers a practically applicable framework for professional development design and institutional learning design policy.

2 Theoretical Background

2.1 Authentic Learning: Conceptual Foundations

The concept of authentic learning is rooted in situated cognition theory, which holds that knowledge is inseparable from the context of its use (Brown et al., 1989; Lave & Wenger, 1991). Building on cognitive apprenticeship principles, Herrington and Oliver (2000) operationalised authentic learning through nine design characteristics of authentic learning environments: an authentic real-world context, authentic ill-defined activities, access to expert performance, multiple roles and perspectives, collaborative knowledge construction, reflection, articulation, coaching and scaffolding, and authentic integrated assessment.

Critically for the present study, Herrington et al. (2014) clarified that these characteristics describe design qualities, not inherent properties of any particular method or technology. A flipped classroom, a virtual reality environment, or an AI-supported tutorial can each facilitate authentic learning when the design ensures engagement with real-world problems, access to expertise, and meaningful assessment; or fail to do so when these conditions are absent. We therefore define authentic learning scenarios as pedagogical approaches, whether method-focused or technology-mediated, intentionally designed to activate students and enable them to meaningfully construct knowledge in contexts involving real-world problems relevant to their disciplinary or professional practice (Donovan et al., 1999). Learning design refers to the deliberate alignment of learning activities with intended outcomes through constructive alignment (Biggs, 2014; Divjak, Grabar et al., 2022).

2.2 The Implementation Gap and Professional Development

Despite theoretical consensus, a well-documented implementation gap persists between pedagogical innovation and classroom practice (Nelson et al., 2019; Rasheed et al., 2020). The Technological Pedagogical Content Knowledge framework (TPACK; Koehler & Mishra, 2006) helps explain this gap: effective use of authentic learning scenarios in technology-mediated environments requires the integration of technological, pedagogical, and content knowledge in ways many educators have not had the opportunity to develop. Institutional constraints like workload, resource limitations, and weak collegial support, further restrict adoption (Bragg et al., 2021; Brown, 2016).

In this study, TPACK serves as an interpretive lens rather than a deductive coding framework: rather than applying TPACK categories a priori to the focus group data, we use the framework post-hoc to interpret the emergent pattern of challenges and enablers, specifically to explain why method-focused and technology-focused scenarios produce structurally different demands on educators. The distinction between pedagogical knowledge demands (dominant in MF scenarios) and technological knowledge demands (dominant in TF scenarios) directly reflects the core TPACK distinction between the pedagogical and technological knowledge components.

A cross-institutional professional development study found that while online professional development can improve educators' competencies with authentic learning scenarios, impacts are shaped by institutional culture and disciplinary context (Rienties et al., 2023).

Taken together, these findings suggest that the question of which authentic learning scenarios exist is secondary to the question of under what conditions educators in specific institutional contexts perceive them as feasible, worthwhile, and amenable to implementation. It is precisely this practitioner-level, cross-institutional understanding that the present study seeks to provide.

3 Methods

3.1 Research Design

This paper reports a two-phase empirical study conducted in 2023. Study 1 used expert consensus methodology to identify applicable authentic learning scenarios. Study 2 used focus group research with practitioners to explore perceived benefits, challenges, and enabling conditions for each scenario. Ethical approval was obtained at the coordinating institution; participation was fully voluntary and informed. Fig. 1 shows the timeline of the two studies.

3.2 Study 1: Expert Consensus Workshop

Thirteen learning design experts (6 male, 7 female) from four institutions participated in a two-day in-person consensus workshop in May 2023. The four institutions represent diverse European higher education contexts: a large multi-disciplinary Croatian university (CU1, >70,000 students), a Finnish research university (FU2, ~14,000 students), a German research university (GU3, >43,000 students), and a large UK distance-learning university (UU4, >200,000 students). Collectively, participants had published over 900 peer-reviewed articles and had experience with learning design across more than 400 courses.

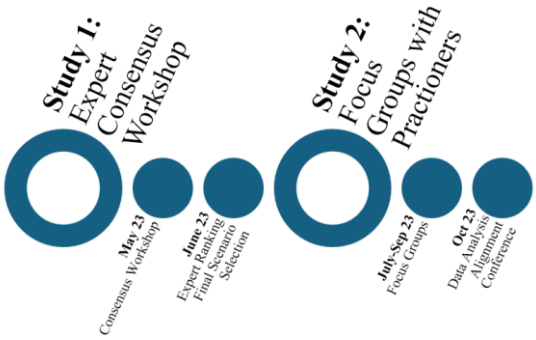


Figure 1. Timeline of Study 1 and Study 2 in 2023.

Using the Consensus Workshop Method (ICA:UK, 2014), participants identified 16 initial scenarios through a two-hour facilitated brainstorming session. Experts then independently ranked all 16 scenarios in an online survey (response rate: 69%) over two weeks, rating relevance to their institution on a 1 (most relevant) to 10 (least relevant) scale. Results were discussed in a one-hour online follow-up meeting, leading to the selection of 11 scenarios[1] with an average rank below 6.0 and cross-institutional applicability as the key selection criteria.

3.3 Study 2: Focus Groups with Practitioners

Between July and September 2023, 13 focus groups were conducted with 95 higher education practitioners (63 female, 32 male) across the four institutions, including experienced university teachers and pre-service educators. Each focus group followed a structured five-step procedure: participants received scenario materials in advance; voted on two scenarios to discuss; engaged in 30–45 minute discussions per scenario guided by prompts on benefits, learning outcomes, and implementation challenges; and the facilitator summarised findings. Sessions were conducted in participants' local language, recorded, transcribed with Amberscript, and fully anonymised.

After that, Data were analysed using Qualitative Content Analysis (Mayring, 2015, 2023) with an inductive approach in which the category system emerged concurrently with analysis. Two coders per institution conducted the coding; inter-coder reliability was verified through joint evaluation conferences. Three superordinate categories (*benefits*, *challenges*, and *enablers*) were set deductively as starting categories across all four institutions.

The uneven distribution of scenario discussions across institutions, e.g. Problem-Based Learning appearing only at CU1 and VR only at CU1 and GU3, reflects genuine differences in institutional priorities and teaching cultures rather than a sampling error. To mitigate the risk of institution-specific patterns skewing the inductive category system, coders at each institution developed their sub-categories independently before a joint cross-institutional alignment conference, where categories were compared, merged, and renegotiated until consensus was reached. These 2-day alignment conference with all coders took place in Finland in October 2023.

4 Results

4.1 Study 1: Expert Rankings and Scenario Selection

Table 1 presents the full ranking results for all 16 scenarios. The 11 scenarios marked 'Yes' were selected for Study 2 based on an average rank below 6.0 and confirmed cross-institutional applicability.

Table 1. Expert ranking results
(n=13; Avg.=average rank (1=most relevant, 10=least relevant); Med.=median; SD=standard deviation; Sel.= selected for Study 2; *excluded despite high rank, see text)

Scenario	Avg.	Med.	SD	Sel.
Blended Learning	2.33	1	2.16	No*
AI in Education	2.38	2	1.60	Yes
Student Engagement in Learning Design	2.43	2	1.51	Yes
Hybrid Teaching (large groups)	3.50	3	2.51	Yes
Flipped Classroom	3.71	2	3.04	Yes
Project-Based Learning (virtual)	3.75	3	2.92	Yes
Hybrid Teaching (small groups)	4.00	4	2.19	Yes
Problem-Based Learning	4.00	4	2.10	Yes
Virtual Reality / Augmented Reality	4.00	3	3.00	Yes

[1] For a detailed description of the 11 selected authentic learning scenarios, see: 10.6084/m9.figshare.32788440

Inquiry-Based Learning	4.14	4	2.54	Yes
Game-Based Learning	4.50	5	2.43	Yes
Simulation (SSP)	5.80	6	3.83	Yes
Authentic assessment tasks supervised by Employers	6.25	7	3.86	No
Fully online Learning	6.50	8	3.73	No
Work-Based Learning (online)	7.71	9	2.93	No
Student placement and shadowing	7.80	8	2.28	No

Blended learning ranked highest overall but was excluded because experts judged the term too broad: most teaching at the three face-to-face institutions could already be described as blended, making it more informative to specify a particular pedagogical approach within blended settings. Fully online learning and work-based learning online were excluded as not applicable across all four institutions, particularly UU4, which as a fully online university could not meaningfully discuss scenarios requiring physical classroom presence. Student Placement and Shadowing and Authentic Assessment Tasks Supervised by Employers were excluded because, despite their conceptual alignment with authentic learning principles, experts judged them as requiring employer partnerships and organisational arrangements that fall outside the direct pedagogical control of higher education institutions. As such, they represent valid forms of authentic learning but constitute a distinct category of work-integrated learning that demands different institutional infrastructure and is beyond the immediate scope of this study's focus on in-course pedagogical design.

The resulting 11 scenarios span method-focused approaches (Flipped Classroom, Game-Based Learning, Inquiry-Based Learning, Problem-Based Learning, Project-Based Learning, Simulation, Student Engagement in Learning Design) and technology-focused approaches (AI in Education, Hybrid Teaching in Large and Small Groups, Virtual Reality/AR).

4.2 Study 2: Distribution and Overview

In total, 34 scenario discussions were conducted across the 13 focus groups. Table 2 shows the distribution per institution.

Table 2. Scenario discussions per institution
(+ = one focus group discussion)

Scenario	CU1 (4 FG)	FU2 (4 FG)	GU3 (2 FG)	UU4 (3 FG)
Flipped Classroom	++	+++	++	++
AI in Education	+	++	++	+
Hybrid – Large	+	+		+
Game-Based Learning		+		+
Hybrid – Small	+	+		
Inquiry-Based	+			+
Problem-Based	++			
Project-Based	+	+		
Simulation (SSP)	+			+
Student Eng. in LD		+		+
VR/AR	+		+	

Flipped Classroom was the most frequently discussed scenario (9 discussions across all four countries), followed by AI in Education (6 discussions). These were also the only two scenarios discussed in all four national contexts, reflecting both widespread prior experience (Flipped Classroom) and strong interest in an emerging topic despite limited practical experience (AI in Education). Table 3 provides a comprehensive overview of the benefits, challenges, and enabling conditions identified for all 11 scenarios.

Table 3. Summary of benefits, challenges, and enablers by scenario
(MF=method-focused; TF=technology-focused)

Scenario (Type)	Benefits	Challenges	Enablers
Flipped Classroom (MF)	Active learning, interaction, motivation	Student non-preparation, teacher workload	Clear expectations, supportive culture
AI in Education (TF)	Individualised support, shifts learning outcomes	Guidelines, legal/ethical concerns	Didactic input, teacher testing, student co-development
Hybrid – Large (TF)	Remote inclusion, digital skills	High preparation, dual-audience demands	In-class assistant, structured planning
Game-Based (MF)	Motivation, flexible methods	Perceived non-educational, design effort	Clear objectives, safe game space
Hybrid – Small (TF)	Collaborative formats, flipped pairing	Technical demands, effectiveness scepticism	Improved infrastructure

Inquiry-Based (MF)	Critical thinking, co-creation	Sustaining motivation, grading flexibility	Practice examples, clear instructions
Problem-Based (MF)	Theory application, motivation via context	Intensive involvement, over-use risks	Curriculum integration, AI support
Project-Based (MF)	Real-world feedback	Needs theory background, large groups	Scenario examples, peer feedback
Simulation/SSP (MF)	Safe, realistic, clinical preparation	Authentic experience design, cost	Combination methods, AI
Student Eng. in LD (MF)	Student responsibility, diverse perspectives	Unfamiliar to teachers, resource constraints	Co-teaching, workshops, clear examples
VR/AR (TF)	Practical content, AI convergence, flexibility	Cost, experience limits, technical barriers	Funding, instructor training

4.3 Detailed Findings: Five Focal Scenarios

To illustrate the range of findings in depth, we present detailed results for five scenarios selected on three complementary criteria: discussion frequency (Flipped Classroom and AI in Education were the most frequently discussed, each in all four countries), thematic and structural variance (Hybrid Teaching and Virtual Reality represent the technology-focused cluster with contrasting maturity profiles), and institutional divergence (Problem-Based Learning was discussed exclusively at CU1, offering the clearest example of discipline-specific contextualisation). Together, these five scenarios span both structural clusters and cover the full spectrum from widely practised to emerging.

4.3.1 Flipped Classroom

The Flipped Classroom was by far the most extensively discussed scenario, with 9 focus group sessions (52 participants) across all four countries. Practitioners widely recognised its potential to enhance active learning during class time and to strengthen teacher-student interaction, shifting face-to-face time from information transmission toward discussion, problem-solving, and application. Several participants described observed gains in student engagement and intrinsic motivation when pre-class materials were carefully designed to build genuine curiosity for the in-class phase.

The most frequently cited challenge was uneven student preparation: when students did not engage with pre-class materials, the in-class phase collapsed, placing considerable pressure on teachers to improvise. This was especially pronounced in large cohorts where monitoring preparation rates is difficult. Practitioners from CU1 noted that the flipped approach worked best in smaller seminar formats, while GU3 and FU2 participants reported more mixed outcomes in large lecture-based courses. Increased workload for teachers, particularly in creating high-quality pre-class video or reading materials, was a near-universal concern.

Key enablers identified included a supportive departmental culture in which flipped teaching is a shared norm rather than an individual experiment, clear communication of expectations to students from the outset, and institutional provision of repositories containing reusable pre-class materials. Participants at UU4 noted that the flipped model closely parallels existing distance-learning design principles, making it relatively transferable to that context.

4.3.2 AI in Education

AI in Education generated the most animated discussions across the study, driven by the rapid emergence of generative AI tools in the months preceding data collection. Unlike most other scenarios, practitioners discussed AI primarily from a position of curiosity and concern rather than established experience: the majority had limited hands-on implementation of AI tools in their teaching and treated the focus group as an opportunity to collectively think through implications.

Perceived benefits centred on the potential for individualised learning support: AI providing tailored feedback at scale, supporting students who fall behind, and enabling personalised learning pathways beyond what any individual teacher could offer. Participants also acknowledged that AI fundamentally challenges traditional assessment design, with several noting that assignments requiring AI-generated answers needed to be reconceived entirely to assess higher-order competencies. This was framed not merely as a challenge but as a productive pressure to reconsider what authentic assessment means.

The primary challenges were institutional and ethical: participants across all four countries called urgently for clear guidelines on acceptable AI use, legal frameworks governing data privacy and copyright, and training to help educators use AI purposefully rather than reactively. A recurring enabler was the proposal to develop AI guidelines collaboratively with students rather than imposing top-down rules, fostering shared ownership and transparency. Participants in GU3 and FU2 also highlighted the value of teachers experimenting personally with AI tools before introducing them into courses, to develop pedagogical judgment about their affordances and limitations.

4.3.3 Hybrid Teaching

Hybrid Teaching, enabling simultaneous participation by students on-site and online, was discussed in seven focus groups across three of the four institutions (CU1, FU2, UU4), in both large-group and small-group formats. Practitioners valued it primarily for its inclusivity benefits: students with mobility constraints, work commitments, or geographic distance could participate without being disadvantaged, and the digital participation modality was seen as developing skills of increasing importance in professional life.

However, hybrid teaching also generated some of the strongest expressions of implementation difficulty in the study. Managing two concurrent audiences, one physical, one virtual, requires sustained dual attention that many participants described as exhausting and cognitively demanding. The risk of one group (typically the online participants) becoming passive observers rather than active participants was repeatedly raised. Participants from FU2 described asymmetric participation patterns in which online students spoke less frequently and were overlooked in spontaneous discussion. Digital infrastructure failures, e.g. particularly camera placement, microphone coverage, and video latency, were cited as significant disruptors, especially for small-group collaborative formats.

The most consistently cited enablers were technical: a dedicated in-class assistant to manage the online stream and flag contributions from remote participants, and a structured lesson plan that explicitly allocates interaction time to both groups. Several participants noted that combining hybrid teaching with flipped classroom principles, so that information transfer happened asynchronously before class, substantially reduced the cognitive load of the in-class facilitation. Improved technical infrastructure, including dedicated hybrid teaching rooms, was identified as the single most important institutional investment.

4.3.4 Problem-Based Learning

Problem-Based Learning (PBL) was discussed in two focus groups at CU1, reflecting the strong tradition of problem-based approaches in health professional education at that institution. Practitioners valued PBL primarily for its capacity to bridge theoretical and applied knowledge: working through concrete, realistic cases motivated students in ways that abstract instruction did not. Also the collaborative group process was seen as developing professional competencies like teamwork, communication, clinical reasoning, alongside content knowledge.

The central challenge identified was the intensity of involvement required from both teachers and students. Effective PBL requires sustained scaffolding: teachers must design or select authentic problems carefully calibrated to current student knowledge, facilitate without over-directing, and provide timely feedback throughout the process rather than only at the end. Participants noted a paradox of motivation: while well-implemented PBL was highly motivating, excessive or poorly scaffolded PBL, where problems are too abstract, too easy, or poorly aligned with assessment, could produce frustration and disengagement. Curriculum integration was therefore identified as an essential condition: PBL worked best when it was embedded coherently across a course module rather than treated as an isolated activity.

Structured problem-scenario templates were highlighted as a practical enabler that reduced the design burden on individual teachers. Participants also raised AI as a potential support tool within PBL: AI could assist students in accessing background information, provide iterative feedback on draft solutions, and help teachers generate or adapt problem scenarios more efficiently. This emergent integration of AI as a support tool, not a replacement for the PBL process, was seen as a promising development.

4.3.5 Virtual Reality and Augmented Reality

Virtual Reality and Augmented Reality (VR/AR) was discussed in three focus groups (CU1, GU3) and generated a notably bifurcated response: participants who had hands-on experience with VR/AR were enthusiastic about its potential; those without direct experience were largely sceptical and uncertain. The primary benefit cited by experienced practitioners was the creation of safe, repeatable learning environments for high-stakes professional practice; particularly in medicine and engineering, where errors in real settings carry serious consequences. Participants from CU1's medical faculty described VR simulation as transformative for procedural skills training.

The most prominent challenges were cost and access: high-quality VR hardware remains expensive, and few institutions have sufficient headsets for whole-cohort deployment. Technical limitations were also noted, e.g. latency, display quality, user discomfort during prolonged use. A further challenge was the limited experience of most teaching staff with VR environments, both as users and as designers, creating a double barrier of personal unfamiliarity and design complexity.

Participants identified a convergence with AI as a particularly promising direction: AI-driven adaptive scenarios within VR environments could personalise difficulty and feedback in ways that static VR content cannot. Enablers at the institutional level included dedicated funding streams for equipment, partnerships with industry or healthcare providers to develop and maintain VR content, and structured instructor training programmes. A programme-level approach, i.e. integrating VR into multiple courses rather than a single elective, was recommended to justify the investment and build staff competence over time.

5 Discussion

The findings confirm and extend existing evidence on authentic learning scenario implementation across European higher education. The two-group structure that emerged inductively from the focus group data, method-focused (MF) and technology-focused (TF) scenarios, maps meaningfully onto the TPACK framework (Koehler & Mishra, 2006). MF scenarios primarily challenge educators' pedagogical knowledge: they require sophisticated facilitation skills, curriculum integration, and learner-centred design regardless of which digital tools are used. TF scenarios, by contrast, place greater demands on technological knowledge and infrastructure, digital barriers rather than pedagogical ones represent the primary challenge. This distinction is particularly evident in the contrasting profiles of Flipped Classroom (MF, high pedagogical demand) and Hybrid Teaching (TF, high technical demand).

This distinction has direct implications for professional development design. Support for MF scenarios should prioritise didactic skill-building, structured learning design templates, and curriculum integration strategies. Support for TF scenarios requires attention to technological infrastructure, digital competence development, and institutional investment. The cross-cutting finding that collections of good-practice examples and learning design templates are key enablers across both groups aligns with Rienties et al. (2023) and supports the use of structured learning design tools grounded in constructive alignment principles (Biggs, 2014; Divjak, Grabar et al., 2022).

AI in Education stands out as qualitatively different from all other scenarios: it is the only one discussed primarily from a position of limited experience, making it prospective and normative in character rather than reflective of established practice. The urgency of calls for institutional guidelines and ethical frameworks around AI aligns with Holmes and Tuomi's (2022) analysis of systemic readiness gaps in higher education's response to AI. The finding that AI was also mentioned as a support tool in PBL, Simulation, and VR discussions suggests that AI's role may be as much infrastructural (cutting across scenarios) as scenario-specific.

The prominence of practical examples, learning design templates, and structured problem scenarios as cross-cutting enablers across all five focal scenarios aligns with Rienties et al. (2023) and supports the use of constructive alignment tools in professional development (Biggs, 2014). Institutional context shaped scenario relevance substantially, confirming Herrington et al.'s (2014) argument that authenticity is realised differently across settings. The cross-institutional variation, particularly UU4's distinct scenario profile as a distance-learning institution, and CU1's stronger orientation toward simulation and PBL, confirms that authentic learning design cannot be approached prescriptively: it requires contextual adaptation.

The following limitations should be noted. First, the sample is restricted to four European universities, limiting transferability to other cultural and national contexts. Second, only educator perspectives were captured; student experiences of these scenarios remain unexplored. Third, the qualitative design yields rich descriptions of perceived conditions but cannot establish causal links between scenario use and learning outcomes.

A fourth limitation concerns the temporal context of the data. All focus groups were conducted in 2023, at a moment when generative AI tools such as large language models had only recently entered mainstream educational discourse. Practitioners' perceptions of AI in Education reflect an early adoption phase characterised by uncertainty, experimentation, and limited institutional guidance. Given the pace of development in this area, the specific challenges and enablers identified for AI in Education are likely to have evolved significantly and should be interpreted as a snapshot of practitioner perspectives at a particular moment rather than stable findings. Future studies should revisit this scenario longitudinally to track how growing practitioner familiarity and emerging institutional frameworks reshape its implementation profile.

6 Conclusion

This paper has identified 11 authentic learning scenarios considered applicable across four European higher education contexts (RQ1) and documented their perceived benefits, challenges, and enabling conditions from 95 practitioners' perspectives (RQ2). Scenarios cluster into method-focused and technology-focused groups with distinct profiles of pedagogical and technical demands. Across both groups, institutional support, differentiated professional development, and practical learning design resources emerge as critical enablers.

Detailed analysis of five focal scenarios (Flipped Classroom, AI in Education, Hybrid Teaching, Problem-Based Learning, and Virtual Reality) illustrates both the specific implementation challenges of each scenario and a set of cross-cutting themes: the need for shared institutional norms, structured design support, and adaptive professional development that goes beyond awareness-raising to skill-building in context. Future research should extend this framework to include student perspectives, examine long-term learning outcomes, and track how AI integration reshapes the authentic learning scenario landscape over time.

Acknowledgements.

The study was supported by the project 'Innovating Learning Design in Higher Education' (iLed), financed from the Erasmus+ Program of the European Union (Erasmus+ KA220-HED - Cooperation partnerships in higher education).

Disclosure of Interests.

The authors have no competing interests to declare that are relevant to the content of this article.

References

Biggs, J. (2014). Constructive alignment in university teaching. HERDSA Review of Higher Education, 1, 5–22.

Bragg, L. A., Walsh, C., & Heyeres, M. (2021). Successful design and delivery of online professional development for teachers: A systematic review. Computers & Education, 166, 104158.

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. Educational Researcher, 18(1), 32–42.

Brown, M. G. (2016). Blended instructional practice: A review of the empirical literature. The Internet and Higher Education, 31, 1–10.

Divjak, B., Grabar, D., Svetec, B., & Vondra, P. (2022). Balanced Learning Design Planning. Journal of Information and Organizational Sciences, 46(2). https://doi.org/10.31341/jios.46.2.6

Divjak, B., Rienties, B., Iniesto, F., Vondra, P., & Žižak, M. (2022). Flipped classrooms in higher education during COVID-19. International Journal of Educational Technology in Higher Education, 19(1), 9.

Dochy, F., Segers, M., Van den Bossche, P., & Gijbels, D. (2003). Effects of problem-based learning: A meta-analysis. Learning and Instruction, 13(5), 533–568.

Donovan, M. S., Bransford, J. D., & Pellegrino, J. W. (1999). How People Learn: Bridging Research and Practice. National Academies Press.

Herrington, J., & Oliver, R. (2000). An instructional design framework for authentic learning environments. Educational Technology Research and Development, 48(3), 23–48.

Herrington, J., Reeves, T. C., & Oliver, R. (2014). Authentic learning environments. In Handbook of Research on Educational Communications and Technology (4th ed., pp. 401–412). Springer.

Holmes, W., & Tuomi, I. (2022). State of the art and practice in AI in education. European Journal of Education, 57(4), 542–570.

ICA:UK. (2014). Consensus Workshop Method Overview.

Koehler, M., & Mishra, P. (2006). Technological Pedagogical Content Knowledge. Teachers College Record, 108(6), 1017–1054.

Kukulska-Hulme, A., et al. (2022). Innovating Pedagogy 2022. The Open University.

Lave, J., & Wenger, E. (1991). Situated Learning: Legitimate Peripheral Participation. Cambridge University Press.

Liu, M., et al. (2024). Educational technology research trends. Computers & Education, 210, 104953.

Mayring, P. (2015). Qualitative Inhaltsanalyse (12th ed.). Beltz.

Mayring, P. (2023). Qualitative Content Analysis. In International Encyclopedia of Education (Vol. 12, pp. 313–322). Elsevier.

Nelson, M. J., Voithofer, R., & Cheng, S.-L. (2019). Mediating factors in technology integration. Computers & Education, 128, 330–344.

Rasheed, R. A., Kamsin, A., & Abdullah, N. A. (2020). Challenges in the online component of blended learning. Computers & Education, 144, 103701.

Rienties, B., et al. (2022). Online professional development across institutions. International Journal of Educational Technology in Higher Education, 20(1), 30.

Rienties, B., et al. (2023). Learning design professional development across four European universities. Teaching in Higher Education, 28(4), 789–806.

Sharples, M., et al. (2012). Innovating Pedagogy 2012. Open University Innovation Report 1.

Tondeur, J., et al. (2012). Preparing pre-service teachers to integrate technology. Computers & Education, 59(1), 134–144.

Synergies of learning design and analytics: Design patterns of higher education courses

Hanni Muukkonen, Egle Gedrimiene
University of Oulu
Faculty of Education and Psychology
Pentti Kaiteran katu 1, 90570, Oulu, Finland
{hanni.muukkonen, egle.gedrimiene}@oulu.fi

Blaženka Divjak, Barbi Svetec
University of Zagreb
Faculty of Organization and Informatics
Pavlinska 2, 42000 Varaždin, Croatia
{bdivjak, basvetec}@foi.hr

Julia Schmitt, Michael Eichhorn, Alexander Tillmann
Goethe University Frankfurt
studiumdigitale
Eschersheimer Landstr. 155/157, 60323 Frankfurt am Main, Germany
{schmitt, eichhorn, tillmann}@studiumdigitale.uni-frankfurt.de

Bart Rienties
The Open University
Institute of Educational Technology
Walton Hall, Milton Keynes, MK7 6AA, UK
bart.rienties@open.ac.uk

Abstract. *Synergies of learning design and learning analytics benefit higher education teachers by enabling structured and visualized presentations of expected learning outcomes, student workload, activities, and assessment, thereby facilitating the (re)design of curricula. This study investigated, first, what learning (re)design aims teachers articulated. Second, twenty learning designs were examined with learning analytics methods to identify patterns and structural features. Finally, we investigated how students (N = 580) evaluated course learning design and analytics for supporting their learning. The study suggests future research combining qualitative, micro-level understanding of learning design, tool support, and macro-level pattern analysis for guiding curriculum renewal.*

Keywords. Learning design, Learning analytics, Higher education, Pattern analysis, Student feedback

1 Introduction

Learning design (LD) has been emphasized as one of the key factors for successful learning (Martin et al., 2019). LD follows a student-centered approach on the development of pedagogical processes, based on contemporary sciences of learning and education, and it is the key for providing efficient, effective, and engaging learning experiences (Wasson & Kirschner, 2020). Empirical evidence supports this: large-scale analyses at the Open University, covering over 150 courses, have shown that well-structured LDs, with a balanced mix of collaborative, independent and assessment activities correlate with higher student satisfaction and better pass rates (Rienties & Toetenel, 2016; Nguyen et al., 2022). Systematic LD frameworks such as the '7Cs' (Conole, 2013) or Balanced Design Planning (BDP) (Divjak et al., 2022) offer structured methods for making pedagogical intentions explicit, thereby facilitating iterativeness and improvement. This study contributes to the understanding of how LD developed with the support of learning analytics can provide educators and learners with insights for enhanced learning experiences and outcomes. Learning design (LD) can be defined as the conscious, systematic process of planning, documenting, and sequencing teaching and learning activities (TLA) (Conole, 2013; Lockyer et al., 2013). In contrast to traditional lesson planning, LD combines educational theory with the practical structuring of learning tasks, thus making the intentionality of teaching visible and shareable (Laurillard, 2012). It is based on the assumption that effective learning environments are the result of decisions about learning outcomes, activity types, resources, and assessment strategies (Divjak et al., 2022). Whether through a LD or in class, clarity in planning helps students develop a strong understanding of what is expected and supports their engagement and learning regulation.

Synergies with learning analytics have been suggested to benefit design work (Macfadyen et al., 2020), especially by focusing on enabling visualized LD including learning outcomes, student workload, ensuring constructive alignment and assessment validity. Studies have shown that visual representations of LD, including workload distribution graphs and constructive alignment maps, help educators identify misalignments and overemphasis, leading to targeted course revisions that improve student engagement and performance (Lockyer et al., 2013; Divjak et al., 2023; Nguyen et al., 2022). A recent survey of 53 teachers from four European countries found that user experience feedback is critical to the further development of LD tools, such as the BDP tool (Divjak et al., 2023). Learning outcomes, constructive alignment, and analytics were highlighted as core components of effective LD. The benefits of systematic LD have been recognized as important for course success (Prieto et al., 2014; Drugova et al., 2024) and innovative tools are currently developed. This study investigated teachers' LD with one LD tool. The study contributes to identifying benefits of systematic LD by combining teachers' and students' lived experiences across four countries with pattern analysis of teaching and learning activities designed in the LD tool used by teachers.

2 Background

2.1 Learning design and analytics to support pedagogical development

Technological tools support the development of LDs by simplifying the planning and implementation of learning activities and promoting interaction between teachers and learners. As examples we mention the BDP concept and tool which supports LD through visual analyses of workload distribution, constructive alignment and assessment validity (Divjak et al., 2023) and edCrumble, a (blended) LD authoring tool (Albó et al., 2022). Based on visualized feedback created by analytics functions, teachers are able to identify discrepancies between intended learning outcomes and assessment practices, workload bottlenecks, and course structures to be iteratively refined (Kauppi et al., 2020). Empirical studies have shown that such data-driven, iterative design processes lead to measurable improvements in student performance, satisfaction, and progress rates (Rienties et al., 2015; Divjak et al., 2025).

LD tools incorporate multiple purposes. First, they offer structure and conceptual guidance to plan the design or redesign of a course, by supporting a teacher in specifying learning activities and their sequences (Divjak et al., 2022). Second, they are intended to document design and sequencing of teaching practices, for the purposes of improving understanding, and evaluating TLAs (Lockyer et al., 2013). Third, the created designs can be examined with learning analytics (design analytics) visualizations or feedback for teachers on specified features of the design, such as time reserved for collaboration in the designed unit or means of feedback for students. Fourth, LDs can be collaboratively advanced or evaluated by teacher peers, designers or students (Lakkala et al., 2015).

Transparent practices, such as clearly communicating objectives, assessment criteria, and expectations, have been recognized repeatedly in the literature as fostering higher levels of student achievement. The time and effort instructors invest in planning and organizing their courses correlates strongly with student learning outcomes (Feldman, 2007). The aim of LD is therefore not just to be a descriptive representation of teaching practice, but a strategic design science (Laurillard, 2012), which integrates educational theory, empirical findings, and analytics to systematically improve the quality and equity of higher education curricula.

2.2 Patterns in learning design

Learning paths and patterns are well-known concepts in learning analytics. Commonly learning paths represent sequenced sets of learning activities which are designed to lead learners towards the achievement of the intended learning outcomes (Martínez-Carrascal et al., 2024). A distinction can be made between designed (or ideal) learning paths, which reflect the intended sequence of activities, and enacted learning paths, which represent the actual sequences of learning activities chosen by students during the learning process (Angrave et al., 2020). This study examines learning paths in LD and looks into learning patterns, which are recurring behaviors or structures, indicating common paths, detours, or cycles (Real & Pimentel, 2025).

When we investigate learning patterns, one perspective that can be taken is to analyze patterns typical for innovative pedagogical approaches. Quality LDs are based on innovative pedagogical approaches such as Flipped Classroom (FC), Project-Based Learning (PBL) and Inquiry Learning (IL). For example, FC incorporated in a course LD is likely to provide a typical pattern: pre-class acquisition, formative assessment, active learning, and assessment (Divjak et al., 2024).

While substantial progress has been made in terms of the conceptual development of LD and learning analytics (Law & Liang, 2020), and recent studies have provided empirical evidence of how LD and learning analytics data could be effectively combined to provide crucial insights into learning processes and outcomes (e.g., Rienties & Toetenel, 2016; Nguyen et al., 2022), most of these studies were nested within one institutional context. Furthermore, most of the studies reviewed by Drugova et al. (2024) were unable to close the loop on LD and learning analytics with actual stakeholders. While Albuquerque et al. (2025) found,

in a range of European countries, important similarities in how teachers designed learning activities, to the best of our knowledge no study has looked at the lived experiences with LD of educators and students in a multi-country context.

2.3 The present study

In this study, as a starting point, higher education (HE) teachers from four European countries were invited to participate in workshops organized within the Innovating Learning Design in Higher Education (iLed) project. Teachers had an opportunity to learn about the concept of LD and to practically work with the Balanced Design Planning (BDP) LD tool (Divjak et al., 2022). They designed or redesigned a course or a part of a course using the BDP tool with the workshop support. A subset of the courses was then implemented in HE study programs and students were invited to evaluate and give feedback on the LD for their course.

The study examined first, the aims teachers set for their LD. Subsequently, teachers created LDs, which were analyzed for recurring patterns and emphasis in teachers' design work. We aimed to discover typical forms of learning outcomes (LOs) teachers specified as well as the patterns in teaching and learning activities (TLA) designed. The final research question was related to the analysis of student evaluations on how access to the LD supported their learning.

RQ1: What aims did teachers set for their LD?

RQ2: What levels of learning outcomes were represented in the LDs, and how were TLAs distributed across these levels?

RQ3: What patterns can be identified in the sequencing of learning types across the designed TLAs?

RQ4: How did students evaluate the implemented course learning designs in terms of how well it supported their learning and how useful it was?

3 Methodology

3.1 Participants and context

Teachers from five HE institutions participated in workshops, as part of the Erasmus+ iLed project, designed to enhance the competencies related to digital pedagogies and tools for innovating HE curriculum. The teachers used the BDP tool to create and develop the LD of their course as well as to update the LD based on what they learned from the workshop, peer feedback, and BDP design analytics. Teachers' reflections were collected during the process. Developed LDs (N = 20) from the workshop were collected for analysis.

The (re)designed courses were implemented by HE institutions and student feedback on the LD of the courses was collected. The courses ranged from 1 to 5 ECTS credits in scope and were mostly organized in line with innovative pedagogies, especially flipped classrooms (see Appendix). Teachers of the (re)designed courses introduced to students the course LD created in the BDP tool. Specifically, the teacher introduced to the students the course overview of the TLAs, the design analytics, and provided students with the tool access link so they could check the LD on their own time. At the end of the course implementation of these 13 courses, students were asked to complete a brief online survey and evaluate the course design from the student's perspective. Students' (N = 580) feedback in the form of two Likert-type questions (if the LD overview and analytics were useful to them) and two open-ended questions were collected (the experienced positive and negative aspects of the course designs). Additional nine courses were included in the pattern analysis of LD content without student evaluations (courses 14-22 in Appendix).

Overall, the courses involved a wide variety of students and course set-ups (small group of 8-9 to a large class of 979). The central content of courses was in the studies of medicine, STEM, and education. All teacher participants gave written informed consent to be included in the study and its reporting, and the students also gave consent to participate in the study. An ethical approval was granted by the ethics committee of the University of Zagreb (FOI).

3.2 LD tool and concept

The BDP concept and tool support educators in LD, which is a student-centric development of pedagogical processes focusing on determining learning conditions, activities, and their sequences (Divjak et al., 2022).

The BDP concept and tool are strongly grounded in learning outcomes as a starting point for LD. The concept focuses on ensuring constructive alignment (Biggs, 1996) between learning outcomes, TLAs, and assessment, as well as the corresponding student workload. In the tool, educators can plan their courses from learning outcomes, to topics, units, and teaching and learning activities, including details like workload, learning types, delivery modes, presence of collaboration and group work, and assessment.

Figure 1. The BDP tool analysis function visualization example (left learner workload and right collaboration).

Another specific feature refers to the learning types that a teacher designs for their course. These TLAs are marked as Acquisition, Discussion, Investigation, Practice, Production, or Assessment (Divjak et al., 2022). The tool also provides design analytics that can support educators in ensuring constructive alignment and developing LDs in line with the intended pedagogical approach (Fig. 1).

3.3 Data analysis

To address RQ1, teacher reflections on their aims for (re)design of courses were thematically reviewed for descriptive purposes. This involved identifying teachers' redesign ideas and key aims of course update described in the brief written reflection notes.

To respond to RQ2, we analyzed the distribution of LO levels (based on 176 LOs) and the distribution of TLA learning types among LO levels (based on 1624 TLAs) in the 22 LDs (including a total of 161 LOs, 1477 TLAs), using the tidyverse packages in R. Note that for analyses which required links between LOs and learning types we excluded the *Assessment* learning type because it is horizontal and can be linked to several LOs. While the overall sample included 1624 TLAs, this subsample included 1150 TLAs.

A Kendall's Tau correlation was performed to determine any correlations between LO levels, TLAs, and specific types of engagement included in the courses (teacher presence, group work, feedback, collaboration, assessment, and delivery mode). We analysed the subset of 1150 TLAs, using the vcd, vcdExtra, rcompanion, ggcorrplot2, and DescTools packages in R.

To respond to RQ3, we examined the sequencing of TLAs within the LDs looking for regularities among samples of sequentially ordered data (Bechini et al., 2023). Particularly, we conducted **sequential pattern mining** (Agrawal & Srikant, 1995) to identify the most frequent patterns in learning type occurrence, using the arulesSequences library in R. Sequential pattern mining belongs to the process mining family, aiming to identify sequences in a dataset, which appear statistically significant number of times to verify interesting recurrences. It is focused on finding regularities among samples of sequentially ordered data (Bechini et al., 2023). It has been used in learning analytics research to identify frequent learning paths and patterns. For this study we used the SPADE algorithm. Support reflects the frequency of a sequence, confidence its predictive reliability, and lift the strength of the association relative to chance (lift > 1 indicating a meaningful association) (Zaki, 2001).

Finally, to respond to RQ4, descriptive statistics was extracted from student feedback on their course LDs and experiences with the tool.

4 Results

Regarding RQ1, based on teachers' reflections, most of the courses in the study were redesigned, instead of being planned from scratch. The updates were in the form of fine-tuning and small changes in already existing TLAs and pedagogies to better reflect student learning needs and course realities, e.g., availability of AI for students or an increase in the number of students in the course. Teachers' redesign ideas were based on students' feedback and teachers' experiences in previous course implementation as well as collective observations in the project workshops. The most frequent aims of course update were as follows:

1. To improve the clarity and structure of the LD of the course.
2. To better align TLAs with LOs.

3. To better plan and distribute student workload.
4. To include more formative assessments (e.g., in seminars).
5. To include AI in some of the tasks for problem solving (e.g., an essay).

Many teachers also emphasized that they continually redesign, update and overwise improve LDs in their courses based on student feedback and curriculum development.

To address RQ2, the frequency of levels of LOs, following a revised version of Bloom's taxonomy proposed by Krathwohl (2002), was analyzed with LO levels presenting remembering (1), understanding (2), applying (3), analyzing (4), evaluating (5), and creating (6). Based on the LD data for all courses (N = 22), a majority of LOs were described through concepts related to applying knowledge (42%), followed by outcomes related to understanding (20%) and analyzing (16%) phenomena under study. The lowest numbers of LOs related to remembering (2%), while evaluating and creating were each present in c. 10% of the data.

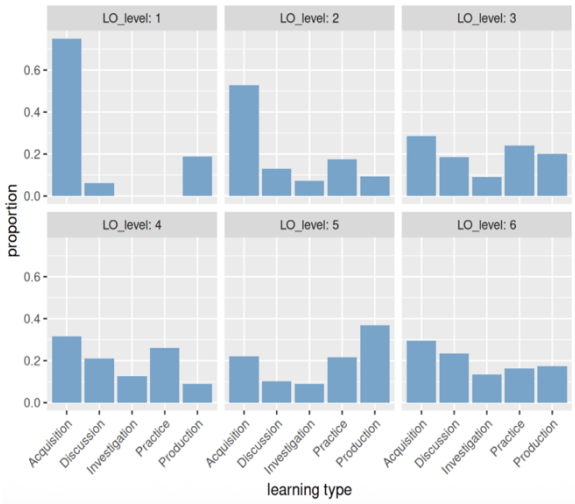


Figure 2. A comparison of learning outcome level (LO_level: 1-6) and learning types present in the learning designs.

As mentioned already above, the learning types describe the TLAs that the teachers had designed and labelled in the LD tool for their courses. Fig. 2 shows how the learning types were related to the LO levels. Acquisition was highly present in TLAs related to the LOs at the levels of *remembering* and *understanding*, while less prominent on higher LO levels. Discussion and Production were present in all LO levels, which highlights that teachers granted value to these activities in all types of designs. Production was especially related to the LO level emphasizing *evaluating*. Investigation was the least included type of learning activity. Except for the lowest LO level, that is *remembering*, Practice was also present in all types of designs, with most emphasis on LO levels *applying*, *analyzing*, and *evaluating*.

The correlation analysis shown in Fig. 3 presents Kendall's τ correlation coefficients in the lower triangle and their statistical significance in the upper triangle ($p < .05$, $p < .01$, $p < .001$). Most coefficients were weak, with only a small number of moderate to strong positive associations observed between pedagogically related variables. There are strong, significant correlations between teacher being present and feedback, teacher being present and synchronous delivery, feedback and synchronous delivery. This indicates that feedback for students was mostly planned to be provided by a teacher in synchronous teaching sessions. However, feedback to students did not correlate with the level of LO, learning type, or assessment. It should be noted that learning type is a categorial variable and for more meaningful insights it would be appropriate to analyze each learning type separately.

Figure 3. Correlations (Kendall) between learning outcome levels, learning types, and various types of engagement.

Regarding RQ3, pattern mining revealed the common sequences and patterns (Table 1). Altogether, 32 sequences with frequency equal to or greater than 2 were identified.

One of the most prominent sequences of learning types in the LD sample turned out to be Acquisition + Assessment + Practice → Practice. This reflects a possible strong emphasis on the flipped classroom approach.

Table 1. Subsequences for consecutive learning types in the learning designs.

rule	support	confidence	lift
<{Practice},{Production}> => <{Production}>	0.0360360	0.5333333	1.4095238
<{Production},{Production}> => <{Production}>	0.0855856	0.5757576	1.5216450
<{Assessment},{Practice}> => <{Practice}>	0.0540541	0.8571429	1.6989796
<{Acquisition},{Assessment},{Practice}> => <{Practice}>	0.0540541	0.9230769	1.8296703
<{Investigation},{Acquisition}> => <{Practice}>	0.0405405	0.5000000	0.9910714
<{Practice},{Investigation},{Acquisition}> => <{Practice}>	0.0405405	1.0000000	1.9821429
<{Investigation},{Acquisition},{Practice}> => <{Assessment}>	0.0360360	0.8888889	2.2424242
<{Practice},{Investigation},{Acquisition},{Practice}> => <{Assessment}>	0.0360360	0.8888889	2.2424242
<{Acquisition},{Practice},{Practice}> => <{Assessment}>	0.0405405	0.6428571	1.6217532
<{Acquisition},{Acquisition},{Acquisition}> => <{Assessment}>	0.0360360	0.5714286	1.4415584
<{Practice},{Investigation}> => <{Acquisition}>	0.0405405	0.6923077	1.1057001

The sequence of Practice + Investigation + Acquisition → Practice has the strongest confidence level (1) and it occurs in about 4% of LDs (0.0405405), suggesting that whenever this sequence occurred, it was consistently followed by Practice. It indicates that problem-based and inquiry-based learning, as examples of modern pedagogical approaches, are present in the analyzed LD sample. This is expected, since the aim of the project exercise was specifically to redesign courses in a way that reflects innovative pedagogies. Similarly, the sequence of Investigation + Acquisition + Practice → Assessment, which also has a high confidence level (0.8888889) but a bit lower support (0.0360360) is consistent with inquiry-based learning as well. Some other patterns, although with weaker metrics (support, confidence), are also consistent with constructivist approaches and active learning, e.g., Practice + Investigation → Acquisition.

The results also show that sequences most frequently end with Practice, Assessment, or Production. While assessment is a traditional end of a learning sequence, sequences that end with practice and production indicate that students often learn through hands-on experiences, which again points to active learning practices.

Surprisingly, while Discussion activities were frequent, they did not appear as part of the most prominent patterns.

Regarding RQ4, 543 students answered the question about the usefulness of LD analytics and visualizations. Students were asked to provide their feedback on the usefulness of the two learning design parts in the BDP tool: a general overview of the teaching and learning activities and analytics. Fig. 4 shows the distribution of responses: 56% agreed that the overview was useful, 23% did not disagree or agree, 6% did not find it useful, and an additional 15% had not checked it afterwards.

Although analytics and visualizations provided very different information compared to general teacher and learning activities in the overview of the course design, students' evaluations of their usefulness were almost identical. From all the students, 55% agreed the analytics part was useful, while 23% remained undecided. At the same time, 7% of students indicated that analytics and visualizations were less useful to them, and 15% had not checked the tool afterwards.

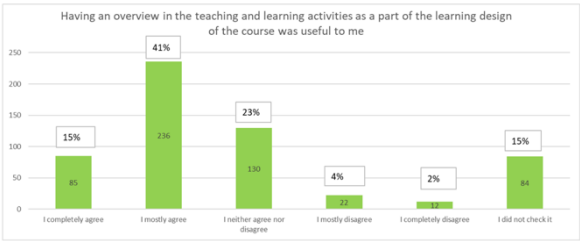


Figure 4. Distribution of students' responses to overview of the teaching and learning activities. "Having an overview of the teaching and learning activities as a part of the LD of the course was useful to me".

In their open answers, students typically highlighted that the design shown with the tool gives the learner a better idea of what they are supposed to learn. They expressed liking and finding it useful to know which lessons were dedicated to engaging with particular learning outcomes. Students also commented on how they perceived the LD helpful and useful for planning the workload, e.g., "It also helps to prepare for the course and understand what is expected" (Student 2).

5 Discussion and Limitations

RQ1 captured which aims teachers had for redesigning their courses. The continual need to redesign, update, and otherwise improve LDs was evident. These align with the review findings of Drugova et al. (2024) that the areas of LD improvements are often based on monitoring of learning to detect problems and addressing these problems with redesign. Particularly teachers were raising the questions of how to include AI-mediated teaching and learning activities in meaningful and ethical ways for problem solving.

The analysis of LOs and learning types (RQ2) showed that while teachers planned a variety of learning activities, knowledge acquisition was dominant, suggesting that more traditional teaching methods were preferred. Simultaneously, the results also reflect a typical flipped classroom design, where for a large part students study independently and are provided with feedback in synchronous classes led by teachers. Our correlation analyses further showed that activities involving teacher presence, feedback, and synchronous engagement were interlinked, emphasizing the importance teachers gave for real-time interaction for formative assessment and learner support.

With the pattern mining (RQ3), we observed an evident emphasis on practice and production signifying that teachers valued active learning and planned opportunities for students to develop their skills. The dominant sequence of learning types observed in the data (Acquisition, Assessment, Practice, Practice) exemplifies the characteristic structure of flipped classroom setups, highlighting that students spend substantial class time practicing knowledge and skills they have acquired independently through preparatory materials. The learning type Discussion was unexpectedly not part of most prominent patterns, possibly reflecting a less structured or explicated use of discussion as activity contributing to learning outcomes. This may reflect a lesser emphasis on facilitation and discourse as aspects included in a LD when designing with a LD tool. This calls for further identification of how Discussion is used in course LDs.

Student evaluations of the learning design (RQ4) enabled us to examine how students valued the information on learning outcomes and the teaching and learning activities for the course. While not everyone took up this opportunity, the majority of students evaluated that the course design and design analytics supported their learning. The results closely align with Feldman's (2007) findings, which emphasize the critical role of teacher preparation and course organization in promoting student engagement and success.

The effectiveness of the patterns found and positive evaluations of the learning designs by students are empirically supported by meta-analyses. van Alten et al. (2019) found that flipped classroom models are particularly effective when classroom time is not reduced, allowing students to engage with content independently before class and then use in-person sessions exclusively for collaborative work and applying knowledge. Their meta-analysis demonstrated medium to large effect sizes in student learning outcomes compared to traditional teaching formats.

The analysis offers suggestions for continued development of LD tools. For example, what would be the typical sequences for flipped learning or inquiry learning which could be benchmarked by teachers in need of on-demand, AI-supported guidance while creating new types of LDs. Further observing and detailing high quality designs (e.g., Lee et al., 2017) and how they are portrayed at system level through the analysis of LO and learning types offers future means to scaffold learning design.

The limitations primarily concern the variability in the quality of LDs, which stems from differences in teachers' pedagogical experience and training. While this variability poses challenges for interpretation, the findings nonetheless reflect the reality of HE, where a wide diversity of teaching qualifications and practices is common. Another source of variability was that not all courses were of the Flipped classroom pedagogy, future studies could identify patterns related to specific pedagogies. Overall, this study included a relatively small number of learning designs with variability of courses (FC, PBL, and IL) from a project-specific sample. This poses limitations on generalizability, but offers a synthesizing view on the variability existing in any HE curriculum development endeavor.

6 Conclusions

The study offers evidence on how teachers understand learning design, e.g., the interplay of learning outcomes and teaching and learning activities as well as how students benefit from LD and design analytics. It points to a continued need for constructive alignment of learning outcomes with teaching and learning activities, especially for more advanced levels of learning outcomes in higher education. It expands on how tools supporting design and analytics can make course learning design more explicit and tangible for collaborative design and professional development. This case study supports that if teachers undergo systemic training in learning design, innovative approaches (e.g., flipped classroom) can be detected by design analytics. Furthermore, the use of visualizations and design analytics may contribute to the quality of learning design and help teachers to resolve design errors in real time, which is in line with prior research (Albó et al, 2022; Schmitz et al., 2017).

The central features of the BDP tool, the general overview of the teaching and learning activities and the design analytics, offered useful information also for students enrolled in those courses. The value was generated by support for own planning, evaluating the workload, and understanding what is expected. These are aspects related to improved regulation of own learning, through planning, monitoring, and reflection. The results are consistent with empirical studies that provide robust evidence of the significant improvement in student learning outcomes that can be achieved through transparent teaching and course structure (Ambrose et al., 2010).

Transparent practices, such as clearly communicating objectives, assessment criteria, and expectations, have been recognized repeatedly in the literature as fostering higher levels of student

achievement (Feldman, 2007). The benefits of better planning, clearer objectives, and more effective self-regulation, are not only noted by users, as documented in our study, but are also supported by robust evidence linking transparency with increased learner achievement (Feldman, 2007). Therefore, the positive results of the last research question are likely to be enhanced by the transparent communication embedded in the course designs. This matches international research findings postulating that increased learning success is related to transparency in higher education (e.g., Rienties et al., 2015).

When students are provided with explicit information about course processes, goals, and evaluation procedures, they report feeling more empowered and experience fewer misunderstandings, which may contribute to better outcomes (Ambrose et al., 2010; Feldman, 2007). Hattie (2013) also emphasizes that teacher clarity is a critical factor for effective learning and achievement. The structure and visualizations of the BDP tool support teaching practices and were also perceived as helpful by the students in our study. Qualitative feedback underscored the helpfulness of knowing lesson objectives and activity distribution. Nevertheless, not all students reported engaging with the overview or analytics, revealing potential gaps in user uptake or perceived relevance. Addressing these gaps could encourage data-informed decision-making.

Further research efforts are needed to examine patterns from both a micro-level perspective of the LDs and the macro-level pattern generation. This will contribute to the understanding of teachers' strategic planning of learning and teaching, design practices, and alignment of curriculum development involving multiple levels of learning outcomes.

References

Agrawal, R. & Srikant, R. (1995). Agrawal, R., & Srikant, R. (1995). Mining sequential patterns. *Proceedings of the Eleventh International Conference on Data Engineering*, 3-14. https://api.semanticscholar.org/CorpusID:12242182

van Alten, D. C. D., Phielix, C., Janssen, J. J. H. M., & Kester, L. (2019). Effects of flipping the classroom on learning outcomes and satisfaction: A meta-analysis. *Educational Research Review, 28*, 100281. https://doi.org/10.1016/j.edurev.2019.05.003

Albó, L., Barria-Pineda, J., Brusilovsky, P., & Hernández-Leo, D. (2022). Knowledge-based design analytics for authoring courses with smart learning content. *International Journal of Artificial Intelligence in Education, 32*, 4–27. https://doi.org/10.1007/s40593-021-00253-3

Albuquerque, F., Rienties, B., & Divjak, B. (2025). Decoding learning design decisions: A cluster analysis of 12,749 teaching and learning activities. *LAK '25: Proceedings of the 15th International Learning Analytics and Knowledge Conference (pp. 407 - 417). ACM.* https://dl.acm.org/doi/10.1145/3706468.3706520

Ambrose, S. A., Bridges, M. W., DiPietro, M., Lovett, M. C., & Norman, M. K. (2010). *How learning works: Seven research-based principles for smart teaching*. Jossey-Bass.

Angrave, l., Zhang, Z., Henricks, G., & Mahipal, C. (2020). Who benefits? Positive learner outcomes from behavioral analytics of online lecture video viewing using ClassTranscribe. In *The 51st ACM Technical Symposium on Computer Science Education* (SIGCSE '20) (pp. 1193 – 1199). ACM. https://doi.org/10.1145/3328778.3366953

Bechini, A., Bondielli, A., Dell'Oglio, P., & Marcelloni, F. (2023). From basic approaches to novel challenges and applications in Sequential Pattern Mining. *Applied Computing and Intelligence, 3*(1), 44-78.

Biggs, J.B. (1996). Enhancing teaching through constructive alignment. *Higher Education, 32*, 347–364.

Conole, G. (2013). *The 7Cs of Learning Design*. https://opennetworkedlearning.wordpress.com/wp-content/uploads/2015/05/the-7cs-of-learning-design.pdf

Divjak, B., Grabar, D., Svetec, B., & Vondra, P. (2022). Balanced Learning Design Planning: Concept and tool. *Journal of Information and Organizational Sciences, 46*(2), 361-375 DOI: https://doi.org/10.31341/jios.46.2.6.

Divjak, B., Badari, J., Grabar, D., Horvat, D., Vondra, P., & Rienties, B. (2023). Enhancing learning design through user experience research: Insights from a survey in four European countries. *Proceedings of the Central European Conference on Information and Intelligent Systems* (pp. 213–221). https://oro.open.ac.uk/99852/

Divjak, B., Rienties, B., Svetec, B., Vondra, P., & Žižak, M. (2024). Reviewing assessment in online and blended flipped classroom. *Journal of University Teaching and Learning Practice, 21*, 7-26. doi: 10.53761/wfmhvt17

Divjak, B., Barthakur, A., Kovanović, V., & Svetec, B. (2025). The impact of learning design on the mastery of learning outcomes in higher education. LAK'25: *Proceedings of the 15th International Learning Analytics and Knowledge Conference* (pp. 726-737). ACM. https://researchoutputs.unisa.edu.au/11541.2/42157

Drugova, E., Zhuravleva, I., Zakharova, U., & Latipov, A. (2024). Learning analytics driven improvements in learning design in higher education: A systematic literature review. *Journal of Computer Assisted Learning, 40*(2), 510–524. https://doi.org/https://doi.org/10.1111/jcal.12894

Feldman, K.A. (2007). Identifying exemplary teachers and teaching: Evidence from student ratings. In Perry, R.P., Smart, J.C. (Eds.) *The Scholarship of Teaching and Learning in Higher Education: An Evidence-Based Perspective*. Springer. https://doi.org/10.1007/1-4020-5742-3_5

Hattie, J. (2013). *Visible learning for teachers: Maximizing impact on learning*. Routledge.

Kauppi, S., Muukkonen, H., Suorsa, T., & Takala, M. (2020). I still miss human contact, but this is more flexible – Paradoxes in virtual learning interaction and multidisciplinary collaboration. *British Journal of Educational Technology, 5*(4), 1101-1116. https://doi.org/10.1111/bjet.12929

Krathwohl, D.R. (2002). A revision of Bloom's Taxonomy: An overview. *Theory Into Practice, 41*(4), 212–218. http://www.jstor.org/stable/1477405

Lakkala, M., Toom, A., Ilomäki, L., & Muukkonen, H. (2015). Re-designing university courses to support collaborative knowledge creation practices. *Australasian Journal of Educational Technology, 31*(5), 521-536.

Laurillard, D. (2012). *Teaching as a design science: Building pedagogical patterns for learning and technology*. Taylor & Francis. https://www.taylorfrancis.com/books/mono/10.4324/9780203125083

Law, N., & Liang, L. (2020). A multilevel framework and method for learning analytics integrated learning design. *Journal of Learning Analytics, 7*(3), 98-117. https://doi.org/10.18608/jla.2020.73.8

Lee, J., Lim, C. & Kim, H. (2017). Development of an instructional design model for flipped learning in higher education. *Education Technology Research and Development, 65*, 427–453. https://doi.org/10.1007/s11423-016-9502-1

Lockyer, L., Heathcote, E., & Dawson, S. (2013). Informing pedagogical action: aligning learning analytics with learning design. *American Behavioral Scientist, 57*(10), 1439-1459. DOI: 10.1177/0002764213479367

Macfadyen, L. P., Lockyer, L., & Rienties, B. (2020). Learning design and learning analytics: Snapshot 2020. *Journal of Learning Analytics, 7*(3), 6-12. https://doi.org/10.18608/jla.2020.73.2

Martin, F., Ritzhaupt, A., Kumar, S., & Budhrani, K. (2019). Award-winning faculty online teaching practices: Course design, assessment and evaluation, and facilitation. *The Internet and Higher Education, 42*, 34-43. https://doi.org/10.1016/j.iheduc.2019.04.001

Martínez-Carrascal, J. A., Munoz-Gama, J., & Sancho-Vinuesa, T. (2024). Evaluation of recommended learning paths using process mining and log skeletons: Conceptualization and insight into an online mathematics course. *IEEE Transactions on Learning Technologies, 17*, 555-568. doi: 10.1109/TLT.2023.3298035

Nguyen, Q., Whitelock, D., & Rienties, B. (2022). Informing learning design in online education using learning analytics of student engagement. In B. Rienties, R. Hampel, E. Scanlon, & D. Whitelock (Eds.), *Open World Learning: Research, Innovation and the Challenges of High-Quality Education* (pp. 189-207). London: Routledge

Prieto, L.P., Asensio-Pérez, J.I., Muñoz-Cristóbal, J.A., Jorrín-Abellán, I.M., Dimitriadis, Y., & Gómez-Sánchez, E. (2014). Supporting orchestration of CSCL scenarios in web-based distributed learning environments. *Computers & Education, 73*, 9-25. https://doi.org/10.1016/j.compedu.2013.12.008.

Real, E., & Pimentel, E. (2025). An educational process mining model on students' paths data from virtual learning environments. *Technology, Knowledge and Learning, 30*, 2281–2306. https://doi.org/10.1007/s10758-025-09824-y

Rienties, B., Toetenel, L., & Bryan, A. (2015). "Scaling up" learning design: impact of learning design activities on LMS behavior and performance. *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge* (pp. 315–319). ACM.

Rienties, B. & Toetenel, L. (2016). The impact of learning design on student behaviour, satisfaction and performance: A cross-institutional comparison across 151 modules. *Computers in Human Behavior, 60*, 333-341. 10.1016/j.chb.2016.02.074

Rienties, B., Divjak, B., Eichhorn, M., Iniesto, F., Saunders-Smits, G., Svetec, B., Tillmann, A., Zizak, M. (2023a). Online professional development across institutions and borders. *International Journal of Educational Technology in Higher Education, 20*, 1 – 1.

Rienties, B., Balaban, I., Divjak, B., Grabar, D., Svetec, B., Vondra, P. (2023b). Applying and translating learning design and analytics approaches across borders. In Viberg, O., & Grönlund, Å. (Eds.) *Practicable Learning Analytics. Advances in Analytics for Learning and Teaching*. Springer, Cham. https://doi.org/10.1007/978-3-031-27646-0_3

Schmitz, M., Van Limbeek, E., Greller, W., Sloep, P., & Drachsler, H. (2017). Opportunities and challenges in using learning analytics in learning design. In É. Lavoué, H. Drachsler, K. Verbert, J. Broisin, & M. Pérez-Sanagustín (Eds.), *Data Driven Approaches in Digital Education. Proceedings of the 12th EC-TEL* (pp. 209-223). Springer. https://doi.org/10.1007/978-3-319-66610-5_16

Wasson, B., & Kirschner, P.A. (2020). Learning design: European approaches. *TechTrends*, *64*, 815–827. https://doi.org/10.1007/s11528-020-00498-0

Zaki, M.J. (2001). SPADE: An efficient algorithm for mining frequent sequences. *Machine learning*, *42*(1), 31-60.

Acknowledgments

This study has been conducted with the support of the iLed project (Erasmus+) financed by the European Union, 2022–1-HR01-KA220-HED-000085818) and TRUELA project (financed by the Croatian Science Foundation, IP-2022–10-2854).

APPENDIX

Table A. Overview of the courses and participants. Flipped Classroom (FC), Project-Based Learning (PBL), and Inquiry Learning (IL).

(Re)designed course and key content	N of students in the course	N of student responses	ECTS	N of teachers in the course	Course delivery	Compulsory /Elective	Pedagogy
1. Information systems	115	88	5	2	Blended, asynch.		FC
2. IT projects	101	17	3	1+1asis.	Blended, asynch.	Elective	PBL
3. Mathematics	415	131	5	3+6asis.	Bended	Compulsory	FC
4. Business	351	219	4	2+2asis.	Blended	Compulsory	FC
5. Education	32	20	5	2+1asis.	Online, synch.	Compulsory	FC
6. Mathematics	38	15	4	1+1asis.	Blended	Compulsory	FC
7. Data analysis	30	18	4	1+1asis.	Blended	Compulsory	FC
8. Medicine	11	11	1	6	Blended	Elective	FC
9. Medicine	15	15	1,5	2	Blended	Elective	FC
10. Medicine	35	11	1	5	Online, synch.	Compulsory	FC
11. Education	24	24	5	2	Blended, synch.	Compulsory	IL
12. E-learning	8	6	-	2	Online	Elective	FC
13. Education	12	5	1	1	Online	Compulsory	FC
14. Medicine	30	-	4				
15. Education	20	-	5				
16. Management	90	-	5				
17. Management	30	-	1,5				
18. Medicine	979	-	NA				
19. Education	9	-	NA				
20. Education	26	-	NA				
21. Management	320	-	NA				
22. Social studies	NA	-	NA				
Total	2691	580					

Strategic Planning and Decision Making

Matea Matic Sosic
Social Influence and e-WOM in Digital Environments: Examining Purchase and Recommendation Intentions through Social Media

Bisera Andrić Gušavac, Mihailo Vučeljić, Milena Popović, Biljana Panić, Dragan Pamučar
Application of Business Analytics in Demand Planning: An Example from the Fast Food Industry

Aleksandar-Jon Stefan, Artur Bjelica
Harnessing friction topology for stage-sensitive checkoutdiagnostics in supplement e-commerce

Klara Šilec, Violeta Vidaček-Hainš, Zrinka Lacković Vincek
Virtual Collaboration in Higher Education: Economic and Competence Implications

Milanka Dimovska, Violeta Cvetkoska
Artificial Intelligence in the Workplace: A Threat or an Opportunity for Better Performance? Insights from The Republic of North Macedonia

Predrag Trpeski, Violeta Cvetkoska, Filip Peovski, Marijana Cvetanoska, Gunter Merdzan, Kristijan Kozheski
Firms' Workforce Planning Orientations Under Structural Skill Shortages in North Macedonia

Marija Gombar, Marija Boban
Not All Technostress Is Equal: Techno-Complexity, Perceived Ease of Use, and Developers' Acceptance of Emerging Development Technologies

Peter Marton
GIS Support for Managerial Decision-Making in Enhancing Cycling Tourism Routes

Krunoslav Matešić, Mirko Palić, Sabina Sačer
Age Differences In The Effectiveness Of Personality-tailored Ads In Social Networks: Evidence from Croatia

Violeta Cvetkoska, Igor Ivanovski, Filip Peovski, Bojan Kitanovikj, Ivona Serafimovska
Circular Migration without Return Infrastructure: Policy and Social-Partner Sensemaking of Skill Shortages, Brain Drain and Labour-Market Governance in a Developing Context

Katarina Pažur Aničić, Vjeran Bušelić
Career Paths of IT Project Managers: Exploring Generic Competencies and Progression Patterns

Petra Vondra, Nina Begičević Ređep, Tomislav Peharda
A Conceptual Framework for AI-Supported Decision Making in Pre-Incubation Programmes

Social Influence and e-WOM in Digital Environments: Examining Purchase and Recommendation Intentions through Social Media

Marija Bečić, Katija Vojvodić, Matea Matić Šošić
University of Dubrovnik, Faculty of Economics and Business
Lapadska obala Nojka Marinovića 7, 20000 Dubrovnik, Croatia
{marija.becic, katija.vojvodic, matea.matic}@unidu.hr

***Abstract**. In contemporary digital environments, consumer decision-making is increasingly influenced by interactive information flows, social influence, and peer-generated content. Using factor analysis and a seemingly unrelated bivariate probit model, this study examines how social influence and electronic word of mouth (e-WOM) affect purchase and recommendation intentions in social media contexts. Based on a purposive sample of 477 respondents, four factors were identified: Recognition, Interaction, Involvement, and Participation. The results indicate that brand recognition, interaction, engagement, and e-WOM have a statistically significant and positive effect on purchase intention, while women and members of Generation Z exhibit a significantly higher propensity to make recommendations.*

Keywords. social media, purchase intention, recommendation intention, e-WOM, social influence, digital environments

1 Introduction

Social media have evolved from communication channels into complex digital environments where users seek information, compare alternatives, develop perceptions, and make purchase decisions through ongoing interactions with brands and other users. In these contexts, digital platforms operate not only as promotional media but also as socio-technical systems that build trust, reduce uncertainty, and influence behavioural outcomes through information exchange, interaction, and peer visibility (Ma, 2024). This perspective is particularly relevant in the broader context of information systems and digital transformation research, where decision-making increasingly takes place in interactive and data-rich environments. Consequently, social media can be understood as digital ecosystems in which information, communication, and social influence jointly guide behaviour.

Previous research indicates that social media significantly transforms brand-related processes by enabling users to participate in the creation, circulation, and evaluation of brand meaning. In digital environments, users are no longer simply passive recipients of brand communication but active participants in the co-creation of brand meanings and perceptions, with important implications for brand management and consumer behaviour (Gensler et al., 2013). At the same time, social media content quality and brand interactivity have been shown to strengthen brand awareness and purchase intention, suggesting that digital platforms influence both cognitive and behavioural responses, and that brand recognition in online environments is formed not only by firm-generated messages but also by the structure and quality of interaction itself (Dabbous & Barakat, 2020). The literature also highlights the importance of interaction, engagement, informational value, and peer-generated influence in explaining behavioural intentions on social media. Higher information quality and brand familiarity are associated with greater involvement in a brand's social media presence and, consequently, higher purchase intention (McClure & Seock, 2020).

User engagement in social media environments has been understood as a multidimensional construct with cognitive, emotional, and behavioural consequences, reinforcing the centrality of user interaction in digital consumer behaviour (Lim et al., 2022). Additionally, users often rely on recommendations, comments, and shared experiences from others when evaluating products, particularly under conditions of uncertainty, and the quality, credibility, usefulness, and adoption of e-WOM information have all been shown to significantly affect purchase intentions (Erkan & Evans, 2016). Overall, these findings indicate that social media function as interactive influence systems in which engagement, involvement, and peer communication contribute to purchase-related decision-making.

Although previous social media research has primarily focused on purchase intention as the main

behavioural outcome, recommendation-related responses have more often been captured indirectly through constructs such as e-WOM, brand advocacy, and brand-support behaviour, rather than examined alongside purchase intention within a common analytical framework. However, the broader engagement and advocacy literature suggests that both intentions may be driven by overlapping mechanisms, including interaction, social influence, and affective brand-related responses. This creates a gap in understanding their joint formation in digital environments (Barger et al., 2016; Choi et al., 2021). The main aim of this paper is to examine how social media-related factors influence purchase intention and recommendation intention among members of Generations Y and Z, with particular emphasis on the role of brand-related perceptions, user interaction, informational involvement, and participation through e-WOM in the digital environment. The paper contributes to the literature by offering an integrated perspective on social media use, combining brand-related, interactional, informational, and participatory elements to explain purchase intention and recommendation intention. It also makes a methodological contribution by applying a seemingly unrelated bivariate probit model, enabling a more nuanced understanding of the interdependence between these two behavioural outcomes.

2 Theoretical Background and Literature Review

Consumers consistently encounter the problem of information asymmetry (Akerlof, 1970), which Dimoka, Hong and Pavlou (2012) extended to online markets demonstrating how consumers in digital environments experience uncertainty about product quality and how various signals, such as descriptions and ratings, reduce this uncertainty. Consequently, consumer decisions in digital environments rely heavily on indirect information signals (Spence, 1973; Dong, 2022) where individuals use observable quality indicators to reduce uncertainty among less informed market participants. On modern digital platforms, these signals include online reviews, ratings, and e.WOM content, which serve as key mechanisms for conveying information about product quality. Although such mechanisms significantly reduce uncertainty, they do not fully resolve the problem of information asymmetry, which persists in digital environments (Dong, 2022; Dimoka, Hong & Pavlou, 2012). To some extent, it can be mitigated by reputation systems and social signals that influence perceptions of quality and consumer trust.

Social media research increasingly regards digital platforms not just as communication tools, but as interactive marketing and information environments where users co-create value, engage with brands, and translate exposure into marketplace responses. Recent syntheses indicate that social media marketing has adopted a more integrated approach, in which engagement, persuasion, and sales outcomes are jointly influenced by content design, user participation, and platform-specific dynamics. In particular, meta-analytic evidence shows that brands owned social media activity affects not only engagement but also sales-related outcomes, challenging the earlier assumption that social media primarily generate visibility rather than conversion-oriented effects (Dwivedi et al., 2021; Liadeli et al., 2023).

Within this broader context, social media have transformed brand-related processes by enabling consumers to participate in the creation, circulation, and reinterpretation of brand meaning, rather than passively receiving firm-generated communication (Gensler et al., 2013). This shift is reflected in empirical evidence showing that perceived social media marketing elements influence consumer-brand engagement and brand knowledge, while content quality and interactivity enhance brand awareness and purchase intention (Cheung et al., 2020; Dabbous & Barakat, 2020). The effectiveness of social media also depends greatly on the extent to which users interact with content and actively engage with brands and other users. In line with this view, studies of platform communication show that post characteristics such as vividness, user-generated content, and presentation cues significantly influence likes, shares, comments, and other forms of online participation (Xu et al., 2023). Together, these insights suggest that behavioural outcomes in social media environments are strongly influenced by the intensity and quality of users' active engagement within the platform.

Informational value is equally important in explaining behavioural intentions in social media environments. Users evaluate content not only for its attractiveness but also for its usefulness, relevance, and ability to support product evaluation and decision-making. Evidence suggests that information quality and brand familiarity strengthen users' involvement with a brand's social media presence and subsequently contribute to more favourable attitudes and stronger purchase intention (McClure & Seock, 2020). Similarly, social media interactions have been shown to affect perceived content quality and purchase intention, indicating that informational relevance and quality are central to the formation of more deliberate users' responses in digital environments (Onofrei et al., 2022).

More recent evidence confirms that information credibility and information adoption remain central pathways through which e-WOM influences online purchase decisions, reinforcing the view that peer-generated communication is not a peripheral feature of social media use but a core mechanism of digital influence and decision support (Ngo et al., 2024). Similarly, recent work suggests that e-WOM continues to evolve as a major force in digital consumer behaviour, with growing importance for how users

interpret, evaluate, and act on online information in social media settings (Liu et al., 2024).

Younger users' cohorts are particularly relevant to research on social media-related behavioural intentions, as their product evaluation and decision-making processes are deeply embedded in digital and platform-based environments. Generation Z, in particular, is widely characterised as a digitally native segment whose responses to social media marketing are strongly influenced by interaction, customisation, trendiness, and e-WOM. Recent studies show that these factors significantly affect purchase intention and related responses among Generation Z users, either directly or through mediating mechanisms such as perceived quality and perceived value (Deepika & Massand, 2025; An & Ngo, 2025). Comparative evidence also indicates that Millennials and Generation Z constitute two major social-media-oriented market segments but may differ in how they process social media content and translate it into behavioural intentions (Ong et al., 2024). This makes younger cohorts especially important for understanding how social media influence both evaluative and behavioural users' responses.

Although the literature provides substantial evidence on the links between social media marketing, engagement, e-WOM, and purchase intention, recommendation-oriented outcomes are still more often addressed indirectly through related constructs such as advocacy and supportive brand behaviour, rather than examined alongside purchase intention within the same analytical framework. Research on brand advocacy shows that credible content delivery, co-creation, and responsiveness can stimulate advocacy-related outcomes in social media environments (Choi et al., 2021). More recent review work further suggests that brand advocacy has become an increasingly important, yet still conceptually fragmented, construct in consumer-brand relationship research, leaving room for studies that examine recommendation-oriented responses more explicitly and in closer connection with purchase-related outcomes (Shimul et al., 2025). Related evidence also indicates that social-media-based community mechanisms can promote advocacy through engagement and brand passion, reinforcing the view that recommendation-oriented behaviour deserves more direct analytical attention in digital environments (Wong et al., 2023).

3 Methodology

Existing literature has primarily examined purchase intention as a key factor in patterns of purchase behaviour, while recommendation intention is often analysed indirectly through constructs such as e-WOM. This creates a gap in understanding how these two behavioural intentions are jointly formed and interconnected within the digital environment. Furthermore, although social media is recognised as an interactive ecosystem shaped by social influence, user engagement, and user-generated content, there remains a lack of empirical evidence on how these factors simultaneously influence both purchase and recommendation intentions, particularly among younger generations. The specific objectives are:

- To examine the influence of selected factors on recommendation intention, with particular focus on social influence and e-WOM.
- To explore the role of sociodemographic variables in shaping the relationship between purchase intention and recommendation intention.
- To assess the interdependence between purchase intention and recommendation intention using a seemingly unrelated bivariate probit model.
- To provide a more comprehensive understanding of how digital interaction, information exchange, and peer influence together shape consumer behaviour in social media environments.

The purpose of this study is to examine the attitudes and behaviours of social media users regarding their purchase intentions and recommendations on social media, as well as the role of brands in influencing users' behaviour. In addition, this study aims to determine the relationship between users' purchase intentions and their activities on social media.

The research was conducted using a survey questionnaire created with Google Forms, distributed between May and October 2025. A total of 506 completed questionnaires were received. As the research focused on generations Y and Z, respondents born before 1981 and after 2007 were excluded from the sample. After these exclusions, 477 validly completed questionnaires were included in the analysis, forming the final sample. The sample was purposive, with respondents selected based on their year of birth, specifically belonging to the targeted generations. The questionnaire consisted of 31 closed-ended questions. Three collected basic demographic data (age, gender, and level of education), while the remaining 28 presented statements to assess attitudes and experiences related to social media. Respondents indicated their level of agreement with the statements using a five-point ordinal scale (1 – completely disagree; 5 – completely agree).

The statements used in the study were based on previously validated questionnaire, organised into several groups. The first group concerns the tendency to share and interact (BE – Brand engagement), examining respondents' willingness to share opinions, feelings, and information about products and services via social media (Thanasi-Boçe et al., 2022). Product brand recognition (BA – Brand awareness) aims to determine the contribution of social media to the recognition of products and brands (Dabbous & Barakat, 2020; Faisal & Ekawanto, 2021). The final group concerns product brand perception and image (BI – Brand image), assessing how social media influence the creation and reinforcement of a positive

image of the product and its brand (Barreda et al., 2020; Faisal & Ekawanto, 2022).

The next group of statements (SMAC – Social media marketing) concerns activities such as posted content and its attractiveness, focusing on perceptions of the interest in product-related content and the ease of expressing opinions about it (Ismail, 2017). The following four statements, related to CS (Communication strategy), measure the informational and cognitive impact of social media on users, specifically how much social media facilitate communication with a company, stimulate interest, and help users understand the features and benefits of a product (Senalasani et al., 2024; Evania et al., 2023; Idris et al., 2017). In addition, three statements on E-WOM concern user activities involving actively following and using other users' posts on social media to gather information, form impressions of products or brands, and make purchase decisions (Ayuni, 2020; Kudeshia & Kumar, 2017; Reza Jalilvand & Samiei, 2012).

Statements on purchase intentions (PI) were adapted from Dabbous and Barakat (2020) to measure users' intention to purchase products via social media, that is, how inclined users are to buy through social media when they need a specific product. To examine the determinants of purchase intent via social media and recommendation intent, the seemingly unrelated bivariate probit model was applied to address two binary dependent variables simultaneously. Both variables are conceptually related, as they arise from a similar set of attitudes, perceptions, and experiences of users, but represent different forms of behavioural intentions. As these are dichotomous outcomes (0/1), classical linear regression is not appropriate, while separate probit models would ignore the possible correlation of unobserved factors affecting both dependent variables simultaneously. The seemingly unrelated bivariate probit model enables joint estimation, with two probit equations estimated simultaneously, assuming that the errors of these equations are mutually correlated and follow a bivariate normal distribution.

The model consists of two equations describing respondents' propensity to buy through a social network (eq. 1) and their propensity to recommend (eq. 2), with the observed binary variables determined by whether the latent propensity exceeds a specified threshold. Each equation includes the sociodemographic variables gender, generation (Generation Y or Z), and education level. Additionally, a set of factor variables derived from the previously described factor analysis, representing concise dimensions of attitudes and perceptions related to social media and shopping experience, is included. The parameters were estimated using the maximum likelihood method, which provides consistent and efficient estimates under conditions of correlated errors.

$$PI_i = \beta_0 + \beta_1 Gender + \beta_3 GenerationZ + \beta_4 Education + \beta_5 Recognition + \beta_6 Interaction + \beta_7 Involvement + \beta_8 Participation + \varepsilon_1 \quad (1)$$

$$RI_i = \gamma_0 + \gamma_1 Gender + \gamma_3 GenerationZ + \gamma_4 Education + \gamma_5 Recognition + \gamma_6 Interaction + \varepsilon_2 \quad (2)$$

The proposed approach enables a deeper understanding of the relationship between purchase intention and recommendation intention, while providing a methodologically robust basis for interpreting empirical findings and applying them in further theoretical and practical analyses. The research employed appropriate statistical methods to process and present the collected data, using the computer programmes SPSS 29.0 and STATA. The following paper will describe the sample on which the research was conducted.

4 Research Results

In this section, two multivariate factor analyses (EFA) were conducted using the statements from the measuring instrument, specifically the survey questionnaire. Cronbach's alpha was used to assess the reliability of the measurement scales for social media brand impact and social media activities. According to the reliability coefficient test, the first measurement scale, which includes the dimensions of Brand Image (BI), Brand Engagement (BE), and Brand Awareness (BA), was found to be highly reliable, with a Cronbach's alpha value of 0.886. The second set of statements, which includes the variables Communication Strategy (CS), Social Media Marketing (SMAC), and e-WOM, also demonstrated high internal consistency, with a Cronbach's alpha coefficient of 0.909, confirming the significant reliability of the social media activities measurement scale. The Bartlett test of sphericity and the Kaiser-Meyer-Olkin (KMO) test were used to validate the measurement scales. The results for the KMO and Bartlett tests of sphericity for the two measurement scales were high: for the social media brand impact measurement scale, the KMO was 0.887, and for the social media activities measurement scale, the KMO was 0.894. For both measurement scales, the Bartlett test of sphericity was statistically significant ($\alpha \leq 0.05$), indicating that factor analysis is appropriate for both measurement scales. Table 1 presents the total variance explained for the social media brand impact measurement scale.

Table 1. Total variance explained for the social media brand impact measurement scale

	Initial eigenvalues			Rotation Sums of Squared Loadings		
	Total	Variance %	Cumulative %	Total	Variance %	Cumulative %
1	5.32	48.4	48.4	4.85	44.1	44.1
2	2.31	69.4	69.4	2.78	25.3	69.4

Extraction Method: Method Principal Component Analysis. Source: Research findings (N=477)

The total variance explained for the social media activities measurement scale is shown in Table 2.

Table 2. Total variance explained for the social media activities measurement scale

	Initial eigenvalues			Rotation Sums of Squared Loadings		
	Total	Variance %	Cumulative %	Total	Variance %	Cumulative %
1	5.94	54.0	54.0	3.87	35.2	35.2
2	1.31	11.9	66.0	3.38	30.8	66.0

Extraction Method: Method Principal Component Analysis. Source: Research findings (N=477)

As shown in the tables for both measurement scales, a high total variance explained by the factors indicates that the extracted factors effectively summarise the main structure of the data, represent the latent constructs well, and support the validity of the measurement instrument. To simplify interpretation of the factors, Varimax rotation was applied, resulting in a clearer structure of factor loadings. For the social media brand impact measurement scale, the total variance explained was 69.39%, and for the social media brand experience measurement scale, it was 65.97%. Based on the Rotated Component Matrix for the social media brand impact measurement scale, two factors were identified with eigenvalues above 1, explaining the fundamental dimensions of social media influence: Recognition and Interaction. The first factor, Recognition (explaining 44.09% of the total variance), reflects users' ability to recognise, recall, and understand product-related information provided on social media. The second factor, Interaction (explaining 25.29% of the total variance), captures users' active participation and social engagement with brand-related content on social media.

For the second measurement scale, two factors were extracted and named: Involvement and Participation. The first factor explained 35.02% of the total variance and describes how social media influences customers' perceptions and engagement with products, including interest, understanding of product features and benefits, and active participation in consuming content. The second factor, which explained 30.76% of the total variance, defines social influence and the sharing of information on social media, such as following other users' posts, making decisions based on their experiences, and sharing recommendations and information with friends. The factors Recognition, Interaction, Involvement, and Participation are included in the analysis of determinants of users' behaviour using seemingly unrelated probit regressions, the results of which are shown in Table 3 below.

Table 3. Seemingly unrelated probit regressions results

	Coeff.	**Std. err.**	**Z**	**Sig.**
Dependent variable - PI				
Gender	-0,0910	0,1556	-0,58	0,559
GenZ	-0,0412	0,1599	-0,26	0,797
Education – ref. cat. High school				
College	-0,0102	0,2248	-0,05	0,964
Graduate	-0,1455	0,1633	-0,89	0,373
Master	-0,474	0,2451	-1,94	0,053
PhD	-1,2854	0,6192	-2,08	0,038
Recogn	0,6312	0,1205	5,24	0,000
Interact	0,3210	0,0771	4,16	0,000
Involv	0,265	0,1062	2,53	0,011
Particip	0,4080	0,1080	3,78	0,000
_cons	0,0875	0,2767	0,32	0,752
Dependent variable - RI				
Gender	0,3873	0,1404	2,76	0,006
GenZ	0,3460	0,1442	2,40	0,016
Education – ref. cat. High school				
College	0,0487	0,2156	0,23	0,821
Graduate	0,1235	0,1620	0,76	0,446
Master	-0,0526	0,2184	-0,24	0,810
PhD	0,0887	0,4199	0,21	0,833
Recogn	0,6085	0,0780	7,80	0,000
Interact	0,2966	0,710	4,18	0,000
_cons	-0,3738	0,2493	-1,50	0,134
/athrho	0,4527	0,1048	4,32	0,000
rho	0,424	0,0859		
Wald test of rho=0: chi2(1) =18.6564 Prob>chi2=0.0000				

Source: Authors' calculation

The key parameter of the model is the rho coefficient, which measures the correlation between the errors of the two equations, that is, the extent to which unobserved factors jointly influence purchase intent and recommendation intent. The statistical significance of the rho parameter was tested using the Wald test of the hypothesis that rho is zero. Rejecting this hypothesis indicates that evaluating the two probit models separately would result in a loss of information and potentially biased conclusions, further supporting the use of the bivariate approach. The results of the model suggest that the intention to purchase via a social

network and the intention to recommend are statistically significantly correlated but are also influenced by partially different groups of factors. This enables the model to identify the effects of individual independent variables more reliably, while accounting for the interdependence of the two behavioural intentions.

Socio-demographic variables generally do not have a statistically significant effect on the intention to buy through social media, suggesting that willingness to buy via social media is relatively homogeneous across genders and generations. However, both gender and generation have statistically significant and positive effects on the intention to recommend. This indicates that Gen Z users and women are more likely to share recommendations and positive experiences in the digital environment.

Education coefficients are statistically significant and negative in the first equation which may indicate greater criticism or caution among highly educated individuals towards shopping through social media. Education has no statistically significant effect in second equation, suggesting that the propensity to recommend is not systematically influenced by the level of education.

Factors included in both equations are statistically significant and positive, which can be explained by their effects on users' information processing, perceived value, trust, and emotional involvement, all of which are central to decision-making in digital environments.

Recognition, based on brand awareness and image, reduces cognitive effort in decision-making and increases the likelihood of product selection. Frequent and multidimensional exposure (visual, informational, and emotional) on social media enhances the mental availability of brands and their perceived value. Additionally, users' willingness to comment and share opinions strengthens purchase intention through psychological involvement, as public expression reinforces identification with the brand and consistency between attitudes and behaviour.

Involvement reflects the perceived informational and communicative value of social media. When users regard these platforms as reliable sources of information and interaction, perceived risk decreases and confidence in purchase decisions increases, facilitating the transition from evaluation to purchase. Participation, including electronic word-of-mouth and user interaction, further reduces uncertainty and strengthens trust by drawing on others' experiences. In this way, social media serve as a collective quality filter, with decisions shaped by aggregated community knowledge. Overall, purchase intention arises from the interplay of information, trust, social influence, and emotional-cognitive engagement within an integrated digital ecosystem.

Brand awareness and image also have a strong positive effect on recommendation intention. A clear and positive perception of a product increases users' confidence in their judgement, which is crucial given the reputational responsibility associated with making recommendations. These brand-related factors enhance the effects of social interaction and e-WOM, as trust built through a positive image is spread via recommendations and content sharing. Consequently, strengthening brand image through social media not only increases purchase intention but also stimulates recommendation intention, extending the reach and effectiveness of marketing activities. To provide a more intuitive interpretation of the probit model coefficients, average marginal effects were calculated (Table 4).

Table 4. Average marginal effects

	Marginal effect	**st. err**	**Sig.**
Gender	0.023	0.036	0.511
GenZ	0.028	0.036	0.440
College	0.004	0.049	0.937
Graduate	-0.014	0.039	0.714
Master	-0.095	0.055	0.082
PhD	-0.249	0.079	0.002
Recogn	0.180	0.026	0.000
Interact	0.090	0.017	0.000
Involv	0.050	0.021	0.016
Particip	0.076	0.021	0.000

Source: Authors' calculation

To provide a more intuitive interpretation of the probit model coefficients, average marginal effects were calculated. Among the variables analysed, Recognition has the strongest effect: a one-unit increase raises the probability of simultaneous purchase and recommendation intention by approximately 18 percentage points. This finding underscores the central role of brand awareness and image in enhancing consumers' readiness to both purchase and recommend products via social media, consistent with previous research (Dabbous and Barakat, 2020).

Participation also has a positive and statistically significant effect, increasing this probability by around 9 percentage points. This highlights the importance of electronic word-of-mouth, peer recommendations, and shared user experiences in reducing uncertainty and strengthening trust (Erkan and Evans, 2016; Liu et al., 2024). Involvement contributes an additional 7.6 percentage points, indicating that informative and engaging content facilitates product evaluation and enhances perceived value. Finally, Interaction increases the probability by approximately 5 percentage points, suggesting that active engagement with brand-related content reinforces consumer–brand relationships.

The results can be interpreted within the theoretical framework of information asymmetry (Akerlof, 1970) and signalling (Spence, 1973). In digital markets characterised by incomplete information, social media

reduce uncertainty by providing multiple quality signals. The strong effect of recognition reflects the role of brand awareness and image as credible signals that lower perceived risk and facilitate decision-making. Participation, through electronic word-of-mouth and peer-generated content, acts as a decentralised signalling mechanism that reveals product quality based on collective user experience. Involvement enhances the interpretation of these signals by increasing the perceived informational value of social media, while interaction reinforces trust through active engagement with brand-related content. Overall, the findings suggest that social media function as integrated signalling environments that mitigate information asymmetry, thereby simultaneously strengthening purchase and recommendation intentions.

5 Concluding Remarks

The research provided insight into the complex dynamics between social media activities and the behavioural intentions of Generations Y and Z. The aim of this study was to examine how different dimensions of engagement on digital media shape purchase intent and willingness to recommend products. The research was based on primary data collected in 2025 through a structured questionnaire administered to a sample of 477 respondents. Factor analysis was conducted, resulting in the extraction of four factors: Recognition, Interaction, Involvement, and Participation. Recognition includes brand recognition and awareness, while Interaction involves active user engagement on social media. For the second measurement scale, two factors were extracted: Involvement, which encompasses perceived information and communication value, and Participation, which refers to behaviours reflecting social influence and information sharing on social media. Regression analysis showed that purchase intent is positively influenced by all extracted factors. These findings suggest that social media are no longer merely promotional channels, but complete ecosystems that reduce perceived risk and facilitate the transition from the evaluation phase to the purchase phase. Interestingly, sociodemographic variables did not play a significant role in purchase intent, except for level of education; highly educated individuals showed a higher level of caution when shopping through social media. In contrast, the intention to recommend is strongly determined by gender and generation. The results indicate that women and members of Generation Z are the most active promoters of brands in the digital environment, identifying them as key drivers of electronic word of mouth.

Despite the robust methodology, this research has certain limitations. The sample was purposive, focusing exclusively on Generations Y and Z, which limits the generalisation of conclusions to other population groups. Additionally, the data were collected within a specific period (May to October 2025), which may reflect market trends subject to rapid changes in the digital environment. The use of only closed questions on the Likert scale limits a more comprehensive understanding of users' subjective motives, which qualitative research could provide. These limitations suggest that future studies could include qualitative and longitudinal research to track the evolution of trust in social media as sales channels. It would also be useful to expand the model by including variables such as product type or the influence of specific platforms (e.g., TikTok versus Instagram) on decision-making, providing a more detailed understanding of modern users' digital behaviour.

This paper makes a methodological contribution by applying a seemingly unrelated bivariate probit model, enabling the simultaneous analysis of purchase and recommendation intentions while accounting for the correlation between them. The results confirm that these two intentions are statistically significantly correlated. The findings are particularly useful for marketers and brand managers in optimising strategies targeting Generations Y and Z, as factors such as interaction and e-WOM directly stimulate purchase intention.

References

George A. Akerlof, 1970. "The Market for "Lemons": Quality Uncertainty and the Market Mechanism," *The Quarterly Journal of Economics, President and Fellows of Harvard College*, vol. 84(3), pages 488-500. https://doi.org/10.2307/1879431

An, G.K., Ngo, T.T.A. (2025). Uncovering the influence of social media marketing activities on Generation Z's purchase intentions and eWOM for organic cosmetics. *PLoS One*, 20(6), e0325953, https://doi.org/10.1371/journal.pone.0325953

Ayuni, R.F. (2020). Bringing virtual communities into a marketing strategy to create purchase intentions in the social media era. *Journal of Indonesian Economy and Business*, 35(2), 112-128. https://doi.org/10.22146/jieb.53261

Barger, V., Peltier, J.W., Schultz, D.E. (2016). Social media and consumer engagement: a review and research agenda. *Journal of Research in Interactive Marketing*, 10(4), 268-287, https://doi.org/10.1108/JRIM-06-2016-0065

Barreda, A.A., Nusair, K., Wang, Y., Okumus, F., Bilgihan, A. (2020). The impact of social media activities on brand image and emotional attachment: A case in the travel context. *Journal of Hospitality and Tourism Technology*, 11(1), 109-135, https://doi.org/10.1108/JHTT-02-2018-0016

Cheung, M.L., Pires, G., Rosenberger, P.J. (2020). The influence of perceived social media marketing elements on consumer–brand engagement and brand knowledge. *Asia Pacific Journal of Marketing and Logistics*, 32(3), 695-720, https://doi.org/10.1108/APJML-04-2019-0262

Choi, Y., Kroff, M.W., Kim, J. (2021). Developing brand advocacy through brand activities on Facebook. *Journal of Consumer Marketing*, 38(3), 328-338, https://doi.org/10.1108/JCM-10-2019-3460

Dabbous, A., Barakat, K.A. (2020). Bridging the online offline gap: Assessing the impact of brands' social network content quality on brand awareness and purchase intention. *Journal of Retailing and Consumer Services*, 53, https://doi.org/10.1016/j.jretconser.2019.101966

Deepika, K.S., Massand, A. (2025). A study on reflective factors of social media marketing activities and its influence on purchase intention of gen z. *Acta Psychologica*, 259, https://doi.org/10.1016/j.actpsy.2025.105459

Dimoka, A., Hong, Y., & Pavlou, P. A. (2012). On Product Uncertainty in Online Markets: Theory and Evidence. *MIS quarterly*, *36*(2), 395-426. https://doi.org/10.2307/41703461

Dong, F. (2022). Evaluation of Product Uncertainty in Online Shopping. *Information Systems and Economics*, 16-21. https://doi.org/10.23977/infse.2022.030603

Dwivedi, Y.K., Ismagilova, E., Hughes, D.L., Carlson, J., Filieri, R., Jacobson, J., Jain, V., Karjaluoto, H., Kefi, H., Krishen, A.S., Kumar, V., Rahman, M.M., Raman, R., Rauschnabel, P. A., Rowley, J., Salo, J., Tran, G.A., Wang, Y. (2021). Setting the future of digital and social media marketing research: Perspectives and research propositions. *International Journal of Information Management*, 59, https://doi.org/10.1016/j.ijinfomgt.2020.102168

Erkan, I., Evans, C. (2016). The influence of eWOM in social media on consumers' purchase intentions: An extended approach to information adoption. *Computers in Human Behavior*, 61, 47-55, https://doi.org/10.1016/j.chb.2016.03.003

Evania, S., Listiana, E., Rosnani, T., Fahruna, Y. (2023). The Effect of Influencer Marketing and Content Marketing on Customer Engagement and Purchase Decisions on Followers. *Asian Journal of Economics, Business and Accounting* 23(2), 12-20. https://doi.org/10.9734/ajeba/2023/v23i2917

Faisal, A., Ekawanto, I. (2021). The role of Social Media Marketing in increasing Brand Awareness, Brand Image and Purchase Intention. *Indonesian Management and Accounting Research*, 20(2), 185-208. https://doi.org/10.25105/imar.v20i2.12554

Gensler, S., Völckner, F., Liu-Thompkins, Y., Wiertz, C. (2013). Managing Brands in the Social Media Environment, *Journal of Interactive Marketing*, 27(4), 242-256, https://doi.org/10.1016/j.intmar.2013.09.004

Idris, J., Abd Rahman, S.S. (2017). Aida Model, Social Media Marketing and Decision to Patronize of Muslim Friendly Hotel. *Journal of Management and Muamalah*, 7(2), 3-15. Retrieved from https://jmm.uis.edu.my/index.php/jurnal/article/view/57

Ismail, A.R. (2017). The influence of perceived social media marketing activities on brand loyalty: The mediation effect of brand and value consciousness. *Asia Pacific Journal of Marketing and Logistics*, 29(1), 129-144, https://doi.org/10.1108/APJML-10-2015-0154

Kudeshia, C., Kumar, A. (2017). Social eWOM: does it affect the brand attitude and purchase intention of brands?. *Management Research Review*, 40(3), 310-330, https://doi.org/10.1108/MRR-07-2015-0161

Liadeli, G., Sotgiu, F., Verlegh, P.W.J. (2023). A Meta-Analysis of the Effects of Brands' Owned Social Media on Social Media Engagement and Sales. *Journal of Marketing*, 87(3), 406-427, https://doi.org/10.1177/00222429221123250

Lim, W.M., Rasul, T. (2022). Customer engagement and social media: Revisiting the past to inform the future. *Journal of Business Research*, 148, https://doi.org/10.1016/j.jbusres.2022.04.068

Liu, H., Jayawardhena, C., Shukla, P., Osburg, V.-S., Yoganathan, V. (2024). Electronic word of mouth 2.0 (eWOM 2.0) – The evolution of eWOM research in the new age. *Journal of Business Research*, 176, https://doi.org/10.1016/j.jbusres.2024.114587

Ma, Y. (2024). A socio-technical analysis of factors affecting consumer engagement in livestream shopping: Evidence from structural equation modeling and fuzzy set qualitative comparative analysis. *Telematics and Informatics*, 87, https://doi.org/10.1016/j.tele.2023.102094

McClure, C., Seock, Y.-K. (2020). The role of involvement: Investigating the effect of brand's social media pages on consumer purchase intention. *Journal of Retailing and Consumer Services*, 53, https://doi.org/10.1016/j.jretconser.2019.101975

Ngo, T.T.A., Bui, C.T., Chau, H.K.L., Tran, N.P.N. (2024). Electronic word-of-mouth (eWOM) on social networking sites (SNS): Roles of information credibility in shaping online purchase

intention. *Heliyon*, 10(11), e32168, https://doi.org/10.1016/j.heliyon.2024.e32168

Ong, Y.X., Sun, T., Ito, N. (2024). Understanding followers' processing and perceptions of social media influencer destination marketing: Differences between #Millennials and #GenZs. *Journal of Hospitality and Tourism Management*, 60, 194-207, https://doi.org/10.1016/j.jhtm.2024.07.003

Onofrei, G., Filieri, R., Kennedy, L. (2022). Social media interactions, purchase intention, and behavioural engagement: The mediating role of source and content factors. *Journal of Business Research*, 142, 100-112, https://doi.org/10.1016/j.jbusres.2021.12.031

Reza Jalilvand, M., Samiei, N. (2012). The effect of electronic word of mouth on brand image and purchase intention: An empirical study in the automobile industry in Iran. *Marketing Intelligence & Planning*, 30(4), 460-476, https://doi.org/10.1108/02634501211231946

Senalasari, W., Setiaji, M., Khrisnamurti, N., Februadi, A. (2024). Optimizing Brand Awareness and Purchase Intention Travel Agents: Communication Strategy through Social Media Marketing. *Journal of Marketing Innovation* (JMI), 4(1), https://doi.org/10.35313/jmi.v4i1.98

Shimul, A.S., Faroque, A.R., Quader E. (2026). Brand advocacy: a scoping review and future research agenda. *Journal of Product & Brand Management*, 35(1), 16-32, https://doi.org/10.1108/JPBM-10-2024-5537

Spence, M. (1973). Job market signaling. *Quarterly Journal of Economics*, 87(3), 355–374. https://doi.org/10.2307/1882010

Thanasi-Boçe, M., Zeqiri, J., Kurtishi-Kastrati, S. (2022). The Role of Social Media Usage and Engagement on Purchase Intentions for Fashion Brands. *International Journal of E-Services and Mobile Applications (IJESMA*), 14(1), 1-14. https://doi.org/10.4018/IJESMA.300269

Xu, Y., Chen, W., Ow, T.T. (2023). The effects of social media posts' characteristics on customer engagement: Evidence from WeChat. *Information & Management*, 60(7), https://doi.org/10.1016/j.im.2023.103854

Wong, A. (2023). How social capital builds online brand advocacy in luxury social media brand communities. *Journal of Retailing and Consumer Services*, 70, 103143, https://doi.org/10.1016/j.jretconser.2022.103143

Application of Business Analytics in Demand Planning: An Example from the Fast Food Industry

Bisera Andrić Gušavac, Milena Popović, Biljana Panić, Dragan Pamučar
Faculty of Organizational Sciences, University of Belgrade, Department of Operations Research and Statistics
Jove Ilića 154, Belgrade, Serbia
{bisera.andric.gusavac, milena.popovic, biljana.panic, dragan.pamucar}@fon.bg.ac.rs

Mihailo Vučeljić
Simplify Outsourcing d.o.o.
Milovana Glišića 2B, Belgrade, Serbia
vuceljicmihailo01@gmail.com

Abstract. *This paper presents an integrated business analytics framework for demand planning in the fast-food industry by combining descriptive, predictive, and prescriptive analytics within a single decision-support system. The framework is validated through a case study of a fast-food restaurant in Serbia, where highly dynamic raw material demand requires accurate forecasting and efficient inventory management. Descriptive analytics identify significant correlations, trends, and seasonal patterns in the data. The obtained forecasts are incorporated into a linear optimization model to support procurement decisions, reduce inventory costs, and minimize raw material shortages. The findings demonstrate that integrating forecasting with inventory optimization enhances operational decision-making and provides a practical, data-driven approach to demand planning. The proposed framework offers both managerial value and a methodological contribution to the application of integrated business analytics in demand planning.*

Keywords: Business analytics, demand management, fast food industry, optimization.

1 Introduction

The fast-food industry operates in a dynamic environment characterized by fluctuating demand, short product life cycles, and perishable raw materials. Therefore, accurate demand planning is essential, as forecasting errors can lead to excess inventory, shortages, higher costs, and operational inefficiencies. Business analytics is increasingly used to support data-driven decision-making and improve demand and inventory management.

Business analytics combines descriptive, predictive, and prescriptive analytics to support managerial decision-making. Descriptive analytics facilitates the exploration of historical data and the identification of patterns, trends, and seasonal effects. Predictive analytics applies statistical and forecasting techniques to estimate future demand, while prescriptive analytics uses optimization models to recommend the most appropriate decisions under defined operational constraints. Although these analytical approaches have been widely studied, previous research has predominantly focused on improving forecasting accuracy, with limited attention given to integrating forecasting results directly into procurement and inventory planning.

To address this research gap, this paper proposes an integrated decision-support framework that combines descriptive, predictive, and prescriptive analytics for demand planning in the fast-food industry. The framework links demand forecasting with inventory optimization, enabling managers to translate demand forecasts into procurement decisions while considering operational constraints.

The proposed framework is validated through a case study of Company X, a fast-food restaurant in Serbia, using historical consumption data for three key raw materials collected between 2019 and 2023. The study aims to examine how the integration of business analytics methods can improve demand planning, inventory management, and operational decision-making. The analysis is conducted in three stages. First, descriptive analytics is used to prepare and explore the data by identifying trends, seasonal patterns, and relationships among variables. Second, predictive analytics evaluates several forecasting models to identify the most accurate approach for demand prediction. Finally, prescriptive analytics integrates the selected forecasting model into a linear optimization model to determine optimal procurement decisions and support inventory planning.

The main contribution of this study is the development of a unified analytical framework that combines demand forecasting and inventory optimization within a single decision-support system. The proposed approach integrates descriptive, predictive, and prescriptive analytics to support

managerial decision-making and advance business analytics applications in demand planning.

2 Fast-Food Demand Planning

Demand management is the process of planning, forecasting, and directing demand for products and services while coordinating key business functions such as marketing, sales, and logistics. The goal is to meet customer needs while optimizing resource utilization, reducing costs, and increasing forecast accuracy. Of particular importance is the optimization of inventory to avoid surpluses and shortages (Farag, 2026).

The demand management process generally comprises five phases: data collection, demand forecasting, balancing supply and demand, demand shaping through pricing and promotional activities, and continuous performance monitoring. Contemporary challenges include forecast uncertainty, rapidly changing consumer behavior, supply chain disruptions, and limited integration of data across organizational functions (Ivanov & Dolgui, 2022; Wamba et al., 2023).

2.1 Specifics of the fast-food industry

The fast food industry continues to grow, fueled by busy lifestyles and a demand for speed (fortunebusinessinsights.com, 2025). However, this growth is accompanied by significant challenges related to managing perishable inventory, volatile demand, and rising operational costs.

The industry's key challenges include changing consumer preferences toward healthier and more sustainable food, labor shortages, rising operating costs, intense competition from fast-casual restaurants and online food delivery platforms, and increasing regulatory and environmental requirements. These trends require restaurant operators to continuously adapt their products, pricing strategies, and operational processes while improving forecasting accuracy and minimizing food waste (Kasavan et al., 2022; Alsuwaidi et al., 2022; Kwok, 2022).

3 Business Analytics and Its Role in Decision Making

Business analytics is the process of transforming data into actions through analysis and insight in the context of organizational decision-making and problem-solving (Liberatore & Luo, 2010). The authors divide it into several phases, and the most common division is: descriptive analytics, predictive analytics, and prescriptive analytics. Descriptive analytics interprets historical data and determines what happened. Predictive analytics should predict the values of variables that are not under the control of the decision maker or are under his partial control, and which will affect the success of the business (Vujošević & Vujošević, 2014). Prescriptive analytics uses optimization, heuristic, and simulation methods for the best courses of action based on the data obtained in the previous two phases.

Today's business is shaped by rapid technological progress and unstable markets, which create complex conditions for decision-making. Organizations produce and store huge amounts of data, which often remain unused (Bumblauskas et al., 2017). Although organizations generate and store unprecedented volumes of data, a substantial proportion remains underutilized for decision-making, limiting the potential value that can be extracted from organizational data assets (Liu, et al., 2023). In this context, it has become important to use business analytics to integrate decision-making processes in order to gain a competitive advantage and a higher level of company performance (Oesterreich et al., 2022; Debeljević & Peovski, 2026). The development of concepts such as “big data” has further accelerated this need. The rapid expansion of large and connected data sets creates new opportunities for innovation and improving the quality of decisions (Dwivedi et al., 2017).

In addition to collecting data, companies face the challenge of how to extract useful insights from data and turn them into actions. The focus is increasingly shifting towards data-driven decision-making. Business analytics provides tools and methodologies to transform raw data into strategic insights, helping managers make effective decisions. Research shows that the use of various types of analytics can help organizations make more intelligent decisions and increase business value from those decisions (Hindle & Vidgen, 2018; Li et al., 2022).

3.1 Limitations and Challenges in Implementation

Despite the significant advantages of business analytics for informed decision-making, its practical application faces several limitations and challenges. One of the main problems lies in the data itself, which is often inconvenient to analyze due to its huge volume, variety of formats, and rapid creation (Jagadish et al., 2014). This leads to various technical challenges in storing and processing data (Jagadish et al., 2014). Data quality is also of great importance because data can be incomplete, inconsistent or inaccurate, which makes it difficult to draw reliable conclusions later (Liu et al., 2023). In addition to technical aspects, a major challenge is the organizational culture itself, as well as the human aspects in accepting analytics. A large number of organizations still do not have a mature culture of data-driven decision-making, so employees may be reluctant to trust analytics or may have difficulties in

understanding analytical tools (Silaban, 2022). There is also a problem of translating analytical results into business actions, where it is not always clear how metrics and reports obtained from analytics can be transformed into concrete strategic decisions (Berman & Israeli, 2022). This lack of clarity leads to the fact that valuable analysis findings are sometimes not implemented, but remain at the report level. Another challenge is the operationalization of analytics, i.e., the integration of analytical models and tools into existing business processes and decision-making systems. Studies have shown that many companies have difficulty truly realizing value from business analytics programs in real-world decision-making conditions (Silaban, 2022).

The combination of these factors results in most organizations still using only a small portion of the available data for decision-making, as previously mentioned (Liu et al., 2023). Overcoming limitations such as data quality, lack of trust in analytics, and lack of integration of analytics teams into business processes is a prerequisite for realizing the full potential of analytics in decision-making.

4 Empirical Study

This paper utilizes data from Company X for the period 2019-2023, which pertains to the demand for three key raw materials that have the greatest impact on the production process. To preserve anonymity, these raw materials are designated as A, B, and C. Company X operates within the fast-food industry and generates approximately 20 million dinars in average monthly net sales. The costs of purchasing raw materials and other production inputs account for between 30% and 35% of total sales revenue. The three raw materials under consideration represent the most significant inputs, with their share accounting for approximately 50% of total procurement costs. Raw materials A and B are ordered on a daily basis due to their short shelf lives, while raw material C is purchased every three to four days, resulting from high inventory turnover and limited storage capacity. Demand fluctuations are driven by various factors, including promotions, local events, weather conditions, and sporting activities.

4.1 Application of Descriptive Analytics

Table 1 shows the results of a descriptive analysis of the consumption of the observed raw materials, which provide a comprehensive picture of the level of consumption and its variability, as well as the stability and patterns of demand during the observed period.

Raw material A shows moderate variability in consumption with a slight positive skewness, which means that occasional months with higher consumption have increased the average. Raw material B has a higher level of average consumption and more pronounced variability, with positive skewness and a flatter distribution. Raw material C has the lowest consumption and the most stable pattern, with values that are closest to a normal distribution and without extreme deviations.

Table 1. Results of descriptive analytics

Measure	Raw material A	Raw material B	Raw material C
Arithmetic Mean	18435.417	25618.033	555.359
Median	17631	25075.500	541.395
Standard deviation	2906.11	4980.308	108.111
Minimum	13753	18395	375.880
Maximum	26665	38933	851.990
Range	12912	20538	476.110
Asymmetry Coefficient	0.982	0.895	0.879
Coefficient flattening	1.108	0.604	0.823
95% confidence interval	±750.729	±1286.550	±27.928
Number of observations	60	60	60

Correlation analysis was conducted to examine the relationship between the observed variables. The results, obtained using the Pearson correlation coefficient and additionally confirmed by scatter plots with regression lines, indicate a high degree of interdependence between the variables.

The correlation coefficients between the observed raw materials range from 0.857 to 0.963, indicating strong positive linear relationships among all variables. Beginning with the lowest correlation, raw materials A and B show $r = 0.857$ ($R^2 = 0.734$), meaning about 73% of the variability in the demand for raw material B is explained by variations in A. Moving to a higher correlation, raw materials A and C present $r = 0.928$ ($R^2 = 0.862$), reflecting a pronounced common movement in their consumption, partly due to technological and process connections. Finally, the highest degree of correlation is seen between raw materials B and C ($r = 0.963$; $R^2 = 0.927$), indicating an almost proportional movement of these variables and a very high level of interdependence in their consumption.

4.2 Application of Predictive Analytics

This chapter presents the results of predictive time series analysis conducted on the demand for raw material A in Company X. The objective was to identify the forecasting model that provides the highest accuracy based on historical demand data. Table 2 summarizes the performance of the evaluated time series models using standard forecast accuracy indicators calculated on the validation dataset. These indicators enabled an objective comparison of model performance and the selection of the most suitable approach for future demand forecasting. The identified model was subsequently considered for practical application in demand planning and data-driven business decision-making.

The results of the conducted analysis indicate that there is no universally optimal time series model that would provide the best results for all observed raw materials. Instead, for each time series, a model was identified that, according to the selected evaluation criteria, achieves the highest level of forecast accuracy. Based on the findings, the models that showed the best performance according to the criterion of mean absolute percentage error (MAPE) for each raw material were applied in the continuation of the research. Specifically, for raw material A, the Holt method proved to be the most accurate, whereas for raw material B, the multiplicative Holt–Winters method demonstrated the highest forecasting accuracy. For raw material C, the damped multiplicative Holt–Winters method achieved the best results (Winters, 1960). Table 4 shows the forecasted consumption quantities for each raw material for 2024, obtained by applying the aforementioned models that stood out as the most efficient within the framework of the conducted analysis.

Table 2. Results of predictive analytics

Time series	Raw material A		Raw material B		Raw material C	
Criterion	*RMSE*	*MAPE*	*RMSE*	*MAPE*	*RMSE*	*MAPE*
Method	Naive method	*Holt*	Multiplicative *Holt-Winters*		Multiplicative *Holt-Winters*	Damped Multiplicative *Holt-Winters*
Error	2916.04	8.79	3254.35	8.01	72.00	7.74

The obtained forecasts for 2024 will be used as input parameters within the framework of prescriptive analytics, with the aim of optimizing orders and supporting decision-making in the process of demand planning and inventory management. In this way, the predictive results represent the basis for further modeling and defining optimal procurement strategies in Company X.

Table 3. Forecasted demand for raw materials in 2024

Month	Raw material A [pcs]	Raw material B [pcs]	Raw material C [L]
January	26542	29975	752.78
February	26844	30297	767.34
March	27146	30612	802.82
April	27448	30922	825.37
May	27750	31225	818.29
June	28052	31522	908.19
July	28354	31811	831.00
August	28656	32092	900.07
September	28958	32365	921.91
October	29260	32630	987.05
November	29562	32920	937.19
December	29864	33314	978.16

4.3 Application of Prescriptive Analytics

To ensure optimal procurement planning and inventory management in Company X, a linear optimization model was developed that incorporates forecasted demand, ordering costs, inventory holding costs, and operational constraints, aiming at minimizing total costs. The model applies inventory optimization and procurement planning principles under variable demand and limited resources, enabling a balance between inventory levels and maintenance costs, while maintaining continuity in production supply. The sets, parameters, and primary components of the model are defined below as the foundation for formulating the optimization problem:

- Time period (months): t = 1,…,T, where T represents the total number of months observed;
- Raw materials: $i \in$ {A,B,C};
- Decision variables: $x_{i,t}$ - quantity of ordered raw materials i in period t;
- Forecasted demand: $\widehat{D}_{i,t}$) - estimated demand for raw material i in period t;
- Ordering costs: $c^p_{i,t}$ - unit cost of ordering the i-th raw material in period t;
- Inventory holding costs: $c^h_{i,t}$ - cost of holding one unit of the i-th raw material during period t;
- Monthly procurement budget: B_t- available budget for procurement of all raw materials in period t;
- Minimum inventory level: S_i - minimum initial inventory level for raw material i;
- Maximum stock level index: $MaxCover_i$ - maximum allowed number of months of stock coverage for raw material I,
- Inventory balance: $I_{i,t}$ – quantity of raw material inventory i at the beginning of period t, obtained through the formula: $I_{i,t} = I_{i,t-1} + x_{i,t} - \widehat{D}_{i,t}$.

Based on the described parameters, a mathematical model of total cost minimization was formulated. The formulation of the objective function and model constraints is as follows:

$$(min) \sum_{t=1}^{T} \sum_{i \in \{A,B,C\}} (c^p_{i,t} * x_{i,t} + c^h_{i,t} * I_{i,t})$$

(s.t.)

$$I_{i,t} = I_{i,t-1} + x_{i,t} - \widehat{D}_{i,t} \ (\forall i,t) \quad 1)$$

$$\sum_i c^p_{i,t} * x_{i,t} \le B_t \ (\forall t) \quad (2)$$

$$I_{i,t} \le MaxCover_i * \widehat{D}_{i,t} \ (\forall i,t) \quad (3)$$

$$I_{i,t} \ge S_i \ (\forall i,t) \quad (4)$$

$$x_{i,t} \ge 0 \ (\forall i,t) \quad (5)$$

$$x_{A,t} \in N, x_{B,t} \in N \quad (6)$$

The objective function minimizes the total cost of purchasing and holding inventory during the observation period. The first group of constraints ensures that the inventory for each raw material at the

end of month t must be equal to the inventory of that raw material at the end of the previous month, increased by the order for month t and decreased by the forecast demand for month t. Constraints (2) ensure that the total value of orders in month t must not exceed the available budget for that month. Constraints (3) limit the inventory according to the corresponding *MaxCover* factor for each raw material. The *MaxCover* factor was introduced due to the perishability of raw materials. Within this constraint, it ensures that the inventory cannot exceed a certain percentage of the monthly demand to prevent spoilage of the ingredients. Constraints (4) ensure that the inventory level at the end of the month for each raw material must be greater than the defined monthly safety stock level S_i. Finally, constraints (5) and (6) represent natural constraints and an integer condition for raw materials A and B, given that they are ordered in pieces.

In the initial model setup, appropriate values are given for the parameters of forecasted demand, unit costs, safety stock level, initial stock, and MaxCover factor for raw materials. Also, an appropriate budget is defined for each month. The currency used throughout the paper is the Serbian dinar (RSD).

For raw material A, the unit ordering costs are $c^p_{A,t}$=53 RSD/piece, while the inventory holding costs are $c^h_{A,t}$ =2.75 RSD/piece. The initial inventory level $I_{A,0}$ is 10000 pieces, while the safety stock level S_A = 2000 pieces. The *MaxCover* factor for raw material A is 0.5. The forecasted quantities of raw material A by month were previously calculated using the Holt method and are shown in Table 4.

For raw material B, the unit ordering costs are $c^p_{B,t}$=48 RSD/piece, while the inventory holding costs are $c^h_{B,t}$ =2.40 RSD/piece. The initial inventory level $I_{B,0}$ is 20000 pieces, while the safety stock level S_B is 2500 pieces. The MaxCover factor for raw material B is 0.5. Forecasted monthly quantities obtained using the multiplicative Holt-Winters method are presented in Table 4.

For raw material C, the unit ordering costs are $c^p_{C,t}$ =250RSD/L, while the inventory holding costs are $c^h_{C,t}$ = 5 RSD/L. The initial inventory level $I_{C,0}$ is 200L, while the safety stock level S_C = 100L. The *MaxCover* factor for raw material C is significantly higher than for raw materials A and B due to its longer shelf life, and is 1.8. The forecasted quantities of raw material C per month were previously calculated using the damped multiplicative Holt-Winters method and are shown in Table 4.

The available budget for ordering raw materials A, B, and C is 3500000 RSD per month. The budget is determined based on the percentage share of the costs of these raw materials in relation to the company's total sales costs, as well as the share of total sales costs in the total monthly sales of Company X. The following chapter will present the results of applying the solver to optimize the previously described model in Excel.

The model is implemented so that the variables are $x_{i,t}$, the target cell represents the total cost in 12 months, and the constraints entered into the solver include: inventory balances for each raw material and month, monthly budget, *MaxCover* and minimum inventory, non-negativity, with $x_{A,t} \in N, x_{B,t} \in N$. Simplex LP was chosen as the solver method. The results of applying the solver are presented below. First, Table 4 shows the obtained decision variables, i.e., the values of the quantities of raw materials A, B, and C that need to be ordered in 2024 for each month. The objective function was 37625237.5 RSD, of which procurement costs were 37480237.5 RSD, and inventory holding costs were 144000 RSD. The budget was not a limiting factor because monthly consumption was less than the budgeted 3.5 million RSD. Also, *MaxCover* constraints were not active since the ending inventory was consistently below the upper limits. Monthly consumption is consistently below the defined budget, which indicates that the budget is somewhat higher than necessary in the current business conditions.

Table 4. Solver results

Raw material A	Quantities by month	Raw material B	Quantities by month	Raw material C	Quantities by month
$x_{A,1}$	18542	$x_{B,1}$	12474	$x_{C,1}$	652.78
$x_{A,2}$	26844	$x_{B,2}$	30296	$x_{C,2}$	767.34
$x_{A,3}$	27146	$x_{B,3}$	30612	$x_{C,3}$	802.82
$x_{A,4}$	27448	$x_{B,4}$	30922	$x_{C,4}$	825.37
$x_{A,5}$	27750	$x_{B,5}$	31225	$x_{C,5}$	818.29
$x_{A,6}$	28052	$x_{B,6}$	31521	$x_{C,6}$	908.19
$x_{A,7}$	28354	$x_{B,7}$	31810	$x_{C,7}$	831.00
$x_{A,8}$	28656	$x_{B,8}$	32092	$x_{C,8}$	900.07
$x_{A,9}$	28958	$x_{B,9}$	32365	$x_{C,9}$	921.91
$x_{A,10}$	29260	$x_{B,10}$	32630	$x_{C,10}$	987.05
$x_{A,11}$	29562	$x_{B,11}$	32919	$x_{C,11}$	937.19
$x_{A,12}$	29863	$x_{B,12}$	33314	$x_{C,12}$	978.16

Based on the obtained results, a clearly defined pattern of the optimal solution can be observed. Given that the budget constraints are not strict, and the purchase prices are constant, the optimal policy identified by the solver boils down to the Just-in-Time approach to inventory management. Within this approach, the initial month (January) is characterized by the use of existing stocks, while from February onwards, quantities corresponding to the forecasted demand are ordered, while maintaining a defined minimum level of safety stocks. In this way, continuity in supply is ensured while minimizing inventory holding costs and efficiently using system resources.

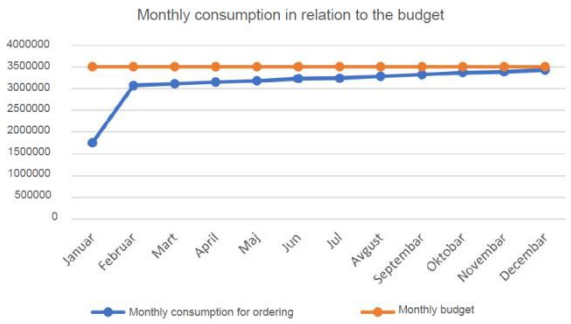


Figure 1. Monthly consumption compared to budget

Fig. 1 shows a comparative overview of actual consumption and available budget by month in the observed period. Based on the visual display, it is possible to see to what extent the realized consumption fits into the planned financial framework, as well as whether there are deviations in individual months.

In the observed period, consumption is below the budget limit, which indicates efficient cost management and stable procurement planning. This display allows management to clearly identify critical points in consumption, monitor financial discipline, and make timely decisions to optimize the budget and control costs.

5 Managerial and Theoretical Implications

The proposed framework provides practical support for managers operating in the fast-food industry by integrating demand forecasting with inventory optimization into a unified decision-support system. The results demonstrate that more accurate demand forecasts enable managers to reduce raw material shortages while avoiding excessive inventories of perishable products. By optimizing procurement decisions, the framework contributes to lower inventory holding costs, improved procurement timing, and more efficient allocation of financial resources.

Furthermore, the use of quantitative forecasting and optimization techniques supports data-driven decision-making, reducing reliance on managerial intuition. Since the framework is based on widely available analytical tools and standard forecasting methods, it can be readily adapted to other restaurants and small- and medium-sized enterprises (SMEs) facing similar demand planning challenges.

This study contributes to the business analytics literature by demonstrating the integration of descriptive, predictive, and prescriptive analytics within a single demand planning framework. Unlike previous studies that primarily focus on forecasting accuracy, the proposed approach establishes a direct connection between demand forecasting and inventory optimization, thereby extending the application of business analytics in operational decision-making.

The research also provides a methodological example of combining time-series forecasting with linear programming in a real business setting.

6 Discussion and conclusion

The results demonstrate that the proposed business analytics framework provides an effective approach to demand planning in the fast-food industry by integrating descriptive, predictive, and prescriptive analytics within a single decision-support system. Descriptive analysis identified clear trends and seasonal patterns in the consumption of key raw materials, providing a reliable basis for subsequent forecasting and optimization. Among the evaluated forecasting techniques, the Holt and Holt–Winters models achieved the highest forecasting accuracy, confirming their suitability for demand characterized by both trend and seasonality. The integration of these forecasts into a linear programming model enabled the optimization of procurement decisions while considering budgetary and storage constraints. Furthermore, scenario analysis confirmed that the proposed model remains robust under changes in key planning parameters, highlighting its practical applicability in dynamic business environments.

From a managerial perspective, the framework supports more efficient procurement planning by reducing the risk of raw material shortages and excessive inventories, thereby lowering inventory-related costs and improving resource utilization. By combining forecasting with optimization, the proposed approach enables managers to make data-driven procurement decisions rather than relying primarily on experience or intuition. Moreover, because the framework is based on widely available analytical methods and software tools, it can be readily adapted to other restaurants and small- and medium-sized enterprises operating under similar demand conditions.

From a theoretical perspective, this study extends the application of business analytics by demonstrating how descriptive, predictive, and prescriptive analytics can be integrated into a unified framework for operational decision-making. Unlike approaches that focus solely on forecasting performance, the proposed methodology establishes a direct connection between demand forecasting and inventory optimization, providing a practical example of integrating forecasting models with mathematical optimization in a real business context.

The study has several limitations. The analysis is based on a single case study and considers only three key raw materials, which may limit the generalizability of the findings. In addition,

forecasting accuracy depends on the quality and availability of historical data, while unexpected market changes, supply chain disruptions, and operational constraints may affect the practical implementation of the proposed model. A further limitation is that a substantial portion of the historical dataset corresponds to the COVID-19 period, during which consumer behavior and demand patterns differed from normal market conditions. Therefore, the forecasting models may partially capture extraordinary fluctuations that are not fully representative of post-pandemic demand patterns. However, the inclusion of 2023 as the validation period partially mitigates this limitation by evaluating model performance under more stable business conditions.

Future research could extend the framework by incorporating a larger number of products, additional operational constraints, and external demand drivers such as weather conditions, promotional activities, or economic indicators. Furthermore, the application of advanced machine learning techniques and multi-objective optimization models may further improve forecasting performance and procurement planning in complex supply chain environments.

References

Alsuwaidi, M., Eid, R., & Agag, G. (2022). Tackling the complexity of guests' food waste reduction behaviour in the hospitality industry. *Tourism Management Perspectives*, 42. doi:10.1016/j.tmp.2022.100963

Berman, R., & Israeli, A. (2022). The value of descriptive analytics: Evidence from online retailers. Marketing Science, *INFORMS*, 41(6), 1074-1096. doi:10.1287/mksc.2022.1352

Bumblauskas, D., Nold, H., Bumblauskas, P., & Igou, A. (2017). Big data analytics: Transforming data to action. *Business Process Management Journal*, 23(3), 703–720. doi:10.1108/BPMJ-03-2016-0056

Debeljević, S., & Peovski, F. (2026). Business Analytics for Ride-Hailing Platforms: Demand Forecasting and Ride-Completion Modeling. *Applied Decision Analytics*, 2(1), 283–294. https://doi.org/10.66972/ada21202625

Dwivedi, Y. K., Janssen, M., Slade, E. L., Rana, N. P., Weerakkody, V., Millard, J., & Snijders, D. (2017). Driving innovation through big open linked data (BOLD): Exploring antecedents using interpretive structural modeling. *Information Systems Frontiers*, 19(2), 197–212. doi: 10.1007/s10796-016-9675-5

Farag, M. I. H. (2026). Bridging Prediction and Prescription: Explainable AI for Managerial Decision Analytics under Uncertainty. *Applied Decision Analytics*, 2(1), 295–319. https://doi.org/10.66972/ada21202626

Fortune Business Insights. (2025). Fast food market size, industry share & COVID-19 impact analysis. Retrieved from https://www.fortunebusinessinsights.com/fast-food-market-106482

Hindle, G. A., & Vidgen, R. (2018). Developing a business analytics methodology: A case study in the food bank sector. *European Journal of Operational Research*, 268(3), 836–851. doi: 10.1016/j.ejor.2017.06.031

Ivanov, D., & Dolgui, A. (2022). The shortage economy and its implications for supply chain and operations management. *International Journal of Production Research*, 60(24), 7141–7154. doi:10.1080/00207543.2022.2118889

Jagadish, H. V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J. M., Ramakrishnan, R., & Shahabi, C. (2014). Big data and its technical challenges. *Communications of the ACM*, 57(7), 86–94. doi: 10.1145/2611567

Kasavan, S., Siron, R., Yusoff, S., & Fakri, M. F. R. (2022). Drivers of food waste generation and best practice towards sustainable food waste management in the hotel sector: A systematic review. *Environmental Science and Pollution Research*, 29, 48152–48167. doi:10.1007/s11356-022-19984-4

Kwok, L. (2022). Labor shortage: A critical reflection and a call for industry-academia collaboration. *International Journal of Contemporary Hospitality Management*, 34(11), 3929–3943. doi:org/10.1108/IJCHM-01-2022-0103

Li, Lei & Lin, Jiabao & Ouyang, Ye & Luo, Xin (Robert). (2022). Evaluating the impact of big data analytics usage on the decision-making quality of organizations. *Technological Forecasting and Social Change*. Elsevier, 175(C). doi:10.1016/j.techfore.2021.121355

Liberatore, M. J., & Luo, W. (2010). The analytics movement: Implications for operations research. *Interfaces*, 40(4), 313–324. doi: 10.1287/inte.1100.0502

Liu, S., Liu, O., & Chen, J. (2023). A Review on Business Analytics: Definitions, Techniques, Applications and Challenges. *Mathematics*, 11(4), 899. doi.org/10.3390/math11040899

Silaban (2022). Challenges in Business Analytics Implementation: A Comprehensive Review Using TOE Framework. Milestone. *Journal of Strategic Management*, 2(2). doi:org/10.19166/ms.v2i2.5953

Vujošević, M., & Vujošević, D. (2014). Business analytics: A challenge to managers, researchers and software engineers. *Info M*, 13(50), 4–14.

Wamba, S. F., Queiroz, M. M., & Trinchera, L. (2020). Dynamics between Blockchain Adoption Determinants and Supply Chain Performance: An Empirical Investigation. *International Journal of Production Economics*, 229(2). doi:org/10.1016/j.ijpe.2020.107791

Winters, P. R. (1960). Forecasting sales by exponentially weighted moving averages. *Management Science*, 6(3), 324–342. https://doi.org/10.1287/mnsc.6.3.324

Harnessing friction topology for stage-sensitive checkout diagnostics in supplement e-commerce

Aleksandar-Jon Stefan
Proton System / Abela Pharm, Belgrade
Faculty of Organizational Sciences, University of Belgrade
Jove Ilića 154, Beograd 11000, Serbia
aleksandarjon@gmail.com

Artur Bjelica
Faculty of Medicine, University of Novi Sad
Hajduk Veljkova 3, Novi Sad 21000, Serbia
Clinical Center of Vojvodina, Novi Sad
Hajduk Veljkova 1-9, Novi Sad 21000, Serbia
artur.bjelica@mf.uns.ac.rs

Abstract. *This study examined checkout disruption in supplement e-commerce using aggregated Google Analytics 4 (GA4) funnel exports from an established supplement brand. The purpose was to develop and test friction topology, a stage-sensitive diagnostic model that locates checkout losses across five funnel stages by device category and first user medium. Transition-loss vectors were computed for overall, device-based and medium-based contexts. The dominant friction peak occurred before cart entry, while device and medium differences emerged after cart entry, at the shipping, payment and purchase stages. The main contribution is a diagnostic model applicable to aggregated GA4 data, without individual user tracking.*

Keywords. Friction topology, GA4, checkout funnel, supplement e-commerce, user progression, device category, first user medium

1 Introduction

Checkout loss has become a major practical and theoretical challenge in digital commerce. Traffic acquisition costs are rising, while privacy regulations and platform measurement limits reduce visibility into user pathways. In practice, Google Analytics 4 (GA4) funnel reports in this context are usually reviewed as static tables during weekly performance meetings, with limited discussion of where exactly users drop out of checkout. In supplement e-commerce, this is compounded by product evaluation, health assurances, delivery expectations and payment trust. The supplement sector was selected as the research context because supplements are credence goods, where trust and reassurance matter at every stage of the checkout process. The sector combines health-related risk perception with routine online retail mechanics, which makes stage-specific friction more pronounced than in low-involvement categories. Funnel-level diagnostics in this sector also remain under-researched compared to fashion or electronics e-commerce. Users may therefore abandon their purchase before reaching the cart, when relevance is not established, or after entering the cart, when interaction costs become too high. Recent work on clickstream analysis, checkout simplification and cart abandonment shows that observed behaviour can reveal purchase disruptions hidden by final conversion rates (Jabr et al., 2023; Kukar-Kinney et al., 2022; Unal & Park, 2023).

Current funnel analytics do not fully address this issue. GA4 Funnel Exploration can show user retention or attrition across predefined stages. Still, its output is usually read as a descriptive report of drop-off rates rather than a structured behavioural model (Google, 2026b). Survey-based models of acceptance, risk and effort explain why users hesitate, but link less directly to the consolidated funnel exports used in day-to-day management (Chopra et al., 2024; Liu & Mensah, 2024; Roy & Shaikh, 2024). A gap exists between behavioural theory and routine analytics practice: the order, position and context of checkout loss are rarely preserved in a concise model for aggregated GA4 data.

This paper aimed to develop and assess friction topology as a stage-sensitive diagnostic model for checkout. Friction was treated not as an isolated factor but as a transition-level quantity from which topology is estimated. The hypotheses were that checkout loss would not be evenly distributed across the funnel, and that device category and first user medium would create distinct topological patterns once users had signalled intent by entering the cart. The analysis focused on a well-known supplement e-commerce brand, using GA4 data segmented by device and medium. The remainder of this paper is organised as follows: Section 2 presents a structured literature review and the development of hypotheses. Section 3 describes GA4 data, stage definitions and formulas. Section 4 presents the overall, device-based, and medium-based results. Section 5 discusses scientific contribution, managerial implications and limitations. Section 6 summarises the achieved contribution and presents future research directions.

2 Literature Review and Hypotheses

2.1 Structured review logic

Four concepts are used throughout this paper. A friction point is a single transition between two adjacent funnel stages at which users abandon the purchase path. Transition friction is the share of users who do not progress through a given friction point (formally defined in Section 3.2). A friction topology is the ordered vector of transition-friction values across all checkout stages for a given context; it describes how loss is distributed along the funnel rather than only how large it is in total. A friction peak is the transition with the highest friction value within a topology vector. A medium-friction topology (or device-friction topology) is the friction topology computed separately for each first user medium or device category.

The literature was screened to establish a basic framework before empirical analysis. Recent academic articles were reviewed across four interconnected domains: clickstream and journey modelling, cart abandonment, mobile and interface effort, and omnichannel acquisition contexts. Priority was given to studies examining behaviour, transition effort, perceived risk, or the structure of online journeys. The review aimed to support theory development and hypothesis formulation rather than to serve as a full systematic literature review. Clickstream and journey studies show that behaviour can be traced into interpretable models of online progression. Jabr et al. (2023) modelled revisiting and purchasing using clickstream data; Kim et al. (2024) incorporated browsing patterns into purchase prediction; and Padilla et al. (2025) approached the customer journey as a source of information under first-party data constraints. These studies support the view that meaningful information is contained in sequences and stages and not only at the final purchase event.

Cart abandonment studies were retained because checkout loss is multi-causal. In recent work, abandonment has been linked to search and comparison, risk perception, online review confusion, cognitive load, choice variety, and device conditions (Chopra et al., 2024; Das Sarma et al., 2025; Kukar-Kinney et al., 2022; Roy & Shaikh, 2024). This stream supports the assumption that an identical abandonment rate can have varying implications depending on the stage at which it occurs. Mobile commerce, online experience, and omnichannel studies were retained since context may alter the task the user is performing. Device differences can affect ease of interaction and form completion (Almeida Lucas et al., 2023), while acquisition context can shift expectation alignment, confidence and readiness to proceed (Fang & Li, 2025; Koronaki et al., 2023; Moliner & Tortosa-Edo, 2024). Therefore, a diagnostic model should preserve not only the amount of loss, but also its location and segment context.

2.2 Research gap and hypotheses

Two gaps can be identified in the retained literature. First, behavioural analytics is often derived from rich clickstream or transaction-level datasets, whereas many e-commerce firms still rely on aggregated GA4 funnel exports for routine decision-making. Second, cart abandonment theory typically explains why checkout loss occurs, whereas standard funnel reporting mainly shows where users drop off the purchase path. The proposed model was therefore developed to bridge these perspectives by representing checkout disruption as an ordered vector of stages derived from aggregated data. What is new here is the combination of three elements: checkout loss is formalised as an ordered friction-topology vector instead of a set of descriptive drop-off rates; the model is built entirely from aggregated GA4 funnel exports, with no user-level tracking; and the evidence shows that the diagnostic value of device and acquisition context depends on the funnel stage rather than acting uniformly.

The checkout funnel was treated as a sequence of distinct tasks. The transition from Page View to Add to Cart was interpreted as an entry and relevance filter. The transition from Add to Cart to Shipping Info marked the start of the active checkout process. The transition from Shipping Info to Payment Info was associated with delivery assessment, identity disclosure and form completion. In contrast, the transition from Payment Info to Purchase represented the final stage of trust and confirmation. Consequently, friction was expected to be stage dependent rather than uniformly distributed across the funnel. Fig. 1 illustrates the conceptual framework of the checkout funnel and the corresponding friction points.

Figure 1. Conceptual framework of the checkout funnel and friction points

In supplement e-commerce, the first transition has a specific diagnostic meaning: users must evaluate product relevance, credibility, dosage, price and delivery expectations before expressing purchase intent, so the Page View to Add to Cart transition was treated as an entry-stage filter for relevance and reassurance rather than a purely technical checkout step. Post-cart transitions were expected to reflect a different type of friction, more directly related to form completion, delivery assessment, payment confidence and final confirmation.

H1. In supplement e-commerce, the dominant friction point is expected at the Page View-to-Add-to-Cart transition, before explicit cart intent is expressed.

H2. Device-based friction topology is expected to show greater differentiation after cart entry than at the Page View-to-Add to Cart transition.

H3. The medium-friction topology is expected to be stage-dependent, with greater differentiation at the entry and late checkout transitions than at the Add to Cart to Shipping Info transition.

3 Methodology

3.1 Data and measurement context

Two GA4 Funnel exports from an e-commerce property owned by an established supplement brand were analysed. The observation period covered January 3, 2026, to April 2, 2026. The first export was segmented by device category, while the second export was segmented by the first user medium. Supplement e-commerce is understood here as the online retailing of dietary supplements (vitamins, minerals, sports nutrition, and other over-the-counter wellness products) through a brand-owned web shop. The analysed property reflects a Serbian supplement e-commerce context, most relevant to customers in Serbia and the wider region. Five funnel stages were retained in the analysis: Page View, Add to Cart, Shipping Info, Payment Info, and Purchase. These stages correspond to GA4 e-commerce funnel measurement logic, which uses add-to-cart, shipping, payment and purchase events to assess transaction progress (Google, 2026a). The operational definitions of the five stages are summarised in Table 1.

Table 1. Operational definition of funnel stages

Funnel stage	GA4 event / step	Diagnostic meaning
Page View	page_view	Entry exposure and product/landing-page relevance
Add to Cart	add_to_cart	Explicit cart intent
Shipping Info	add_shipping_info	Delivery, identity, and form-entry threshold
Payment Info	add_payment_info	Payment readiness and trust threshold
Purchase	purchase	Final transaction completion

The analysis was conducted as a single-property diagnostic study. It was not intended to provide a universal benchmark for the supplement e-commerce sector. Exported active-user counts were used to compute transition completion and transition friction across various stages and contexts. Minor segments retained in the source tables were not heavily emphasised when user numbers were too low for reliable interpretation. All case data were anonymised at the reporting level. No personally identifiable information was used, and no individual browsing history was reconstructed. The analysis relied on active-user counts exported from the analytics interface, so the model should be treated as a decision-support representation rather than as a behavioural trace of identifiable users.

The analysis focused on the transition between adjacent funnel stages as the primary unit of measurement, because a total conversion rate collapses the funnel into a single summary outcome. By measuring transitions, the specific points of loss are retained, and each device or medium context can be represented as a topology vector that captures the distribution of loss across the funnel.

3.2 Friction topology formulas

Let $n_{i,c}$ be the number of active users observed at stage i for context c. The transition completion rate $r_i(c)$ was calculated as the ratio between the following and the current stage.

$$r_i(c) = \frac{n_{i+1,c}}{n_{i,c}}, \quad n_{i,c} > 0 \tag{1}$$

Transition friction was defined as the complement of the completion rate, calculated as one minus the proportion of users who progressed to the next stage.

$$F_i(c) = 1 - r_i(c) \tag{2}$$

Friction topology for each context was represented as a vector of ordered stage-level friction values.

$$T(c) = \langle F_1(c), F_2(c), F_3(c), F_4(c) \rangle \tag{3}$$

Context differences were represented as transition-specific contrasts between two topology vectors.

$$\Delta_i(a, b) = F_i(a) - F_i(b) \tag{4}$$

The transitions consisted of Page View to Add to Cart, Add to Cart to Shipping Info, Shipping Info to Payment Info, and Payment Info to Purchase. Formulas (1) and (2) follow the standard definitions of step-completion and drop-off rates used in funnel reporting (Google, 2026b). Formulas (3) and (4) were introduced by the authors: the topology vector extends the drop-off rate into an ordered, comparable representation of the whole funnel, and the contrast in (4) operationalises context differences at the level of individual transitions.

3.3 Hypothesis testing and diagnostic criteria

Each hypothesis was evaluated against the structure of the topology vectors and supported by transition-level statistics. H1 was considered supported if the Page View to Add to Cart transition showed the highest friction value in the overall topology vector. H2 was considered supported if device-based contrasts

(formula 4) and the associated chi-square diagnostics were larger at post-cart transitions than at the entry transition. H3 was considered supported if medium-based differentiation appeared at the entry and late checkout transitions but not at the Add to Cart to Shipping Info transition. For each transition, a chi-square test of independence was performed on the counts of users who did and did not progress, with Cramer's V reported as the effect size and a significance level of $\alpha = .05$.

The diagnostic statistics were interpreted with caution. Large sample sizes can produce statistically significant differences even when the practical effect is minimal, while small segments may yield unstable rates. Cramer's V values of approximately .10, .30, and .50 were used as rough guides for weak, moderate, and strong associations, in line with recent recommendations to avoid rigid universal benchmarks. Automated statistics and dashboards can likewise overstate the importance of small differences when sample sizes are large, and understate meaningful patterns when segments are small (Gandhi et al., 2024). Accordingly, chi-square tests and Cramer's V served as supplementary indicators rather than causal evidence, while the main interpretation relied on the overall pattern of the friction topology.

4 Results

4.1 Overall checkout topology

For brevity, the following stage abbreviations are used in Tables 2 to 5: Page View (PV), Add to Cart (ATC), Shipping Info (SI), Payment Info (PI) and Purchase (PUR). The total funnel consisted of 154,639 active users at Page View, 3,395 at Add to Cart, 1,045 at Shipping Info, 452 at Payment Info, and 317 at Purchase. The largest levels of friction occurred before cart entry, as shown in Table 2. Friction was 97.8% from Page View to Add to Cart, compared with 69.2% from Add to Cart to Shipping Info, 56.7% from Shipping Info to Payment Info, and 29.9% from Payment Info to Purchase.

H1 was supported. The main friction peak appeared at the Page View to Add to Cart transition, before users expressed cart intent. This is significant because in supplement e-commerce, this transition filters relevance and trust factors like credibility, dosage, price, delivery, and brand before adding to the cart. Thus, the largest friction peak isn't a checkout issue, as it happens before users start the checkout process.

Table 2. Overall checkout topology

Transition	Start	Next	Completion	Friction
PV-ATC	154,639	3,395	2.2%	97.8%
ATC-SI	3,395	1,045	30.8%	69.2%
SI-PI	1,045	452	43.3%	56.7%
PI-PUR	452	317	70.1%	29.9%

4.2 Device-based topology

The device topology indicated that mobile and desktop friction at the first transition was almost identical (Table 3, Fig. 2). The mobile friction from Page View to Add to Cart was 97.8%, and the desktop friction was 97.7%. The topology changed after the cart entry. On desktop, friction was markedly higher from Add to Cart to Shipping Info, but much lower from Shipping Info to Payment Info and from Payment Info to Purchase. At the final transition, desktop friction was 10.5%, while mobile friction was 32.7%.

H2 was supported. Device category explained little of the entry stage peak; however, it became diagnostically meaningful after cart entry. The notable disparity in end-stage friction suggests that the final checkout transition should be inspected more carefully for mobile users. This pattern is best treated as a diagnostic signal, not as evidence of a specific causal mechanism. Future user experience (UX) testing would be needed to determine whether the observed difference is related to interface design, payment-gateway behaviour, perceived security, screen size, or other unobserved factors. Tablet observations were excluded because only three users reached the Shipping Info stage, and none progressed further; therefore, the resulting 100.0% SI-PI friction was treated as a small-segment artefact.

Table 3. Device-based friction topology

Device	PV-ATC	ATC-SI	SI-PI	PI-PUR
Total	97.8%	69.2%	56.7%	29.9%
Mobile	97.8%	68.6%	58.3%	32.7%
Desktop	97.7%	73.6%	39.4%	10.5%

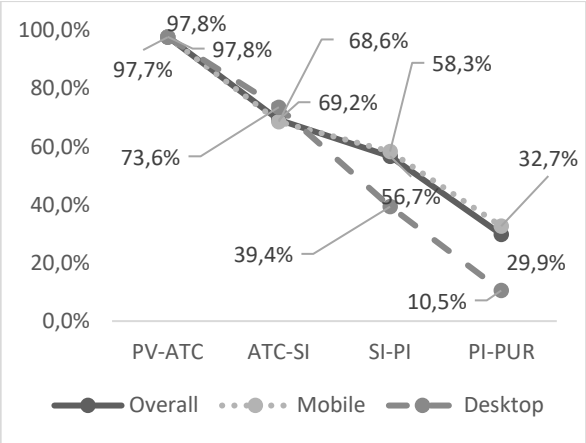


Figure 2. Device-based friction topology vectors

4.3 Medium-based topology

The medium-based topology was more heterogeneous (Table 4). Direct and email traffic had lower entry friction than most other media, while organic and cost-per-click (CPC) traffic had higher Page View-to-Add-to-Cart friction. However, at the final stage, CPC exhibited comparatively low Payment Info-to-

Purchase friction, while paid-social and social showed higher late-stage loss. The acquisition context affects both the readiness to add items to the cart and the willingness to complete later checkout tasks.

H3 was supported with a qualification. Medium-based differences were detected at the Page View to Add to Cart transition and at the late checkout transitions, especially Shipping Info to Payment Info and Payment Info to Purchase. However, the Add to Cart to Shipping Info transition was not clearly differentiated by the first user medium. Acquisition context does not influence all checkout transitions equally. The transition to shipping may represent a high-demand checkout threshold, where delivery disclosure, data-entry effort, and commitment to proceed create friction across media, regardless of the initial acquisition source.

The first user medium is better treated as a stage-sensitive context signal than as a fixed effect of the traffic source. Differences at the entry stage may reflect variation in relevance, audience qualification, or readiness to add items to the cart. In contrast, late-stage differences may indicate varying levels of reassurance required for delivery, payment, or final commitment. These interpretations are diagnostic rather than causal, because the GA4 exports do not include variables such as ad creative, product details, basket value, or user motivation.

Table 4. Selected medium-based friction topology

Medium	PV-ATC	ATC-SI	SI-PI	PI-PUR
paid-social	97.8%	68.2%	68.1%	46.0%
organic	98.9%	69.8%	51.5%	22.2%
cpc	98.3%	65.3%	57.4%	14.3%
direct	95.1%	72.9%	44.9%	24.0%
referral	96.2%	69.4%	57.6%	35.7%
email	94.4%	72.5%	25.0%	23.8%
social	92.1%	65.8%	44.0%	50.0%

4.4 Transition-level diagnostics

Transition diagnostics were performed to determine whether these visual topology differences were also visible in the transition tables (Table 5). Device differences were not highlighted at Page View to Add to Cart, but were present at Shipping Info to Payment Info and Payment Info to Purchase. For the selected media, differences were detected in Page View to Add to Cart and in late checkout transitions. Context effects were stage-specific rather than homogeneous across the funnel. Most Cramer's V values were weak, even when the corresponding chi-square tests were statistically significant.

The diagnostics in Table 5 can be read against the hypotheses. The absence of a significant device difference at PV-ATC ($p = .232$, $V = .003$) confirms that the entry-stage peak is not a device phenomenon, in line with H2. Device associations at SI-PI ($V = .106$) and PI-PUR ($V = .153$) are statistically significant but weak: device category shifts post-cart completion rates in a consistent direction without dominating them. Medium-based associations are strongest at the late transitions ($V = .220$ and $.261$, approaching moderate strength) and absent at ATC-SI ($p = .089$), which matches the stage-dependent pattern proposed in H3. Taken together, context effects are real but modest, and they concentrate after cart entry rather than spreading uniformly across the funnel.

Table 5. Transition diagnostics used as supporting evidence. n.s. = not significant.

Context	Step	p	Cramer V	Note
Mobile-desktop	PV-ATC	.232	.003	n.s.
Mobile-desktop	SI-PI	<.001	.106	weak
Mobile-desktop	PI-PUR	.001	.153	weak
Selected media	PV-ATC	<.001	.080	very weak
Selected media	ATC-SI	.089	.054	n.s.
Selected media	SI-PI	<.001	.220	weak / moderate
Selected media	PI-PUR	<.001	.261	weak / moderate

5 Discussion

5.1 Conceptual and methodological contribution

All three hypotheses were supported, with one qualified. H1 was supported because the Page View to Add to Cart transition showed by far the highest friction (97.8%), reflecting an entry-stage relevance and reassurance filter. H2 was supported because device differences were absent at entry but significant and directionally consistent after cart entry, where interaction cost becomes task-relevant. H3 was supported with a qualification: medium-based differentiation appeared at the entry and late transitions but not at Add to Cart to Shipping Info, suggesting that the shipping threshold imposes a uniform demand across acquisition contexts. The paper's main scientific contribution is modelling supplement checkout loss as a friction topology. The point is not simply that friction is unevenly distributed across stages; it is that loss can be represented as an ordered pattern in which early and post-cart friction carry different diagnostic meanings. The topology preserves three properties that are usually weakened in descriptive funnel reporting — the location, sequence, and context of loss — so a segment is represented not by a single conversion outcome but by a vector showing how loss is distributed across the purchase path.

There is also a methodological contribution. The model can be derived from aggregated GA4 Funnel exploration exports, so it suits settings without user-level tracking, BigQuery exports or experimental arrangements — particularly small and medium-sized firms that rely on standard GA4 interfaces but still need to locate stage-specific checkout disruption.

Empirically, the case clarifies where the model is most informative. The largest loss occurred before cart entry, while the most context-sensitive differences

appeared after it. This matters because supplement purchases involve product credibility, dosage reassurance, price sensitivity, delivery reliability and payment confidence at different points. Early loss may therefore signal a relevance or reassurance filter, whereas late loss may reflect interaction cost, delivery assessment, perceived risk, payment hesitation or difficulty with final confirmation (Mittal, 2025; Roy & Shaikh, 2024). A single conversion rate would not preserve this distinction. Several earlier studies point in the same direction. The concentration of loss before explicit commitment is consistent with Kukar-Kinney et al. (2022), who found that hesitation and deliberation drive abandonment before checkout tasks begin, and with Mittal (2025), who positions early-journey evaluation as a distinct psychological phase. The higher mobile friction after cart entry mirrors Almeida Lucas et al. (2023), who reported lower task ease on mobile platforms during form-intensive steps. The stage-dependent role of acquisition context refines the journey-level perspective of Koronaki et al. (2023) and Padilla et al. (2025): its diagnostic value varies by funnel position rather than acting as a stable segment effect. The contribution relative to these studies is methodological as much as empirical, because the same stage-sensitive reading is obtained here from aggregated export data rather than from clickstream or survey data.

The medium-based results refine the role of acquisition context: the first user medium did not act as a constant segment effect, and its diagnostic value changed by stage. Entry-stage differences suggest that acquisition context shapes initial relevance and readiness to enter the cart, while late-stage differences point to payment confidence, delivery evaluation and final commitment and do not behave uniformly across media. The weak differentiation at the Add to Cart to Shipping Info transition suggests that some checkout thresholds may be overly demanding across contexts.

5.2 Managerial use and practical implications

For supplement e-commerce merchants, friction topology can be used in recurring analytics reports to track where checkout losses concentrate over time. A weekly or monthly comparison of topology vectors can show whether friction is increasing at the entry, shipping, payment, or final confirmation stage. For UX teams, this helps prioritise what to inspect first: mobile usability, form complexity, delivery clarity, checkout transparency, or payment-stage confidence (Ho et al., 2022; Saha et al., 2023). In terms of the customer journey, the model does not redesign the journey itself; it shows where it breaks down. By identifying the stage at which most users are lost, merchants can direct interventions at the actual point of disruption rather than reworking the entire funnel, and each diagnostic cycle shortens the path from observed loss to targeted improvement.

The medium topology can separate traffic quality from checkout experience for marketing teams (Koronaki et al., 2023). When entry friction is high within a medium, managers should first examine landing-page relevance, campaign-message congruence and audience qualification; when friction is concentrated in the final stages, they should focus on payment reassurance, price transparency, shipping disclosure and trust cues (Mittal, 2025; Roy & Shaikh, 2024). The model can thus serve as a low-cost prioritisation tool before more expensive testing begins.

To make the model's diagnostic value more actionable, Table 6 translates the observed friction-topology patterns into practical recommendations for supplement e-commerce merchants.

Table 6. Practical recommendations for supplement e-commerce merchants

Observed friction pattern	Diagnostic interpretation	Recommended merchant action
High Page View to Add to Cart friction	Users are not sufficiently convinced before expressing cart intent	Improve product-page relevance, dosage clarity, benefit communication, reviews, trust badges and delivery visibility before cart entry
Similar entry friction across devices	Device category does not explain the main pre-cart loss	Focus first on product relevance, offer clarity and reassurance rather than device-specific redesign at the entry stage
Higher mobile friction after cart entry	Mobile users may face greater interaction cost, payment hesitation or confirmation difficulty	Audit mobile checkout forms, button visibility, loading speed, payment gateway behaviour and final confirmation screens
High Add to Cart to Shipping Info friction	Users hesitate when checkout requires identity, delivery or commitment information	Reduce form burden, show delivery options earlier, clarify shipping costs and allow guest checkout where possible
High Shipping Info to Payment Info friction	Delivery conditions or form completion may create resistance	Improve delivery transparency, validate forms in real time, reduce unnecessary fields and show estimated delivery time clearly
High Payment Info to Purchase friction	Final-stage trust, payment confidence or unexpected cost issues may block completion	Strengthen payment reassurance, display total cost before payment, highlight secure payment methods and check for payment-gateway errors
Medium-specific entry friction	Some traffic sources may bring less qualified or less aligned users	Review campaign-message congruence, landing-page fit, audience targeting and product-offer alignment
Medium-specific late-stage friction	Users from some channels may need more reassurance before final purchase	Adapt checkout reassurance, retargeting messages and trust cues by acquisition context

These recommendations should not be interpreted as causal evidence of specific checkout problems, but

rather as a prioritisation framework for deciding which checkout stages to audit first.

5.3 Limitations

Several limitations should be acknowledged. First, the analysis relied on a single GA4 property, one established supplement brand and one observation window, so the findings are case-based rather than a sector-wide benchmark. Second, only aggregated counts were available; user-level paths, repeated visits, timing between steps, product categories, basket values, coupon usage, stock availability, shipping costs, payment errors and form-error events were not included. Third, the first user medium and device category followed GA4 breakdown logic, so users who changed devices or channels mid-journey could not be re-attributed in the exported tables. Fourth, the model did not explicitly incorporate psychological constructs.

Cognitive load, perceived ease of use, perceived risk, and trust were employed solely as theoretical interpretative frameworks, not as measured variables. Fifth, no causal inference was attempted; the observed differences identify where friction concentrates, not why it occurs. Future research should integrate friction topology with GA4 BigQuery exports, session-level timestamps, product-level variables, form-error logging, and controlled checkout experiments. In practice, the model should also be used alongside event audits, tag validation, and regular comparisons with order-management records. At the same time, these limitations concern the empirical case rather than the method itself. The model itself is not tied to this property: formulas (1) to (4) require only stage-level user counts so that the same diagnostic can be applied to any e-commerce funnel with ordered stages, including other product categories, other markets, differently defined funnels, and analytics platforms other than GA4. Multi-brand and multi-sector replication would establish whether the entry-stage friction peak and the post-cart context sensitivity observed here are sector-typical or property-specific.

6 Conclusion

This paper develops a friction-topology model for GA4 checkout export data from a well-known supplement e-commerce brand. Instead of a single conversion rate, aggregated counts were used to construct transition-loss vectors, allowing precise localisation of drop-offs by stage, device, and acquisition channel. Across the three hypotheses, H1 and H2 were supported, while H3 was supported with a qualification regarding the Add to Cart-to-Shipping Info transition.

The largest friction peak occurs before cart entry, while differences across devices and media become clearly visible only after users add products to the cart. The model can be implemented directly from standard GA4 Funnel exploration exports, without user-level tracking or BigQuery access, making it practical for analysts, UX, and marketing teams working with routine GA4 data.

Further work should test the model across multiple brands, longer observation periods, and more granular event-level data. Future validation should add event-level variables such as basket value, promotion type, payment method and form errors.

Acknowledgements

The authors employed Large Language Models (Gemini Plus and ChatGPT) solely to assist with language editing and improving the readability of the manuscript. The final content has been fully reviewed and approved by the authors, who assume full responsibility for the integrity of the work.

References

Almeida Lucas, G., Lunardi, G. L., & Bittencourt Dolci, D. (2023). From e-commerce to m-commerce: An analysis of the user's experience with different access platforms. Electronic Commerce Research and Applications, 58, 101240. https://doi.org/10.1016/j.elerap.2023.101240

Chopra, I. P., Jebarajakirthy, C., Jain, T., & Maseeh, H. I. (2024). Electronic shopping cart abandonment: What do we know and where should we be heading? Electronic Markets, 34(1), 25. https://doi.org/10.1007/s12525-024-00697-0

Das Sarma, A., Sarkar, J. G., & Sarkar, A. (2025). E-tail format, cognitive orientation and device: Shaping variety's impact on online cart abandonment. Journal of Consumer Marketing, 42(1), 72–92. https://doi.org/10.1108/JCM-01-2024-6537

Fang, Y.-H., & Li, C.-Y. (2025). Avoiding customer cross-channel behavior in omnichannel retailing: An uncertainty reduction theory perspective. Electronic Markets, 35(1), 49. https://doi.org/10.1007/s12525-025-00794-8

Gandhi, L., Manning, B. S., & Duckworth, A. L. (2024). Effect Size Magnification: No Variable Is as Important as the One You're Thinking About—While You're Thinking About It. Current Directions in Psychological Science, 33(6), 347–354. https://doi.org/10.1177/09637214241268222

Google. (2026a). Ecommerce setup Q&A. In Google Analytics Help. https://support.google.com/analytics/answer/14143583

Google. (2026b). [GA4] Funnel exploration. In Google Analytics Help.

https://support.google.com/analytics/answer/9327974

Ho, X. H., Nguyen, D. P., Cheng, J. M. S., & Le, A. N. H. (2022). Customer engagement in the context of retail mobile apps: A contingency model integrating spatial presence experience and its drivers. Journal of Retailing and Consumer Services, 66, 102950. https://doi.org/10.1016/j.jretconser.2022.102950

Jabr, W., Ghoshal, A., Cheng, Y., & Pavlou, P. (2023). Maximizing Online Revisiting and Purchasing: A Clickstream-Based Approach to Enhancing Customer Lifetime Value. Journal of Management Information Systems, 40(2), 470–502. https://doi.org/10.1080/07421222.2023.2196778

Kim, S., Shin, W., & Kim, H.-W. (2024). Predicting online customer purchase: The integration of customer characteristics and browsing patterns. Decision Support Systems, 177, 114105. https://doi.org/10.1016/j.dss.2023.114105

Koronaki, E., Vlachvei, A., & Panopoulos, A. (2023). Managing the online customer experience and subsequent consumer responses across the customer journey: A review and future research agenda. Electronic Commerce Research and Applications, 58, 101242. https://doi.org/10.1016/j.elerap.2023.101242

Kukar-Kinney, M., Scheinbaum, A. C., Orimoloye, L. O., Carlson, J. R., & He, H. (2022). A model of online shopping cart abandonment: Evidence from e-tail clickstream data. Journal of the Academy of Marketing Science, 50(5), 961–980. https://doi.org/10.1007/s11747-022-00857-8

Liu, Y., & Mensah, I. K. (2024). Using an extended technology acceptance model to understand the behavioral adoption intention of mobile commerce among Chinese consumers. Information Development, 02666669241277153. https://doi.org/10.1177/02666669241277153

Mittal, B. (2025). The psychology of online shopping cart abandonment: A scrutiny of the current research framework and building an improved model of the online shopper journey. Electronic Commerce Research, 25(2), 777–803. https://doi.org/10.1007/s10660-022-09667-0

Moliner, M. A., & Tortosa-Edo, V. (2024). Multirooming: Generating e-satisfaction throughout omnichannel consumer journey design and online customer experience. Journal of Research in Interactive Marketing, 18(3), 349–369. https://doi.org/10.1108/JRIM-05-2023-0149

Padilla, N., Ascarza, E., & Netzer, O. (2025). The customer journey as a source of information. Quantitative Marketing and Economics, 23(3), 379–418. https://doi.org/10.1007/s11129-024-09287-y

Roy, R., & Shaikh, A. (2024). The impact of online consumer review confusion on online shopping cart abandonment: A mediating role of perceived risk and moderating role of mindfulness. Journal of Retailing and Consumer Services, 81, 103941. https://doi.org/10.1016/j.jretconser.2024.103941

Saha, S. K., Duarte, P., Silva, S. C., & Zhuang, G. (2023). The Role of Online Experience in the Relationship Between Service Convenience and Future Purchase Intentions. Journal of Internet Commerce, 22(2), 244–271. https://doi.org/10.1080/15332861.2022.2045767

Unal, M., & Park, Y.-H. (2023). Fewer Clicks, More Purchases. Management Science, 69(12), 7317–7334. https://doi.org/10.1287/mnsc.2023.4716

Virtual Collaboration in Higher Education: Cross-Cultural Competences and Economic Implications

Klara Šilec, Violeta Vidaček-Hainš, Zrinka Lacković Vincek
University of Zagreb
Faculty of Organization and Informatics
Pavlinska 2, 42000 Varaždin, Croatia
{ksilec, vvidacek, zlackovi}@foi.unizg.hr

Abstract. *This paper examines virtual collaboration between students of the University of Zagreb Faculty of Organization and Informatics (FOI) and the Carl H. Lindner College of Business, University of Cincinnati, based on a bilateral university agreement. The activity was implemented as a voluntary activity within the elective course Strategic Human Resource Management in the second year of the graduate study programme Economics of Entrepreneurship. The study is based on a qualitative content analysis of student reflections collected across three academic cohorts, comprising 74 FOI students and 62 University of Cincinnati students, for a total of 136 students. The analysis focused on students' use of digital tools to support intercultural interaction and collaboration. The results show that digital tools enabled communication and teamwork, while the activity contributed to intercultural learning, greater confidence in English communication, personal reflection and the development of relevant digital and collaborative competences that support participation in the global labour market and the digital economy*.

Keywords. internationalization of higher education, human capital development, digital economy, cross-cultural management, virtual collaboration, global workforce competencies

1 Introduction

In the contemporary era, marked by rapid technological development and global interconnectedness, higher education is undergoing a profound transformation. This transformation is closely linked to the changing requirements of the global labour market and the increasing demand for digitally skilled and internationally competent human capital. Higher education institutions are faced with the imperative of providing global, intercultural, and collaborative experiences that enable students to solve complex problems in diverse sociocultural contexts (Gutiérrez-González et al., 2023). Traditional forms of physical student mobility, such as the Erasmus programme, although highly valuable, are often limited by financial barriers, organizational challenges, and, as the recent pandemic demonstrated, global health crises (Demaj & Alushllari, 2023; Polyakova & Galstyan-Sargsyan, 2021), which also creates inequalities in access to international experience and consequently affects the distribution of human capital. In this context, the concept of Industry 4.0 provides a suitable framework for implementing international experiences from home, contributing to the goals of environmental sustainability and the global accessibility of education (Gutiérrez-González et al., 2023).

As a response to these challenges, Collaborative Online International Learning (COIL) has emerged as an innovative and inclusive strategy of internationalization (Liu, 2023), with increasing relevance for the development of competencies required in international business and knowledge-based economies. COIL, developed by Jon Rubin, represents a model that connects classrooms of two or more universities in different countries, enabling students and teachers to work jointly on projects within their regular courses (Huertas-Abril & Palacios-Hidalgo, 2024; Polyakova & Galstyan-Sargsyan, 2021, Rubin, 2016). The primary advantage of this model is the use of digital technologies as a fundamental tool for pedagogical interaction, thereby opening global opportunities within the curriculum that were previously limited (Borger, 2022). Virtual collaboration projects have also proven effective in developing global awareness and encouraging students to engage with the Sustainable Development Goals (Jansen Van Vuuren et al., 2025).

A key benefit of this approach is that, within COIL, cultural differences are not diminished by online collaboration but become more apparent through virtual exchange. Therefore, the development of intercultural communicative competence is at the core of COIL initiatives (Borger, 2022). To interpret the issues and interaction patterns within this type of collaboration, researchers commonly use well-established theoretical approaches, particularly Hofstede's cultural dimensions framework, originally developed in Culture's Consequences (Hofstede, 1980) and later expanded in subsequent work (Hofstede et al.,

2010). Dimensions such as power distance and individualism versus collectivism can help explain potential misunderstandings in digitally mediated intercultural collaboration, especially when students work across different cultural, institutional, and communication contexts (Gruber & Bailey, 2021, Hofstede, 1980; Hofstede et al., 2010). This paper focuses on the synergy between digital technology, the COIL methodology, and Hofstede's cultural dimensions framework, with particular attention to their role in fostering students' human capital and capacity for effective collaboration in international and culturally diverse contexts (Rubin, 2016; Hofstede et al., 2010). Its aim is to explore how digital technologies enable virtual collaboration and how cultural dimensions contribute to the development of human capital and competencies relevant for the global labour market. The empirical part of the paper is based on qualitative research, specifically an analysis of student reflections written by FOI students in COIL-based virtual collaboration activities conducted across three cohorts, with a different group of students taking part each year. In this way, the paper seeks to gain deeper insight into students' experiences, perceived challenges, advantages, and patterns of virtual intercultural collaboration, while also exploring how such experiences contribute to the development of skills relevant for their future participation in the labour market.

The paper is structured as follows: after the theoretical framework, the research methodology is presented, followed by the analysis and interpretation of results, while the final section provides concluding remarks and recommendations for the implementation of virtual activities based on the COIL approach in higher education.

2 Theoretical Framework

This chapter presents the theoretical framework of virtual collaboration in higher education, with an emphasis on the role of digital technologies and the COIL approach in connecting students from different geographical and cultural environments. Special attention is given to Hofstede's cultural dimensions as a framework for understanding intercultural differences in virtual teams. The final part of the chapter describes the application of virtual collaboration within the course Strategic Human Resource Management through collaboration with the University of Cincinnati.

2.1 Virtual Collaboration and Digital Technologies

Virtual collaboration in higher education is defined as a process in which students from different geographical locations work together to achieve shared learning goals by using information and communication technologies (Demaj & Alushllari, 2023). Digital technologies have transformed pedagogical interaction, enabling the creation of contemporary learning environments that transcend physical boundaries (Borger, 2022). The key characteristics of COIL include teamwork, problem solving, and the development of intercultural awareness through "learning at home" (Huertas-Abril & Palacios-Hidalgo, 2024; Vahed, 2024). In COIL projects, platforms such as Zoom, Slack, and Skype become primary tools for engaging students in joint projects, promoting interdisciplinary learning and digital literacy (Medina et al., 2025; Vahed, 2024). The use of these tools not only facilitates communication but also supports the development of 21st-century "super skills," reduces anxiety in foreign-language communication, and increases students' confidence (Liu, 2023). Research shows that the design of collaborative activities, including pedagogical strategies and the selection of ICT tools, directly affects the success of online learning experiences (Reyes et al., 2025). Unlike traditional virtual mobility, COIL is integrated into the curriculum, making it cost-effective and accessible to a wider group of students who cannot afford physical travel (Liu, 2023; Polyakova & Galstyan-Sargsyan, 2021). This model not only strengthens intercultural competencies but is also considered a sustainable solution for the internationalization of higher education in the post-pandemic period (Sharma & Panackal, 2025).

2.2 Hofstede's Cultural Dimensions in the Virtual Context

Hofstede's framework, consisting of six dimensions of national culture: power distance, individualism/collectivism, masculinity/femininity, uncertainty avoidance, long-term orientation, and indulgence/restraint, continues to represent a foundation for understanding cultural differences (Fleșeriu et al., 2023; Kunwar, 2021; Hofstede et al., 2010). In the context of virtual collaboration, these dimensions significantly influence team dynamics, which is particularly relevant in multinational companies and globally distributed teams. For example, the dimension of individualism versus collectivism is reflected in the way students approach responsibility within a group (Kunwar, 2021; Zheng et al., 2025). Power distance defines the extent to which inequalities are accepted, which in COIL projects may influence interaction between students and professors or within student pairs from cultures with different hierarchical structures (Kunwar, 2021). From an economic perspective, understanding these dimensions contributes to more effective management of international human resources. Although Hofstede's model has been criticized for its tendency toward generalization, it still provides practical guidelines for navigating complex local challenges and adapting communication strategies in a multicultural digital

environment, which is essential for achieving high-quality outcomes of intercultural learning (Molina González et al., 2026; Fukkink et al., 2024).

2.3 Virtual International Cooperation as a Cost-Effective Mechanism for Developing Human Capital

International educational experiences contribute to graduate employability through the development and strategic use of multiple forms of capital, including human, social, cultural and agentic capital (Dai & Pham, 2024). These insights are particularly relevant for COIL-based learning, as internationally oriented educational environments can strengthen students' preparedness for culturally diverse labour markets by fostering professional skills, intercultural sensitivity, social networks, and career-related agency (Dai & Pham, 2024). Virtual international collaboration supports the development of students' human capital by enhancing cultural intelligence, intercultural communication competence, and preparedness for work in international and culturally diverse professional contexts (Hackett et al., 2023). The relatively lower importance attributed to academic performance suggests that labour market integration is not determined solely by grades or formal educational achievement (Pham et al., 2024). This indicates that graduates' employability depends not only on what they know academically, but also on how effectively they can communicate, collaborate, understand organisational culture, and strategically mobilise different forms of capital in professional contexts (Pham et al., 2024). Given that physical mobility involves only a small proportion of students, approximately 1% globally and around 10% in the United States and the European Union, COIL offers considerable potential as a cost-effective and more inclusive mechanism of internationalisation (Gelashvili & Blanco, 2024).

3 Methodology

The focus of this paper is on the synergy of digital technology, COIL methodology, and Hofstede's framework in the context of developing competencies relevant to contemporary labour markets. The aim is to explore how digital tools facilitate virtual collaboration and how cultural dimensions shape the learning experience in an online international environment. Given the stated objectives, research questions were formulated. Accordingly, the following two research questions were defined:

RQ1: What is the role of digital technologies in virtual collaboration at the level of higher education institutions?

RQ2: How do students perceive the experience of communicating with students from other cultures?

To address these research questions, a qualitative content analysis was conducted on student reflections, with the aim of identifying recurring themes related to digital collaboration, intercultural communication, and the perceived development of relevant skills. Hofstede's framework was used as an interpretative background for discussing intercultural differences rather than as a predefined coding framework.

3.1 Research Procedure

A virtual collaboration activity with COIL elements was implemented within the Strategic Human Resource Management course. Participation was voluntary, and students were organized into international teams. The activity did not constitute a fully integrated COIL model because the partner courses retained separate curricula and learning outcomes. Nevertheless, it incorporated key COIL elements through structured international teamwork, comparative analysis, and digitally mediated interaction. The collaboration is based on working on a shared topic, namely the comparison of cultures, business practices, and human resource management through three virtual meetings. The outcome of the collaboration is a group report containing individual reflections of team members.

For the purposes of this paper, qualitative research was conducted, focusing on the analysis of students' reflections on virtual collaboration with COIL elements with students from the University of Cincinnati in the United States. Content analysis traditionally developed as a quantitative method based on predefined categories, statistical processing of textual data, and an emphasis on reliability and generalizability. As a response to the limitations of such an approach, qualitative content analysis emerged and is described as a flexible method for the systematic coding and interpretation of textual data in order to identify patterns and themes (Serafini & Reid, 2023).

The research was designed as a retrospective study covering reflections created within COIL activities conducted in the course Strategic Human Resource Management over three consecutive academic years, with different groups of students included in the analysis. The course is an elective course and is taught in the second year of the graduate study programme Economics of Entrepreneurship. Data collection and analysis were carried out through qualitative content analysis. Student reflections were read several times, and relevant text segments were assigned initial codes through an inductive coding procedure. The coding was conducted by one researcher. To strengthen the credibility and trustworthiness of the analysis, the procedure was systematic and iterative: similar codes were compared, refined, and grouped into broader thematic categories, while the consistency of the categories was checked by revisiting the original reflection excerpts. Examples of the coding process are

provided in Section 4 to increase analytical transparency.

3.2 Sample Description

The research is based on group papers written by graduate students of Economics of Entrepreneurship at the Faculty of Organization and Informatics (FOI) after their virtual collaboration with students from the University of Cincinnati. The observed activity included three cohorts of students: 74 FOI students and 62 University of Cincinnati students, 136 students in total. The units of analysis were group reflections written by FOI students, which included individual reports and shared impressions of group members. All available and fully submitted papers produced within the observed activity were included in the analysis. As a result of this activity, a total of 19 group papers by students from FOI were produced. The number of group members varied between 6 and 9, with the ratio of FOI and UC students being almost equal, although FOI students were slightly more represented. The demographic analysis refers only to FOI students, since complete demographic data were available for this group. In the academic year 2023/2024, seven groups participated in the activity, with a total of 58 students, of whom 30 were from FOI and 28 from the University of Cincinnati. Among the FOI students, there were 21 domestic students and 9 international students, of whom 7 were from Indonesia, 1 from the Netherlands, and 1 from France. Regarding the gender structure of FOI students, 24 participants were female (80%), while 6 participants were male (20%). In the academic year 2024/2025, seven groups also participated, with a total of 48 students. Of these, 29 students were from FOI, including 18 domestic and 11 international students from Indonesia, while 19 students participated from the University of Cincinnati. Among FOI students, 6 participants were male (20.69%), while 23 participants were female (79.31%). In the academic year 2025/2026, five groups participated in the activity, and the collaboration included 15 FOI students and 15 students from the University of Cincinnati. This cohort did not include any international students among the FOI participants. All 15 FOI participants were female (100%). FOI students represented a relatively homogeneous group in terms of age, with an average age of 23 years and no major deviations among participants. Demographic characteristics of the University of Cincinnati students were not analysed due to incomplete demographic data. However, it was known that they came from different years of study and different study programmes, which made their academic background more heterogeneous than that of the FOI students.

4 Results and Interpretation

This chapter presents the main findings of the study based on the qualitative analysis of students' reflections. Its purpose is to provide insight into how students experienced the international virtual collaboration and which aspects of the activity emerged as most relevant in their reflections.

4.1 Thematic Categories of Student Reflections

Based on the content analysis, eight categories were proposed, encompassing coded units, that is, students' observations arising from the experience of international virtual collaboration. To provide insight into the inductive coding process, selected examples are presented to illustrate how meaning units from students' reflections were transformed into initial codes and then grouped into broader thematic categories. For example, the statement "This virtual collaboration was really valuable for me because it helped me step out of my comfort zone" was initially coded as increased self-confidence and was later grouped under the thematic category Personal Growth and Development. A statement referring to nervousness about speaking English was coded as language-related insecurity and assigned to Communication and Cultural Factors, whereas a statement describing a better understanding of differences in business practices was coded as cultural awareness and assigned to Intercultural Learning.

Table 1 shows that the most represented categories were the technological and technical aspect, intercultural learning, and personal growth and development, while the least represented category was previous experience.

The obtained results suggest that students focused most of their attention, as expected, on the collaboration itself, namely on the use of digital technologies in the process of connecting with students from the United States, on understanding cultural differences and specificities, as well as on their own internal processes, such as increased self-confidence, perceived nervousness before the beginning of the collaboration, and the development of various skills.

Regarding the comparison of frequencies across cohorts, it should be taken into account that two fewer teams participated in the third cohort, which also resulted in fewer individual reflections.

The following part of the chapter briefly presents all identified categories, while the technological and technical aspect and communication and cultural factors are discussed in separate sections. This is because the activity was implemented in a virtual environment, where digital tools played a central role in enabling collaboration. Communication and cultural factors were also examined separately because they directly reflect the intercultural nature of the activity and its relevance for students' communication

confidence and future participation in international professional contexts.

Intercultural learning refers to students' observations related to learning about cultures, business practices, and approaches to human resource management, as well as understanding cultural similarities and differences among Croatian, American, and Indonesian cultural frameworks. The results suggest that the collaboration encouraged the development of intercultural understanding through the comparison of different cultural patterns, professional approaches, and opportunities for connecting based on observed similarities. Although the categories of communication and cultural factors and intercultural learning are conceptually related, they were treated separately in the analysis. Communication and cultural factors refer to concrete aspects of communication, such as English-language expression, language-related insecurity, and perceived cultural barriers during interaction. Intercultural learning, on the other hand, refers to broader developmental outcomes, including cultural awareness, openness toward other cultures, broadening of perspectives, and understanding differences in HRM practices.

The category of personal growth and development shows that participation in the project contributed not only to academic learning, but also to the development of competencies relevant to future employment. Students' reflections indicate a gradual increase in communication confidence, reduced anxiety in intercultural interaction, and stronger flexibility, adaptability, organizational skills, and cultural sensitivity. These competencies are important for functioning in multicultural workplaces, participating in global virtual teams, and addressing future business problems through collaboration and effective communication. This is consistent with Liu (2023), who found that digital communication tools can reduce anxiety in foreign-language communication, as well as with employability research emphasizing communication, teamwork, adaptability, and problem-solving as important labour market competencies (Pham et al., 2024; Tang et al., 2024).

The experience of international virtual collaboration was generally perceived positively, with students expressing satisfaction, gratitude, and interest in repeating a similar activity. However, they also recognized challenges such as unequal engagement and insufficient preparedness of some participants, which affected the dynamics of collaboration.

The organizational aspects of collaboration, primarily schedule coordination and dealing with time differences, were included in the category of time management. Both negative and neutral statements were present in the reflections: some students reported difficulties in coordinating meeting times, while others emphasized that there were no significant problems related to time organization. These findings show that time coordination represents an important, but not necessarily equally challenging, issue for all participants. At the same time, they point to the importance of planning, flexibility, and agreement in virtual international projects. This is relevant from an employability perspective, as Pham et al. (2024) emphasize that professional skills and personal qualities are important for effective functioning after entering the labour market.

A less represented but still important category is networking and career development, which includes students' observations related to making new acquaintances, recognizing the importance of global collaboration in the educational process, practical preparation for the beginning of a career, and gaining insight into business opportunities and the work environment. The results indicate that students recognized international virtual collaboration not only as an educational experience, but also as an opportunity for professional networking and the initial development of career-related competencies.

Table 1. Frequency of thematic categories across cohorts
Source: Authors' own analysis of student reflections

Category	**Absolute Frequencies**			**TOTAL**
	Academic Year			
	23/24	**24/25**	**25/26**	
Technological and Technical Aspect	78	65	38	**181 (31.81%)**
Intercultural Learning	45	59	26	**130 (22.84%)**
Personal Growth and Development	21	35	21	**77 (13.53%)**
Overall Experience and Challenges	18	29	19	**66 (11.6%)**
Time Management	18	15	3	**36 (6.33%)**
Communication and Cultural Factors	17	9	7	**33 (5.8%)**
Networking and Career Development	16	11	2	**29 (5.1%)**
Previous Experience	8	7	2	**17 (2.99%)**
TOTAL	**221**	**230**	**118**	**569 (100%)**

The least represented category was previous experience, indicating that prior international or virtual collaboration was not a central focus of student reflections. For some students, the project was their first experience of this kind, which may explain their stronger emphasis on communication challenges, adaptation, and personal development.

4.2 The Role of Digitalization in the Internationalization of Teaching Activities

In the observed activity, students engaged in virtual collaboration with COIL elements by using various information and communication tools, among which email emerged as the most frequently used tool. Although previous research on COIL projects highlights platforms such as Zoom, Slack, and Skype as important tools for student engagement, collaboration, and the development of digital literacy (Medina et al., 2025; Vahed, 2024), the findings of this research indicate a different pattern of digital communication tool use.

As shown in Fig. 1, Zoom, Google Meet, and WhatsApp appear most prominently in the participants' reflections, while Slack and Skype were not identified as relevant communication channels within the analysed project experience. In addition, Google Docs also played an important role as a supplementary tool supporting meeting organization and collaborative work on content. The use of these tools enabled students to maintain continuous contact with peers from the partner institution and to effectively overcome spatial barriers.

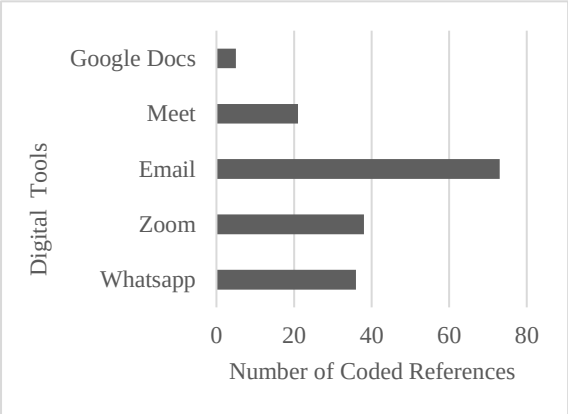


Figure 1. Digital tools used for communication and collaboration (N=173 coded references)
Source: Authors' own analysis of student reflections

Students' impressions indicate that digital tools had both functional and pedagogical value, as they enabled the organization of communication, the exchange of information, and collaboration on joint tasks, while also contributing to the perceived usefulness of online communication and the development of digital literacy. At the same time, students identified internet-related issues and technical difficulties as the main challenges, which shows that the success of virtual collaboration depends on the quality of technical infrastructure and the stability of digital platforms. Nevertheless, information and communication tools played a key role in the implementation of virtual collaboration with COIL elements, as they served not only as technical means of communication but also as a space for collaboration, idea exchange, and connecting students from different national and institutional contexts.

4.3 Communication and Cultural Factors in Virtual Collaboration

According to the results presented in Fig. 2, improvement in English-language communication skills was the most frequently identified aspect, accounting for 20 of the 33 coded references within this category (61%). Fear of language and cultural barriers accounted for 10 coded references (30%), indicating that some reflections emphasized initial insecurity and nervousness related to communication in a foreign language and interaction across cultures. Lack of English-language proficiency was identified in only three coded references (9%), suggesting that explicitly reported language difficulties were considerably less frequent than concerns or initial discomfort related to communication.

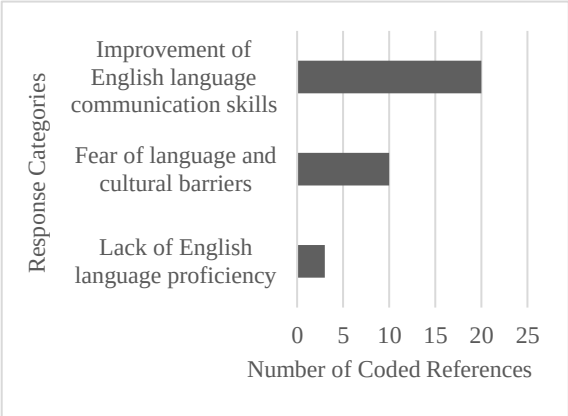


Figure 2. Communication and cultural factors (N=33 coded references)
Source: Authors' own analysis of student reflections

Based on this, it can be concluded that the communication aspects of this collaboration represented both a challenge and a developmental opportunity. On the one hand, students faced a certain level of insecurity related to possible language barriers, while on the other hand, this very experience contributed to the development of their English-language competencies and greater confidence in using English in an international context. This is relevant for students' future employability, as Tang et al. (2024) identify English proficiency as one of the highest-rated contributors to job success and performance, while Pham et al. (2024) also report sufficient English proficiency as the most important factor for obtaining employment among international graduates.

5 Conclusion and Future Work

Overall, the findings indicate that virtual collaboration with COIL elements contributes to the development of economically relevant competencies and strengthens students' readiness for participation in the global and digital labour market. The activity may also be interpreted as a relatively cost-effective approach because it enables international collaboration without the travel and accommodation costs associated with physical mobility. However, the identified economic implications are indirect and interpretative rather than empirically measured financial effects.

In response to the first research question digital tools played a key role in the implementation of the activity, as they enabled communication, information exchange, and joint work among students from different national and institutional contexts, while also overcoming time and geographical barriers. Although students identified technical difficulties, internet connection issues, and scheduling challenges, collaboration and adaptation enabled them to overcome these obstacles while developing competencies relevant to international business and the global labour market.

In response to the second research question, students generally perceived the experience of communicating with students from other cultures positively. The experience strengthened their intercultural understanding, confidence in using English, and communication and organizational skills relevant to multicultural professional environments.

The methodological limitations of the research primarily arise from the way students were included in the examined activity. The course in question is an elective course that is not taken by all students of the cohort, and the virtual collaboration activity within the course was also optional rather than mandatory. Therefore, the number of participants was relatively small, which partially limits the possibility of generalizing the findings to the broader student population.

The limitation regarding generalizability is also related to the fact that this case does not represent the original COIL model, but rather virtual collaboration with COIL elements. Unlike a fully integrated COIL approach, in which the teaching objectives, learning outcomes, and activities of partner courses are more strongly aligned, in this research the syllabi and learning outcomes at FOI and the University of Cincinnati were only partially aligned. This circumstance affects the interpretation of the results, as the findings refer to a specific form of virtual collaboration rather than to a complete COIL model.

An additional limitation concerns the use of content analysis, as the interpretation of student papers may include a certain degree of subjective assessment by the researcher. Although the analysis was conducted systematically, the possibility of interpretative bias cannot be completely excluded.

The obtained results may be useful for courses with similar content, as well as for future cohorts of the course Strategic Human Resource Management, especially in the context of planning and implementing virtual collaborations that can help students develop skills needed for their professional careers. In accordance with the ethical principles of scientific research, all results were processed at the group level and anonymized.

Recommendations for future research include increasing the number of participants and conducting research within the framework of the original COIL model, that is, on courses from different universities and countries that have aligned syllabi and learning outcomes. Future research could also include virtual collaboration within a single course involving three or more different universities, which would enable a comparison of students' experiences across different institutional and cultural contexts. In addition, future studies could apply content analysis to students' reflections or communication records using both digital analytical tools and manual coding conducted by researchers. Comparing the results obtained through these two approaches would make it possible to identify the advantages and limitations of each method. Finally, future research could examine how students who participated in virtual international learning experiences later cope with real international settings, particularly in the context of the international labour market and professional communication.

Acknowledgments

This work was supported by the European Union through the NextGenerationEU programme, under the call for Institutional Research Projects financed from source 581 – Recovery and Resilience Facility (2025), within the project StartUp.Score: Development of a Model for Deciding on the Readiness of an Entrepreneurial Venture for Incubation.

References

Borger, J. G. (2022). Getting to the CoRe of Collaborative Online International Learning (COIL). Frontiers in Education, 7. https://doi.org/10.3389/feduc.2022.987289

Dai, K., & Pham, T. (2024). Graduate employability and international education: an exploration of foreign students' experiences in China. Higher Education Research and Development, 43(6), 1227–1242. https://doi.org/10.1080/07294360.2024.2325155

Demaj, E., & Alushllari, G. (2023). Assessing Virtual Collaborative Learning (VCL) as an Innovative Teaching and Learning Approach: A Case

Application in Albania. International Journal of Management, Knowledge and Learning, 12. https://doi.org/10.53615/2232-5697.12.s49-59

Fleșeriu, C., Lifintsev, D., Jerónimo, N. A., & Tamulevičiūtė-Šekštelienė, A. (2023). Students' Perceptions on YouTube as an Educational Tool in Teaching Cross-Cultural Communication. Pedagogika, 151(3), 105–126. https://doi.org/10.15823/p.2023.151.6

Fukkink, R., Helms, R., Spee, O., Mongelos, A., Bratland, K., & Pedersen, R. (2024). Pedagogical Dimensions and Intercultural Learning Outcomes of COIL: A Review of Studies Published Between 2013–2022. Journal of Studies in International Education (Vol. 28, Number 5, 761–779). https://doi.org/10.1177/10283153241262462

Gelashvili, M., & Blanco, G. (2024). COIL as a Way to Enhance Internationalization: Balancing Evidence and Proliferation. International Higher Education, (116), 28–30. Retrieved from https://www.ejournals.bc.edu/index.php/ihe/article/view/17375

Gruber, A., & Bailey, A. C. (2021). Examining EFL students' pluricultural and plurilingual development during intercultural virtual exchanges between Colombia and Germany. Global Journal of Foreign Language Teaching, 11(1), 14–29. https://doi.org/10.18844/gjflt.v11i1.5148

Gutiérrez-González, S., Coello-Torres, C. E., Cuenca-Romero, L. A., Carpintero, V. C., & Bravo, A. R. (2023). Incorporating Collaborative Online International Learning (COIL) into Common Practices for Architects and Building Engineers. *International Journal of Learning, Teaching and Educational Research*, *22*(2), 20–36. https://doi.org/10.26803/ijlter.22.2.2

Hackett, S., Janssen, J., Beach, P., Perreault, M., Beelen, J., & van Tartwijk, J. (2023). The effectiveness of Collaborative Online International Learning (COIL) on intercultural competence development in higher education. International Journal of Educational Technology in Higher Education, 20(1). https://doi.org/10.1186/s41239-022-00373-3

Hofstede, G. (1980). Culture's consequences: International differences in work-related values. Sage.

Hofstede, G., Hofstede, G. J., & Minkov, M. (2010). Cultures and organizations: Software of the mind: Intercultural cooperation and its importance for survival (3rd ed.). McGraw-Hill.

Huertas-Abril, C. A., & Palacios-Hidalgo, F. J. (2024). Collaborative international online learning for the development of intercultural awareness: an experience with pre-service language teachers. Journal for Multicultural Education, 18(1–2), 67–80. https://doi.org/10.1108/JME-09-2023-0093

Jansen Van Vuuren, W., Heymans, Y., Pool, J., & Strosnider, C. (2025). Virtual Exchange to Develop Global Citizenship and Promote Sustainable Development Goal 3.4. Open Praxis, 17(2), 409–420. https://doi.org/10.55982/openpraxis.17.2.776

Kunwar, J. B. (2021). The Influence of Culture on Collaborative Learning Practices in Higher Education. Journal of Intercultural Management, 13(2), 81–106. https://doi.org/10.2478/joim-2021-0062

Liu, Y. (2023). Overview of the Impact of Collaborative Online International Learning on Learners. SHS Web of Conferences, 157, 04011. https://doi.org/10.1051/shsconf/202315704011

Medina, A. S., Lopez, A. L., Cervantes, M. A. R., Reyna, E., Ortega, K. M., Anota, R., Legarreta, N., Holguin, D. V., & Loera-Valencia, R. (2025). Enhancing Global Competencies: Exploring the Impact of Collaborative Online International Learning (COIL) on Student Wellbeing and Academic Development in Latin American Universities. Proceedings of the 2024 the 16th International Conference on Education Technology and Computers, ICETC 2024, 454–461. https://doi.org/10.1145/3702163.3702455

Molina González, J. A., Lamino, P., & Diaz, J. (2026). Hofstede Cultural Dimensions #1: Introduction. EDIS, 2026(1). https://doi.org/10.32473/EDIS-WC502-2026

Pham, T., Soltani, B., & Singh, J. K. N. (2024). Employability capitals as essential resources for employment obtainment and career sustainability of international graduates. Journal of Further and Higher Education, 48(4), 436–448. https://doi.org/10.1080/0309877X.2024.2344771

Polyakova, O., & Galstyan-Sargsyan, R. (2021). Sustainable Higher Education via Telecollaboration: Improving Plurilingual and Pluricultural Competence. *Integration of Education*, *25*(4), 544–561. https://doi.org/10.15507/1991-9468.105.025.202104.544-561

Reyes, E. R., Arana-Cuenca, A., & Martínez, A. I. M. (2025). Recommendations for Designing Collaborative Activities in Online Higher Education: A Systematic Review. Journal of *Technology and Science Education*, *15*(1), 4–17. https://doi.org/10.3926/jotse.2818

Rubin, J. (2016). The collaborative online international learning network. In R. O'Dowd & T. Lewis (Eds.), *Online intercultural exchange: Policy, pedagogy, practice* (pp. 263–272). Routledge.

Serafini, F., & Reid, S. F. (2023). Multimodal content analysis: expanding analytical approaches to content analysis. *Visual Communication*, *22*(4), 623–649. https://doi.org/10.1177/1470357219864133

Sharma, A., & Panackal, N. (2025). Charting the course of digital collaboration: a bibliometric analysis of COIL literature. Cogent Education, 12(1), 2477369. https://doi.org/10.1080/2331186X.2025.2477369

Tang, M., Pham, T., & Saito, E. (2024). Contributions of Employability Capitals to the Employability of Chinese International Graduates in Australia. Journal of Studies in International Education. https://doi.org/10.1177/10283153241262464

Vahed, A. (2024). Increasing proximities and reducing distances through COIL Virtual Exchanges. The Independent Journal of Teaching and Learning, 19(1), 10–32. https://doi.org/10.17159/ijtl.v19i1.18846

Zheng, F., Zhao, C., Yasmin, F., & Sokolova, M. (2025). Hofstede's cultural dimensions and proactive behavior as the antecedents of entrepreneurial innovativeness. Acta Psychologica, 256. https://doi.org/10.1016/j.actpsy.2025.104948

Artificial Intelligence in the Workplace: A Quantitative Analysis of Employee Attitudes and Readiness

Milanka Dimovska
Faculty of Economics-Skopje, Ss. Cyril and Methodius University in Skopje
Boulevard Goce Delcev 9 v, 1000 Skopje, North Macedonia
milankadimovska@outlook.com

Violeta Cvetkoska
Faculty of Economics-Skopje, Ss. Cyril and Methodius University in Skopje
Boulevard Goce Delcev 9 v, 1000 Skopje, North Macedonia
violeta.cvetkoska@eccf.ukim.edu.mk

Abstract. *The incorporation of artificial intelligence into all segments of modern society is inevitable, and its development will present many opportunities but also challenges that we will need to respond to without avoiding them. Artificial intelligence is affecting the labour market, and some employees fear that artificial intelligence will replace them, but it will be their daily support and assistance in performing work tasks. For this purpose, employees through a series of training programs and acquisition of additional skills, should become familiar with artificial intelligence as soon as possible and accept the benefits of its implementation. This research, conducted through an online anonymous survey, analyses responses based on a sample of 121 respondents from the Republic of North Macedonia and aims to find out how familiar respondents are with artificial intelligence, whether they see it as a threat or as an opportunity for creating new jobs and improving work performance, and how prepared they are to adapt to the changes caused by the application of artificial intelligence in the workplace. The research is presented through an analysis of the demographic and AI-related variables that actually represent the survey questions. Descriptive analysis was conducted to summarize respondents' demographic characteristics and response distributions across the Likert-scale items, followed by Spearman's rank-order correlation analysis in Python to examine relationships between selected AI- related variables. Within this sample, the results show that respondents perceive artificial intelligence in the workplace positively.*

Keywords. artificial intelligence, employee, workplace

1 Introduction

The advancement of artificial intelligence has shown accelerated growth in recent years, and its recent achievements improve prediction (Agrawal, Gans and Golfarb, 2024). The software that is embedded in computer systems allows artificial intelligence to be used in all processes where decisions need to be made because it resembles human-like intelligence (Demirel and Topcu, 2024). When using artificial intelligence, work processes are optimized, work tasks are simplified, and human influence is reduced, which can be used to perform other work tasks (Narang, Vashisht and Bajaj, 2024).

Innovation affects the improvement of economic development in society, and artificial intelligence through the automation of business processes improves the decision-making process and at the same time influences changes in the way of working (Akçetin, 2023). The significant increase in digital services leads to the formation of huge amounts of "big data", which average human intelligence cannot process (Demirel and Topcu, 2024). Automation enables institutions to improve profitability and performance while reducing reliance on human labour (Ris, Stankovic and Avramovic, 2020).

2 Artificial Intelligence in the Workplace

The development of artificial intelligence brings many unanswered questions related to the method of decision-making and choices, which results in the need for each company to make a strategic decision about implementing artificial intelligence in its operations (Daidai and Tamnine, 2023). Decisions made using artificial intelligence, which is increasingly present in business, differ from conventional ideas about management and decision-making by humans because a problem with transparency arises primarily in situations where cooperation and efficient decision-making involving both parties is inevitable, because a person makes a choice and bears responsibility for its actions, while the integration of artificial intelligence into the decision-making process shifts responsibility (Ahmed, 2024). Consequently, no matter how capable a program is to operate and use the most modern technology for decision-making, it still cannot replace human thought, especially for complex and realistic decisions (Al-Mahameed, 2024). An unethical

approach to the use of artificial intelligence can result in an increased risk of cyberattacks and data misuse (Ahmed, 2022).

The human is the most important link in the functioning of companies, so investing in human resources is the main strategy of managers who direct all their capacities towards developing ways to motivate employees and increase their performance, and the most important role in this is played by the development of the reward management system (Güngör, 2011). The connection of big data and machine learning with the human factor is the basis for the proper use of artificial intelligence for employee retention and engagement (Rao, Chitranshi and Punjabi, 2020). The application of artificial intelligence can increase competition and improve innovation, and organizations should focus on using artificial intelligence that will improve the quality of employees' intellectual work and enable their advancement (Kwasek et al., 2024). Although artificial intelligence improves the work of organizations, challenges also arise when it comes to integrating the innovations brought by AI, leading to a more serious approach by organizations in implementing intelligent systems (Olan et al., 2022). In parallel with the development of artificial intelligence, it is essential to embrace change and develop skills, rather than competing with intelligent systems (Bernard, 2024).

The benefits of using artificial intelligence in work processes where repetitive tasks occur have a positive outcome, but if companies incorrectly apply artificial intelligence, there is a risk of more harm than good, which requires an ethical and friendly relationship between the human factor and artificial intelligence (Ahmed, 2022). Whether AI will positively impact or pose a threat to traditional ways of working depends primarily on the organizational capabilities and approach of leaders (Gabelaia, Bagočiūnaitė and Navickienė, 2024). In the traditional hierarchical form of organisation where work tasks are established with a high degree of standardization and compliance, there is a risk that artificial intelligence will be used primarily as a replacement for human labour (Benhamou, 2020). Employees' fear of job loss due to AI and robots can be reduced if employees increase their innovative and creative abilities in dealing with challenges (Rampersad, 2020).

3 Research Problem

The ongoing global changes and the rapid development of society inevitably affect the labour market and workplace conditions. Artificial intelligence is changing the entire life of a person, leaving traces even on the workplace where individuals spend most of the day. Humans as a social being need daily communication and interaction with other people. Artificial intelligence is changing the traditionally established working relationship, which will inevitably lead to the reorganization of existing jobs and the creation of new ones. Therefore, artificial intelligence should be seen as a support in performing work tasks and making better decisions, which leaves room for further growth and development of new skills and knowledge of employees.

Employees in the workplace show different attitudes towards accepting artificial intelligence. The differences that appear depend on many factors that directly influence the attitude of the employee towards changing work tasks brought by modern technology. This research based on a sample of 121 respondents from the Republic of North Macedonia, aims to investigate and analyse the attitudes of respondents regarding their thoughts on the application of artificial intelligence in the workplace, i.e. their attitudes towards adapting to artificial intelligence, readiness and desire to learn new skills.

4 Research Questions

For the purpose of a more detailed analysis, the following research questions were defined within this research:

RQ1: To what extent respondents are familiar with and understand how artificial intelligence functions?

RQ2: Do respondents perceive artificial intelligence as a threat to their jobs or as an opportunity for creating new jobs?

RQ3: Do respondents believe that the application of artificial intelligence can enhance and increase work productivity and efficiency?

RQ4: What is the attitude of respondents towards learning new technologies and tools enabled by artificial intelligence?

RQ5: What is the attitude of respondents towards stress from using artificial intelligence?

5 Methodology

The data for the analysis was obtained through an online survey that was conducted via Google Forms between 18th February and 27th February 2026. The survey consisted of two sections. The first section of the survey focused on respondents' demographic information, such as age group, gender, level of education, field of employment, role/job position and years of work experience. The second section included 14 questions regarding AI awareness and understanding, perceived threats to job, perceived benefits and opportunities of artificial intelligence, stress-related emotional responses and intentions to learn and use AI-powered technologies and tools. The second section of the questionnaire consisted of statements using a five-point Likert scale, with response options ranging from strongly disagree to strongly agree. The questions from both sections were

close-ended, each allowing respondents to select only one answer. The collected survey data were cleaned and prepared for analysis using Power BI. The responses from the second section of the questionnaire were coded into numerical values ranging from 1 to 5, allowing subsequent statistical analysis in Python. The coded data were used to calculate Spearman's rank-order correlation coefficients to examine the relationships between the AI-related variables.

The survey was completed by 123 respondents, but only 121 participants provided full responses which were used for analysis. The survey indicated that the responses were anonymous and that the collected data would be used only to analyse the overall sample, not individual participants

6 Analysis and Interpretation of Results

The analysis begins with overview of the demographic characteristics of the respondents. Respondents' awareness, understanding and opinions on artificial intelligence and its impact on their roles/job positions were analysed regarding the demographic characteristics of the respondents. In addition, an analysis was performed among the second section questions i.e. between the AI related variables in order to explore the relationship between AI knowledge, perceived benefits and threats, preparedness for AI integration and stress-related emotional responses.

6.1 Demographic analysis

The sample consisted of 121 respondents, of whom 78 were female and 43 were male, regarding the gender. The most represented age group was respondents aged 29 to 39 years, comprising 59 participants. In contrast, respondents from the older age groups, particularly those over 50 years of age, were less represented. The majority of respondents (62) held a bachelor's degree, 39 held a master's degree, 8 held a PhD, and the remaining respondents were high school graduates. The respondent's field of employment showed that most of the participants worked in the sectors finance and IT. In terms of work experience 41 respondents reported having between 4 and 8 years of work experience, followed by respondents with 9 to 15 years of experience.

6.2 Analysis of demographic characteristics and AI-related variables

The analysis of familiarity with the concept of artificial intelligence as seen from Fig. 1 below showed that the majority of participants demonstrated a positive level of awareness. When analysed by age group, respondents aged 29-39, reported the highest level of familiarity with artificial intelligence, with 38 respondents agreeing and 10 strongly agreeing.

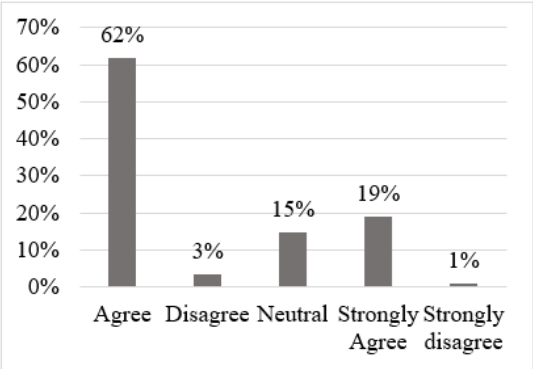


Figure 1. Familiarity with the concept of artificial intelligence

The analysis showed that respondents generally understand how artificial intelligence works with 59% agreeing and 17% strongly agreeing. When analysed across fields of employment, respondents from finance and IT showed the highest level of agreement ("agree" and "strongly agree"). The majority of respondents indicated an agreement with the statement that they understand how AI is used for their job position with 60% agreeing and 12% strongly agreeing. When analysed across fields of employment, respondents from finance and IT showed the highest level of agreement ("agree" and "strongly agree").

The results reported that most respondents didn't perceive artificial intelligence as a threat to their job position, with 47% disagreeing and 7% strongly disagreeing. Analysed by age group, participants aged 29-39 showed the highest level of disagreement, with 26 respondents disagreeing and 4 strongly disagreeing, followed by respondents aged 40-50. On the other hand, analysed by years of work experience, respondents with 4-8 years of experience reported highest level of disagreement, with 24 disagreeing and 4 strongly disagreeing, followed by respondents with 16-30 and 9-15 years of experience.

Out of the 121 respondents 40% of them disagreed and 5% strongly disagreed that artificial intelligence will reduce employment opportunities for their job position in the future. When analysed by field of employment, respondents working in finance and IT recorded highest representation in both "agree" and "disagree" responses regarding whether artificial intelligence will reduce employment opportunities for their role in the future. For the statement that artificial intelligence will create new job opportunities in the future, the agreement was notable among respondents working in the fields of finance and IT. Overall, 50% of the respondents agreed and 18% strongly agreed that AI will create new job opportunities in the future.

A total of 55% of the respondents disagreed and 9% strongly disagreed with the belief that their role/job position will be replaced by AI in the future. When

analysed by years of work of experience, respondents with 4-8 years of work experience reported the highest level of disagreement with the statement that their role/job position will be replaced by artificial intelligence in the future. In contrast agreement with the statement was relatively low across all groups indicating limited concern that artificial intelligence will replace respondent jobs in the future.

More than half of the respondents indicated a positive level (50% agreeing and 23% strongly agreeing) of preparedness to adapt to the changes caused by the application and use of artificial intelligence in the workplace. When analysed by level of education, respondents holding a Bachelor's degree showed the highest level of agreement with 34 agreeing and 14 strongly agreeing, followed by respondents with a Master's degree, among who 17 agreed and 11 strongly agreed with the statement.

When analysed by age group, respondents aged 29-39 reported the highest level of disagreement that using artificial intelligence in the workplace will make them feel stressed with 31 disagreeing and 7 strongly disagreeing. Based on the years of experience, the levels of disagreement were notable across respondents with 4-8 years of experience, followed by respondents with 9-15 and 16-30 years of experience. Overall, as seen from Fig. 2 below the results indicate that using artificial intelligence in the workplace is not widely perceived as a source of stress as 50% of the respondents disagreed and 12% strongly disagreed.

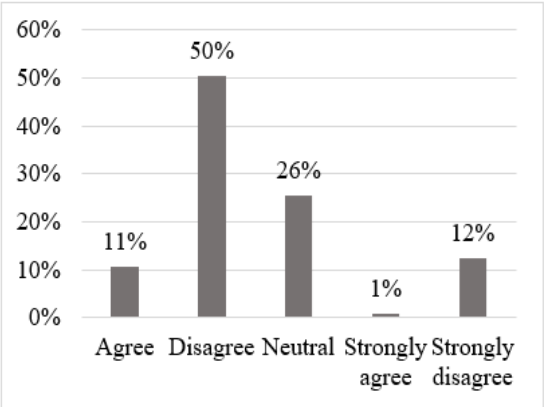


Figure 2. Perceptions of stress associated with the use of artificial intelligence in the workplace

Among the participants who believed that artificial intelligence can enhance and increase work productivity and efficiency, agreement was notable among respondents working in finance and IT. Taken together the results showed that the largest portion of respondents were positive (64% agreed and 19% strongly agreed) that artificial intelligence can enhance and increase work productivity and efficiency. When analysed by level of education, respondents across all educational groups (high school graduate, bachelor's degree, master's degree and doctorate degree) believed that artificial intelligence can enhance and increase work productivity and efficiency.

As seen from Fig. 3 the willingness to learn to use technologies and tools enabled by artificial intelligence resulted in positive responses with 55% of the respondents agreeing and 29% strongly agreeing. Based on the analysis by age group, agreement was notable by respondents aged 29-39 and based on the analysis by level of education, agreement was notable among respondents with Bachelor's degree, followed by respondents with a Master's degree.

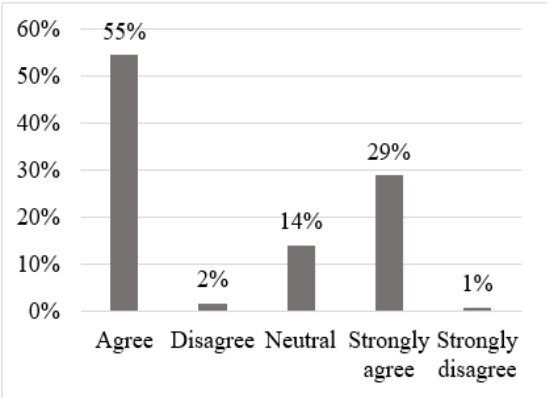


Figure 3. Willingness to learn AI-enabled technologies and tools

Based on the analysis by field of employment, respondents working in finance and IT, indicated agreement with both statements that learning AI skills will improve their job performance and that they will use artificial intelligence as a support for their daily activities. Overall, the results showed that 55% of the respondents agreed and 28% strongly agreed that by learning AI skills they will improve their job performance, and as for the statement whether respondents will use AI as a support for their daily activities 51% of the respondents agreed and 29% strongly agreed. Regarding the belief that AI can make biased decisions, respondents expressed mixed opinions, with significant portion of them remaining neutral or 40% of the respondents, followed by 29% who agreed and 20% who disagreed. The remaining respondents showed low percentages of strongly agreeing and strongly disagreeing responses (6% and 5% respectively).

6.3 Analysis of AI-related variables

The Spearman's rank-order correlation analysis revealed a significant positive association between understanding how artificial intelligence works and the preparedness to adapt to the changes caused by the application and use of artificial intelligence in the workplace ($\rho = 0.343$, $p < .001$). This indicates that participants with higher understanding of AI tended to report greater preparedness to AI adoption. The magnitude of the correlation suggests a weak-to-

moderate relationship, indicating that AI understanding is associated with preparedness to AI adoption, although other factors may also contribute to adoption behaviours.

The Spearman's rank-order correlation analysis revealed a significant negative association between familiarity with the concept of artificial intelligence and perceived stress from using AI in the workplace ($\rho = -0.210$, $p = .021$). This indicates that participants with greater familiarity with AI tended to report lower agreement that AI contributes to workplace stress. However, the strength of the relationship was weak.

A Spearman's rank-order correlation analysis revealed a significant positive association between perceived AI-related job threat and the belief that AI will replace role/job positions in the future ($\rho = 0.524$, $p < .001$). This indicates that participants who perceived AI as a greater threat to employment were also more likely to believe that AI will replace their job positions in the future. The correlation represents a moderate positive relationship.

The Spearman's rank-order correlation analysis regarding the analysis between perceived stress from using artificial intelligence and feeling prepared to adapt to the changes caused by the application and use of AI in the workplace revealed a significant negative association ($\rho = -0.268$, $p = .003$). This indicates that participants who reported greater stress associated with AI tended to report lower preparedness for adapting to AI-related changes. The relationship was weak in magnitude.

A Spearman's rank-order correlation analysis revealed a significant strong positive association between respondents' willingness to learn to use technologies and tools enabled by artificial intelligence and the statement that learning AI skills will improve job performance ($\rho = 0.785$, $p < .001$). This indicates that participants who expressed greater willingness to learn AI were also more likely to believe that AI-related learning would enhance their job performance.

7 Discussion of Results

With the help of the conducted analysis, answers were obtained regarding the defined research problem and the five research questions. From the conducted analysis of the collected database obtained through an anonymous online survey it can be concluded that respondents view artificial intelligence as an opportunity for enhancing and increasing work productivity and efficiency and creating new job opportunities in the future, rather than as a threat to their job. The majority of participants were familiar with the concept of artificial intelligence and understood how AI functions. The negative relationship between AI familiarity and perceived AI-related stress indicates that greater exposure to and understanding of AI may help reduce concerns associated with AI technologies. The willingness to learn to use the technologies and tools enabled by artificial intelligence and to use it as a support for daily activities indicates the motivated side of the respondents to move forward in the rapidly developing world of digitalisation and automation. The analysis of results showed that respondents who perceived artificial intelligence as a threat to their job position tended to think that artificial intelligence will replace their jobs. Overall, the results indicated a relatively low level of perceived job insecurity related to artificial intelligence.

The findings of this analysis can help organizations in preparing their employees in terms of education and acquiring skills for using AI-powered tools.

A limitation of this research is the small group of respondents and the lower participation of respondents from older age groups.

In the future, research can be conducted that will identify the differences and similarities in the application of artificial intelligence in the workplace between employees in the private and public sectors in the Republic of North Macedonia.

8 Conclusion

The inevitable need for the adoption of artificial intelligence represents a challenge from which the positive benefits should be accepted and the negative implications minimized. The training and familiarization of employees with the functioning of AI-powered tools will improve trust in the systems and reduce prejudices about their adoption. Daily repetitive work tasks and processing of large databases will help employees devote themselves to work tasks that require a greater degree of management and decision-making skills. It is of great importance that artificial intelligence is applied in a fair, transparent, and ethical manner, because if algorithms are not monitored and controlled, elements of bias may appear. Organizations that use artificial intelligence in combination with the human factor will be in the best position to create a work environment where employees feel motivated to work towards achieving organizational and personal goals.

References

Agrawal, A., Gans, J.S. & Golfarb, A. (2024). Prediction machines, insurance, and protection: An alternative perspective on AI's role in production. Journal of The Japanese and International Economies, 72, 101307, https://doi.org/10.1016/j.jjie.2024.101307.

Ahmed, F. (2022). Ethical Aspects of Artificial Intelligence in Banking. Journal of Research in Economics and Finance Management, 1(2), 55-63, https://doi.org/10.56596/jrefm.v1i2.7.

Ahmed, R. (2024). How to Strengthening or Weakening the Procedures of Corporate Governance: An Artificial Intelligence Perspective. Journal of Management, Accounting, General Finance and International Economic Issues, 3(2), 643-663.

Akçetin, E. (2023). Türkiye Can Leverage the Power of Innovation, Artificial Intelligence, and Fintech to Enhance Productivity. Academic Journal of Information Technology, 14(55), 335-364, https://doi.org/10.5824/ajite.2023.04.002.x.

Al-Mahameed, M. (2024). Towards a Legal Framework for Corporate Governance Amid Artificial Intelligence. Corporate Law & Governance Review, 6(3), 113-121, https://doi.org/10.22495/clgrv6i3p12.

Benhamou, S., (2020). Artificial Intelligence and the Future of Work. Revue d'économie industrielle. 169 | 1er trimestre 2020,57-88, https://doi.org/10.4000/rei.8727.

Bernard, M. (2024). What jobs will AI replace first? Retrieved from: https://www.forbes.com/sites/bernardmarr/2024/06/17/what-jobs-will-ai-replace-first/, accessed on February 10th 2026.

Daidai, F. & Tamnine, L. (2023). Artificial Intelligence and Corporate Governance. 18th International Conference on Persuasive Technology, Eindhoven University of Technology, Eindhoven, The Netherlands, April 19th - 21st.

Demirel, S. & Topcu, M. (2024). The Impact of Artificial Intelligence Applications on Digital Banking in Turkish Banking Industry. Wiley Advances in Human-Computer Interaction, 2024(1), 9921363, https://doi.org/10.1155/2024/9921363.

Gabelaia, J., Bagočiūnaitė, R. & Navickienė. V. (2024). Exploring the Intersection of Artificial Intelligence, Organizational Behaviour, and Communication Dynamics in the Modern Workplace. International Journal of Management, Knowledge and Learning, 13, 71-80, https://doi.org/10.53615/2232-5697.13.71-80.

Güngör, P. (2011). The Relationship between Reward Management System and Employee Performance with the Mediating Role of Motivation: A Quantitative Study on Global Banks. Procedia Social and Behavioral Sciences, 7th International Strategic Management Conference, 24 (2011) 1510–1520.

Kwasek, A., Kocot, M., Kocot, D., Maciaszczyk, M. & Rogozińska-Mitrut, J., (2024). European Research Studies Journal, XXVII (2), 118-130.

Narang, A., Vashisht, P. & Bajaj, S.B. (2024). Artificial Intelligence in Banking and Finance. International Journal of Innovative Research in Computer Science and Technology, 12(2), 130-134, https://doi.org/10.55524/ijircst.2024.12.2.23

Olan, F., Ogiemwonyi Arakpogun, E., Suklan, J., Nakpodia, F., Damij, N. & Jayawickrama, U. (2022). Artificial intelligence and knowledge sharing: Contributing factors to organizational performance. Journal of Business Research, 145, 605–615, https://doi.org/10.1016/j.jbusres.2022.03.008.

Rampersad, G. (2020). Robot Will Take Your Job: Innovation For an Era of Artificial Intelligence. Journal of Business Research, 116, 68-74. https://doi.org/10.1016/j.jbusres.2020.05.019.

Rao, S., Chitranshi, J. & Punjabi, N. (2020). Role of Artificial Intelligence in Employee Engagement and Retention. Journal of Applied Management-Jidnyasa, 12(2), 42-60.

Ris, K., Stankovic, Z. & Avramovic, Z. (2020). Implications of Implementation of Artificial Intelligence in the Banking Business with Correlation to the Human Factor. Journal of Computer and Communications, 8(11), 130-144, https://doi.org/10.4236/jcc.2020.811010.

Firms' Workforce Planning Orientations Under Structural Skill Shortages in North Macedonia

Predrag Trpeski, Violeta Cvetkoska, Filip Peovski, Marijana Cvetanoska, Gunter Merdzan, Kristijan Kozheski

Faculty of Economics – Skopje, Ss. Cyril and Methodius University in Skopje

Blvd. Goce Delcev 9v, Skopje, Republic of North Macedonia

{predrag.trpeski, violeta.cvetkoska, peovski, marijana.cvetanoska, gjunter.merdzan, kozeski}@eccf.ukim.edu.mk

Abstract. *Small, transition economies face chronic skill shortages and persistent emigration, which call for important decisions about workforce planning. Little is known about how firms make these decisions. Based on the resource-based human-resource architecture and semi-structured interviews with 29 companies in North Macedonia, this study develops a four-type typology of strategic planning orientations, reactive importers, domestic developers, conditional considerers and structural innovators, structured by planning horizon and labor source. The study extends the human-resource architecture to a structurally depleted labor market and finds that even proactive firms treat emigration as a contingency rather than a planning premise, a result that the strategic HRM literature does not predict.*

Keywords. strategic workforce planning, skill shortages, strategic human resource management, labor migration, resource-based view

1 Introduction

North Macedonia faces a structural tension in its labor market. On one side, prolonged emigration of working age residents depletes the domestic skill base while on the other, firms face persistent shortages of qualified workers across manufacturing, hospitality and the information and communication technology sectors. Recent comparative evidence confirms that skill shortages, skill mismatches and migration pressures jointly strain workforce stability in the country (Cvetkoska et al., 2025). For individual firms, these pressures are not abstract. Throughout this paper, skill shortage refers specifically to a sustained excess of employer demand over domestic supply for occupational profiles, while skill mismatch denotes the misalignment between the qualifications workers hold and those employers require. The two conditions frequently co-occur but are analytically distinct (Cvetkoska et al., 2025). Together they translate into unfilled vacancies, lost productivity and difficult decisions about where to source talent. Such planning issues require firms to decide which capabilities should be developed internally or obtained externally, and which depend on cooperation with education providers, public institutions, intermediaries or foreign labor markets.

Strategic planning theory draws a distinction between organizations that deliberately anticipate resource needs and build capabilities and those that only react to scarcity when acute. This distinction in proactive versus reactive workforce planning is fundamental to strategic human resource management (SHRM), defined by Wright and McMahan (1992) as the pattern of planned human resource deployments and activities used to enable a firm to achieve its strategic objectives. But much of the empirical literature on SHRM is based on large firms in advanced economies. There is little research on the actual workforce planning practices of firms in a small economy affected by pronounced emigration and whether strategic use of migrant labor is part of these practices.

We address this gap by utilizing a reflexive thematic analysis of firms' perceptions obtained through semi-structured interviews with 29 companies operating in North Macedonia. The study is positioned to gain insights into how firms (predominantly SMEs) strategically plan their workforce in the context of skill shortage and where recruitment of migrant workers figures within the strategic options they consider. Thus, the contribution is twofold as we provide an empirically grounded typology of workforce planning orientations and analyze the conditions that push firms towards reactive rather than proactive strategies.

The study is conceptualized in the following manner. Section 2 provides theoretical background to the study while Section 3 is devoted to the methodological approach which includes the research questions, analytical framework and the survey design. In Section 4 we present the main findings, followed by their discussion in Section 5. The study's conclusions are drawn in Section 6.

2 Theoretical Background

According to the resource-based view (RBV), firms can achieve sustained competitive advantage through resources that are valuable, rare, inimitable and non-substitutable (Barney, 1991). Human resources qualify when their skills and tacit knowledge are not easily replicable, a logic that Wright, McMahan and McWilliams (1994) explicitly extended to the workforce. Thus, the RBV raises the central strategic interest of this study i.e., whether in the face of scarcity, human capital should be developed internally or acquired externally by firms. This was formalized as a human resource architecture by Lepak and Snell (1999) in which firms allocate employment across internal development, acquisition, contracting and alliance modes based on the strategic value and uniqueness of the skills involved. The framework offers a direct template for classifying firm responses to shortage, as internal development equivalent to training and education partnerships, and acquisition equivalent to external recruitment which potentially includes the recruitment of migrant workers.

The distinction between proactive and reactive planning is supported by the strategic talent management literature that stresses the systematic identification of pivotal positions and deliberate building of talent pools (Collings & Mellahi, 2009) and work on workforce planning under uncertainty (Jaillet, Loke & Sim, 2022). High-performance HR practices have been associated with significantly lower turnover and higher productivity (Huselid, 1995), suggesting that planning orientation has real implications in terms of performance.

When companies resort to external acquisition of migrant labor, integration becomes a strategic concern. Research on skilled migrant integration shows that firms cannot count on state actors to support long-term integration and must provide language, regulatory, and social support themselves (Hajro et al., 2019; van Riemsdijk & Basford, 2022), while recognition of foreign credentials also affects whether migrant skills are actually used (Pecoraro & Wanner, 2019). In the Western Balkans, the noted chronical problem with the ongoing emigration of educated youth and the long-term scarring of labor market entrants (Petreski, Mojsoska-Blazevski & Bergolo, 2017) undermine the domestic supply that firms would otherwise develop, while misalignment between vocational education and employer needs exacerbates these shortages (Bartlett & Pagliarello, 2016).

These contributions are theoretically rich, but a clear gap exists. The RBV and strategic HRM literature have evolved mainly in large firms in mature labor markets where domestic skills supply, even if tight, is not structurally depleted by emigration. By contrast, the literature on the Western Balkans documents the scale of outflows and their labor market consequences but does not consider how individual firms respond strategically. To date, no study has bridged these two bodies of work to ask how firms in a small, emigration-affected economy translate, or fail, the logic of the human resource architecture into practice when the domestic talent pool is simultaneously depleted from above by credential emigration and from below by VET misalignment. The study contributes three theoretical insights to address this gap. First, it extends the human resource architecture of Lepak and Snell (1999) to a structurally depleted labor market, showing that the make-or-build choice is not a free strategic choice, but is constrained by emigration, institutional capacity and sector-specific regulatory factors that are not part of the original framework. Second, it adds planning horizon as a second organizational dimension alongside labor source, resulting in a more granular typology than architecture alone allows. Finally, it reveals that the proactive-reactive distinction, which is central to strategic HRM, behaves differently under structural shortage than the literature predicts that even proactive firms default to reactive postures with respect to the one condition, which is emigration, that most threatens their resource base. These contributions are developed through abductive analysis of interviews with 29 companies in North Macedonia, a context whose combination of SME dominance, acute emigration and EU candidate-status accession pressure makes it both empirically distinctive and theoretically generative.

3 Methodology

3.1 Instrument and Sampling

The instrument used for this study was a semi-structured interview consisting of 22 questions across five thematic blocks i.e., the company profile, assessment of labor availability and shortage occupations, the effectiveness of the education and training system, the threat of emigration, and strategic responses to shortage, which includes salary and adjustments to working conditions, internal training, cooperation with vocational education and training (VET) providers and higher education institutions (HEIs), as well as the recruitment of migrant workers.

The final sample comprises N = 29 companies operating in North Macedonia, purposively selected for sectoral and size diversity with voluntary participation. Company representatives (predominantly senior managers with direct responsibility for human resources, typically HR directors, CEOs or owners) were interviewed between May and August 2024. A combination of in-person and online surveying was used, with in-person interviews being limited to 1 hour. They were conducted in Macedonian language however, quotations presented here were translated into English by the research team and checked against the source transcripts by a bilingual researcher to preserve meaning. Thus, quotations are reproduced from these checked translations rather than being

paraphrased. The sample spans manufacturing (34.5%), information and communication technology (24.1%), financial services (20.7%), hospitality and tourism (13.8%), and the electricity, gas, steam, and air conditioning supply (6.9%) sectors. In terms of size, the sample ranges from micro-firms of four employees to large enterprises exceeding 2,900 employees. Specifically, Table 1 provides the firm size distribution within the sample, indicating a dominant proportion of small and medium sized enterprises (SMEs). To preserve confidentiality and enable traceability within the study, we reference each company by an anonymized code ranging from MKD_01 to MKD_29.

Table 1. Firm size distribution

Size (no. of employees)	n	% of total	Company codes (MKD_X)
Micro (< 10)	1	3.5%	14
Small (10-49)	11	37.9%	06, 07, 15, 16, 18, 19, 22, 23, 24, 26, 29
Medium (50-249)	9	31%	01, 08, 09, 12, 20, 21, 25, 27, 28
Large (≥ 250)	8	27.6%	02, 03, 04, 05, 10, 11, 13, 17

3.2 Research Questions

Within this exploration, the research problem is the identified gap between the rhetoric of strategic workforce planning and its practice in firms operating under chronic emigration and skill shortages. The study is guided by three research questions:

1. What strategic approaches do firms in North Macedonia use to address skill shortages?
2. To what extent are these approaches proactive or planned rather than reactive i.e., improvised?
3. Where does the recruitment of migrant workers fit within firms' strategic options, and what conditions constrain it?

The three questions are deliberately sequenced. Prior to any evaluations, it is necessary to establish what strategies exist in this context as opposed to assuming that the categories developed in the SHRM literature transfer directly to a small economy. The second research question introduces the evaluation, by testing those strategies against the proactive-reactive distinction that the theoretical background identifies as the key marker of strategic intent. The third narrows down to the specific mechanism that has emerged as the most visible response by companies to structural labor scarcity in North Macedonia.

3.3 An Integrative Analytical Framework

Combining the theoretical background and the research questions proposed, we develop the analytical framework that organizes the present study. The human resource domain supplies one axis i.e., the dominant labor source, ranging from internal development of domestic talent to external acquisition that may extend to migrant labor. The proactive-reactive distinction from strategic talent management and workforce planning theory supplies the second axis within the typology or the planning horizon. The combination of the two axes leads to four positions a company may take. Each of these correspond to a different combination of architectural mode and planning stance. Such framework is depicted in Fig. 1, which also positions the four firm types, which were derived from the interview data. The analysis was abductive rather than purely inductive, where the two dimensions of the framework, i.e. planning horizon and dominant labor source, were derived deductively from the resource-based and strategic-planning literature, while the placement and characterization of the four types within that framework emerged from iterative coding of the interview material. The typology is therefore a structured interpretation of the data rather than a claim of grounded emergence. Subsequently, each position is then populated with cases in the following section.

Interview responses were analyzed thematically. Coding focused on five strategic dimensions derived from the human-resource architecture (Lepak & Snell, 1999) specifically how shortages are identified, the planning horizon, internal development through training, collaboration with VET and higher education institutions as well as external acquisition, which primarily focuses on migrant workers recruitment. Coding was conducted independently by two researchers, with discrepancies resolved through discussion until consensus was reached; no formal inter-rater reliability coefficient was calculated, consistent with the interpretive rather than content-analytic approach adopted. Thematic development continued until no new strategic orientations were observed across the remaining interviews. Based on this, responses to the migrant recruitment theme were classified by a general perspective of active use, active consideration, conditional openness, or rejection which are then cross tabulated against each firm’s broader planning orientation. Thus, Table 2 presents the coding matrix that guided type assignment, illustrating how each dimension was operationalized and which empirical signals were treated as indicative of each level. Consequently, a two-dimensional typology was constructed, which is organized by planning horizon (reactive vs. proactive) and dominant labor source (domestic vs. international).

Table 2. Typology of workforce planning orientations

Dimension	Reactive	Moderate	Proactive
Shortage identification	Vacancy pressure only, without monitoring	Informal awareness, exit conversations	Formal KPIs, exit interviews, labor market tracking
Planning horizon	No written strategy	Salary and training adjustments	Formal talent pipeline, VET/HEI partnerships on dual education
Internal development	Ad hoc, in-house only	Some structured training, no institutional ties	Systematic cooperation with VET providers and universities
Emigration response	No strategy, contingency framing	Wage retention efforts, no formal plan	Retention embedded in HR strategy, structured monitoring
Migrant recruitment stance	Active use without integration plan	No use, sector or legal barrier cited	Planned use or deliberate international sourcing

The framework is diagnostic, not normative, by design. It does not assume that internal development is always better, because there are some sectors, at some scales, for which external acquisition is the sensible response. What the framework reveals is whether a firm's sourcing choices are the result of thoughtful design or a haphazard response to shortage. Both the resource-based and resource-dependence traditions would predict that firms facing a known, sustained threat to their core resource would make explicit plans for it. Testing whether that prediction holds in a structurally depleted labor market is the empirical task the framework allows. One aspect of the framework requires a special comment. The lower-left quadrant, which is the reactive planning with a domestic-only labor source, is theoretically possible but empirically shown to be unoccupied in this sample. No firm depended exclusively on isolated domestic hiring without either investing in some internal development or resorting to migrant labor when domestic supply failed. We believe that this absence is not accidental. This suggests that a steady state of reactive hiring from a structurally depleted domestic pool is not sustainable as firms that cannot develop domestically and do not recruit internationally are eventually forced to either invest in training, import labor or exit the market. The empirical vacancy of the quadrant thus supports the theoretical expectation that structural scarcity forces strategic choice rather than permitting indefinite shifts.

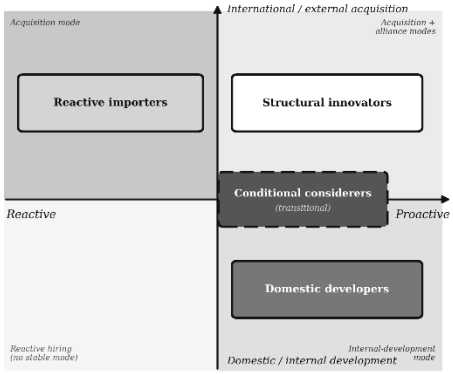


Figure 1. The resource-based human-resource architecture mapped to strategic workforce planning, locating the four firm types

4 Findings

The severity and uneven distribution of shortage were initially confirmed by the interviews. Assessments ranged from comfortable to acute. A dairy processor (MKD_15) rated the availability of suitably qualified labor at the very bottom of the scale, reporting that all employees arrive without prior knowledge and are trained on the spot. An electrical manufacturer (MKD_18) described availability of skilled labor as very difficult, noting the absence of local vocational schools producing adequate profiles. A hospitality firm (MKD_28) summarized the sectoral position bluntly:

> Overall, as a company and a sector, we are in serious trouble. Availability is currently very difficult. (MKD_28)

On the contrary, several financial sector and large manufacturing companies (for example MKD_03 and MKD_12) reported minor difficulties, attributing this to their perception as "*desirable employers*". We find that shortages are concentrated in lower-skill manufacturing, hospitality as well as specialized technical roles, with moderate cases within the financial sector.

Companies were almost unanimous in their perceptions that material rewards and compensations as well as internal training are their primary instruments in tackling shortages. However, since their strategic orientations notably diverged, we identify four distinct types. Table 3 summarizes the typology which is in line with the analytical framework presented in Fig. 1.

Table 3. Typology of workforce planning orientations

Type	Horizon	Labor source	Company examples (MKD_X)
Reactive importers	Reactive	Migrant	01, 19, 28
Domestic developers	Proactive	Domestic	02, 05, 10, 21
Conditional considerers	Proactive	Boundary; domestic development with latent international acquisition	20, 23, 24
Structural innovators	Proactive	International / remote	11, 16, 27

To describe each type, we need to account for the whole sample. Fourteen of the 29 firms fit neatly into one of the four types. The other 15 concentrated in financial services, mid-sized manufacturing and ICT are in a separate analytical middle ground. They plan more deliberately than reactive importers, but they lack the formal talent pipelines of domestic developers, and they do not use or seriously consider migrant labor. They occupy the proactive-domestic zone of the framework but have not made the sustained institutional commitments that characterize domestic developers. Instead of deliberately assigning them to fit into an existing type or to a fifth category based on limited evidence, we view this group as the modal centre of the sample and those are companies whose planning is adequate for normal conditions but not for the structural scarcity they face. Their prevalence (approximately 52% of the sample) suggests that the most common organizational response to skill shortage in this context is neither deliberate strategy nor improvised reaction, but a moderate, domestically oriented pattern that postpones difficult economic choices. The four named types are the analytically distinctive positions at the edges of this distribution, and it is where strategic differentiation is most visible and most consequential.

4.1 Type 1: Reactive Importers

The first group considers migrant labor as a stopgap, to be used if domestic recruitment has failed, with little planning for integration. These companies are mostly in hospitality and are in low skill manufacturing. The most obvious example is a hospitality firm (MKD_28) which described an active but improvised contracting arrangement:

> Yes, we have already contracted with two companies that bring labor from Nepal. We expect 40 new hires soon. The documentation takes a really long time, but I heard positive experiences from colleagues in the sector who concluded such contracts a few months ago and already have foreigners working in their premises with good results. It remains to be seen what our experience will be in this domain. (MKD_28)

Similarly, a dairy-processing firm (MKD_01) explained that it had applied for ten workers from Nepal while waiting eight months for approval, presenting this as an answer to an immediate gap in manual labor rather than a planned capability. Another hospitality operator (MKD_19) said it had already hired workers from India and Albania and was waiting for more to come. A second dairy processor (MKD_15) was explicit that import is its default coping mechanism, having rated local labor availability at zero:

> As a company, we solve the lack of workers by importing workers from abroad. Until now, we have employed two workers from India in our company and two more from Nepal should arrive. (MKD_15)

For these firms, recruitment of migrants is reactive as it is triggered by acute vacancy pressure, agency mediation, and conducted without a stated integration or retention plan. The mode corresponds to the acquisition quadrant of the human-resource architecture, but without the complementary alliance or development activity that would render it strategic.

4.2 Type 2: Domestic Developers

A second, larger group of companies pursues a deliberately long horizon strategy of building domestic talent with an explicit rejection of migrant recruitment labeled as unnecessary. A pharmaceutical manufacturer (MKD_10) noted:

> Our company has a main focus on employing residents from North Macedonia. Only in this way will population migration be reduced. (MKD_10)

This group is characterized by formal education partnerships. One food manufacturer (MKD_02) mentioned that they pioneered university scholarships fifteen years ago while an energy company (MKD_05) has a dual-education project for electrical fitters. An insurance company (MKD_21) directly expressed the make-or-buy logic of the RBV, saying that it is better to create staff rather than acquire them ready-made, even though it is difficult, long and expensive. That same insurer summed up the developmental philosophy perfectly:

> We don’t get the frame ready, we have to create it. (MKD_21)

For these firms, internal development is not the second-best option, but the heart of strategy which is pursued over a long horizon through deliberate institutional partnerships. They fall within the internal-development quadrant of the framework and plan proactively within that quadrant.

4.3 Type 3: Conditional Considerers

A third group occupies a transitional position. They are proactive in their domestic development holding migrant recruitment as a latent option not yet activated, these firms sit at the boundary between the domestic-developer and structural-innovator quadrants rather than a quadrant of their own. They have no migrant recruitment strategy in place but would adopt one should domestic options prove insufficient. The fruit-and-vegetable processor (MKD_24) was most explicit, reporting that the admission of skilled migrant workers is in the plan and expressing hope it would help fill positions where professional staff are lacking, while flagging administrative, language and cultural challenges. A confectionery producer (MKD_23) and an ICT firm (MKD_20) expressed similar conditional openness. At the domestic level, these firms plan proactively, but they see migrant labor as a contingency rather than an active channel.

4.4 Type 4: Structural Innovators

A smaller group rethinks the sourcing problem strategically by widening the talent pool through international and remote channels. An ICT firm (MKD_27) framed emigration not as a loss to be contained but as a reason to redesign sourcing:

> Broadening international talent pools, offering remote or hybrid job opportunities and attractive benefits along with the highly competitive salaries can be great strategies in this regard. (MKD_27)

Another ICT firm (MKD_16) reported openness to hiring migrants for online cooperation in cases where a need arises, while a multinational subsidiary (MKD_11) already employs foreign nationals. For these firms, the distinction between domestic and migrant labor is seemingly dissolving into a single international talent strategy.

4.5 The Reactive Default in Emigration Planning

Across all four types, interviews revealed almost a total absence of formal strategy for managing emigration. Companies that have sophisticated internal development systems believe there is no such thing as a dedicated plan. One insurer (MKD_09) said that it does not have a concrete strategy and does not see the need to develop one, as it has not faced the problem yet and will develop coping strategies in case of future threats. Similarly, an electrical manufacturer (MKD_18) and an ICT firm (MKD_16) said they had no formal, written strategy. This means that even proactive domestic developers are reactive in their planning to emigration itself, treating it as a contingency to be addressed once realized rather than a foreseeable structural condition. It should be noted that structural innovators have the clearest evidence drawn from a small number of ICT firms. The type should therefore be read as an analytically identified tendency whose prevalence across other sectors and firm sizes requires careful confirmation in a larger sample.

5 Discussion

5.1 Answering the Research Questions

The typology directly answers the first two research questions. The Macedonian companies use a common baseline of reward and internal training, but they are different in strategic orientation along two axes based on how far ahead they plan and whether they look for talent at home or abroad. Following Lepak and Snell's (1999) human-resource architecture for mapping firms, domestic developers fall into the internal development quadrant, reactive importers into the acquisition quadrant, and structural innovators into a hybrid of acquisition with alliance and contracting modes.

For the third research question, migrant recruitment surfaces as a real, though mostly marginal, strategic option. Very few firms are doing this actively and, where they are, it is overwhelmingly in a reactive mode, triggered by acute shortage of lower skill roles. More strategic use was often prevented by three conditions. First, administrative and legal friction. Documentation and work-visa procedures were lengthy, with one (MKD_01) citing an eight month wait. Second is the integration burden. One electrical manufacturer (MKD_18) commented that migrant workers are totally dependent on the firm for accommodation, food, transport and healthcare, with an uncertain tenure, as "*the economy is not yet ready*". Third, the sector-specificity and language barriers must be emphasized. Financial sector firms, such as MKD_03 for instance, found migrant recruitment to be incompatible with locally regulated, and importantly, Macedonian language intensive work. These conditions match the limiting factors and the integration-support gap identified by van Riemsdijk and Basford (2022).

The most strategically important finding is the reactive default in emigration planning. Although the structural driver of shortage in the Western Balkans is emigration (Petreski et al., 2017), firms tend to treat it as a contingency rather than a planning premise. The strategic HRM literature would predict that firms exposed to a known, sustained threat to their core resource would explicitly plan for it (Wright & McMahan, 1992), while the resource dependence tradition would predict active buffering against environmental scarcity. This is suggested by the fact that even sophisticated domestic developers do neither, suggesting that the lack of formal emigration planning is structural rather than a simple capability deficit. It indicates that the binding constraint is not the intention of managers, even besides the companies' clear investment in talent, but the lack of frameworks and

institutional support to enable them to shift from reactive responses to anticipation. This is not an obvious consequence.

A second observation concerns the role of technology as a strategic substitute. Two firms framed automation as an alternative to both domestic recruitment and migrant labor. For instance, a confectionery producer (MKD_23) noted that a large part of its work can increasingly be performed by machines, with another manufacturer pursuing the same logic. Automation within the framework effectively shifts a firm's position by eliminating its dependence on the labor source axis altogether. This option is more readily available to capital-intensive firms than to service-intensive firms and helps explain why hospitality firms gravitate toward the reactive-importer position while some manufacturers can remain domestic developers.

The framework also explains why migrant recruitment is so rarely strategic. To move from the reactive-importer to the structural-innovator position, complementary investments are needed in integration support, alliance relationships and remote-work infrastructure. The acquisition mode does not supply that. The great majority of firms that lacked these complements remained reactive even when they were actively importing labor. A handful of structural innovators (notably ICT firms, such as MKD_27) had already built the remote and international infrastructure that enables external acquisition to work as deliberate strategy.

5.2 Conditions Shaping Strategic Choices

Locating the firm types within the framework raises the question of what pushes a firm toward one quadrant rather than another. The interviews point to three recurring drivers. The first is sectoral capital intensity. Capital-intensive manufacturers can substitute automation for labor and sustain internal development, gravitating toward the domestic-developer position. On the other hand, labor-intensive hospitality firms cannot and are pushed toward reactive imports of migrant workers. The second is the regulatory and linguistic specificity of the work. Financial sector firms (for example MKD_03 and MKD_21) judged migrant recruitment incompatible with locally regulated, language-intensive roles, which anchors them in the domestic quadrants irrespective of the level of planning sophistication. The third driver is firm scale and employer brand. Large and well-known employers (MKD_04, MKD_10, MKD_12) reported that their positioning in the labor market produces less identified shortages, allowing a proactive, development-led posture, whereas small firms facing the same shortage lack that buffer and react.

These determinants are largely structural, rather than discretionary, which redefines the central paradox. This reactive default is not merely a failure of managerial foresight but the predictable result of firms' location in the labor market such that the complements necessary for proactive strategy i.e., capital, brand, regulatory fit, or even institutional support are missing. In such case, structure conditions strategy and policy that changes those conditions (e.g. easing work-permit administration or strengthening sector-aligned vocational training) would do more to move firms towards proactive planning.

5.3 Limitations

The findings have several limitations. The study design is qualitative and the sample is small and purposive, so the typology is analytical and theoretical rather than statistically generalizable, as the relative size of each type in the wider population is unknown. Evidence is cross-sectional and self-reported as perceptions by company representatives which may exaggerate the coherence of the stated strategies. Additionally, the study only captures the employer perspective as migrant workers experience, and whether employer strategies succeed from the receiving side, are outside the scope of this study. Finally, the analysis is limited to a single economy and, even though it is representative of the Western Balkans, sectoral and institutional particularities limit direct transferability to other contexts. Future research could test the typology quantitatively across a larger firm sample, track firms longitudinally to observe movement between positions as well as incorporate migrant worker perspectives to assess whether employer strategies succeed from their side. Extending the framework to other Western Balkan economies would establish how far the four-type structure travels beyond North Macedonia.

6 Conclusion

This study examined how 29 firms in North Macedonia strategically plan their workforce under skill shortage, and where migrant labor fits in such context. It contributes a four-type typology i.e., reactive importers, domestic developers, conditional considerers, and structural innovators, which are organized by planning horizon and labor source. It shows that migrant recruitment is mostly a reactive instrument constrained by administrative friction, integration burden, and sector specificities.

The findings suggest that for managers, considering emigration as a planning assumption rather than a contingency and coupling any recruitment of migrants with a deliberate integration plan would change firms from reactive to proactive positions. The structural innovators in the sample demonstrate the move is possible: they built remote-work infrastructure and international talent pipelines before acute shortage forced their hand, turning external acquisition from a stopgap into a planned capability. Partial automation combined with ongoing internal training provides a

different route for firms in labor-intensive sectors to avoid the reactive importer position.

The analysis suggests to policymakers that the binding constraints are predominantly structural and hence, addressable by policies. Cutting administrative friction in work permit procedures, where the months-long obstacle that has been cited repeatedly, would reduce the cost of strategic migrant recruitment. Improving the alignment between vocational education and employer needs would broaden the domestic-development option, while providing support for integration services would lighten the burden that currently falls entirely on individual firms. Importantly, each such measure changes the structural conditions that determine in which quadrant a firm might realistically be placed, a finding that calls on strategic HRM theory to incorporate institutional depletion of domestic labor supply as a boundary condition to the proactive-reactive planning distinction.

Acknowledgments

The authors gratefully acknowledge the companies that participated in the interviews and the wider Horizon Skills4Justice project team whose data collection made this analysis possible.

References

Barney, J. B. (1991). Firm resources and sustained competitive advantage. *Journal of Management, 17*(1), 99-120. https://doi.org/10.1177/014920639101700108

Bartlett, W., & Pagliarello, M. C. (2016). Agenda-setting for VET policy in the Western Balkans: Employability versus social inclusion. *European Journal of Education, 51*(3), 305-319. https://doi.org/10.1111/ejed.12182

Collings, D. G., & Mellahi, K. (2009). Strategic talent management: A review and research agenda. *Human Resource Management Review, 19*(4), 304-313. https://doi.org/10.1016/j.hrmr.2009.04.001

Cvetkoska, V., Trpeski, P., Ivanovski, I., Peovski, F., İmrol, M. H., Babadoğan, B., Ecer, H., Görür, D. Z., Selvi, U., Hunde, A. B., Gemeda, F. T., Dubi, Y. B., Melnyk, S., Lytvynchuk, A., & Tereshchenko, H. (2025). Comparative analysis of skill shortages, skill mismatches, and the threats of migration in labor markets: A sectoral approach in North Macedonia, Türkiye, Ethiopia, and Ukraine. *Social Sciences, 14*(5), 294. https://doi.org/10.3390/socsci14050294

Hajro, A., Stahl, G. K., Pisani, N., & Caprar, D. V. (2019). Acculturation, coping, and integration success of international skilled migrants: An integrative review and multilevel framework. *Human Resource Management Journal, 29*(3), 328-352. https://doi.org/10.1111/1748-8583.12233

Huselid, M. A. (1995). The impact of human resource management practices on turnover, productivity, and corporate financial performance. *Academy of Management Journal, 38*(3), 635-672. https://doi.org/10.5465/256741

Jaillet, P., Loke, G. G., & Sim, M. (2022). Strategic workforce planning under uncertainty. *Operations Research, 70*(2), 1042-1065. https://doi.org/10.1287/opre.2021.2183

Lepak, D. P., & Snell, S. A. (1999). The human resource architecture: Toward a theory of human capital allocation and development. *The Academy of Management Review, 24*(1), 31-48. https://doi.org/10.5465/amr.1999.1580439

Pecoraro, M., & Wanner, P. (2019). Does the recognition of foreign credentials decrease the risk for immigrants of being mismatched in education or skills? In I. Steiner & P. Wanner (Eds.), *Migrants and expats: The Swiss migration and mobility nexus* (pp. 161-186). Springer International Publishing. https://doi.org/10.1007/978-3-030-05671-1_7

Petreski, M., Mojsoska-Blazevski, N., & Bergolo, M. (2017). Labor-market scars when youth unemployment is extremely high: Evidence from Macedonia. *Eastern European Economics, 55*(2), 168-196. https://doi.org/10.1080/00128775.2016.1261631

van Riemsdijk, M., & Basford, S. (2022). Integration of highly skilled migrants in the workplace: A multi-level framework. *Journal of International Migration and Integration, 23*(2), 633-654. https://doi.org/10.1007/s12134-021-00845-x

Wright, P. M., & McMahan, G. C. (1992). Theoretical perspectives for strategic human resource management. *Journal of Management, 18*(2), 295-320. https://doi.org/10.1177/014920639201800205

Wright, P. M., McMahan, G. C., & McWilliams, A. (1994). Human resources and sustained competitive advantage: A resource-based perspective. *The International Journal of Human Resource Management, 5*(2), 301-326. https://doi.org/10.1080/09585199400000020

Not All Technostress Is Equal: Techno-Complexity, Perceived Ease of Use, and Developers' Acceptance of Emerging Development Technologies

Marija Boban
University of Split, Faculty of Law
Department of Information Science and Statistics
Domovinskog rata 8, 21000 Split, Croatia
marija.boban@pravst.hr

Marija Gombar
General Staff of the Armed Forces of the Republic of Croatia
Sarajevska 7, 10000 Zagreb, Croatia
gombar.ma@gmail.com

Abstract. *This study examines techno-complexity as a usability-related barrier to developers' acceptance of emerging development technologies. Survey data from 237 Croatian software-development professionals show that techno-complexity reduces perceived ease of use, while perceived ease of use increases intention to use. The negative indirect effect confirms that complexity weakens acceptance through a usability pathway rather than as a uniformly negative direct predictor. The contribution lies in isolating techno-complexity from techno-insecurity and techno-uncertainty in a developer-specific context. Respondents evaluated development tools, programming languages, databases, and operating systems, not a specific AI tool; AI-related conclusions are therefore presented only as implications.*

Keywords. technostress, techno-complexity, developers, technology acceptance, perceived ease of use, emerging development technologies

1 Introduction

Software development work is increasingly shaped by new or substantially changing technologies, including development tools, programming languages, databases, and primary operating systems. More recently, generative AI and AI-assisted coding tools have intensified this broader transformation by changing how developers search for solutions, write code, evaluate outputs, and interact with development environments (Barke et al., 2023; Nguyen-Duc et al., 2025; Russo et al., 2024; Weisz et al., 2025). However, the empirical scope of this study is broader than AI-assisted software engineering and does not concern a named tool, platform, or AI system. Respondents evaluated new technologies used in their own areas of software-development work. Their acceptance therefore depends not only on expected technical benefits, but also on whether such technologies are perceived as understandable, manageable, and sufficiently easy to integrate into existing work practices.

Technology acceptance research has long emphasised the importance of perceived ease of use and perceived usefulness in explaining users' intention to adopt information technologies (Davis, 1989; Venkatesh & Davis, 2000). In this paper, the focus is placed on perceived ease of use because techno-complexity is expected to operate primarily as a usability-related barrier. In software development contexts, this issue becomes particularly relevant because new tools often require continuous learning, adaptation to changing workflows, and evaluation of outputs whose correctness, security, or reliability may not be immediately transparent. Studies on AI-assisted programming show that developers may benefit from code generation and automation, but they also face usability, trust, security, and validation challenges (Pearce et al., 2025; Perry et al., 2023; Prather et al., 2024; Vaithilingam et al., 2022). Therefore, understanding technology acceptance among developers requires attention not only to perceived benefits, but also to the cognitive and practical burden created by technological complexity. Technostress provides a useful lens for analysing this burden. Earlier research conceptualised technostress as a multidimensional phenomenon associated with overload, invasion, complexity, insecurity, and uncertainty in technology-intensive work (Ragu-Nathan et al., 2008; Tarafdar et al., 2019). More recent studies suggest that technostress should not be treated as a uniform barrier, especially in contexts shaped by AI and rapidly evolving digital tools (Chang et al., 2024; Högemann et al., 2025; Nastjuk et al., 2024). Some forms of technological pressure may primarily affect usability perceptions, while others may relate to professional insecurity, uncertainty, or adaptation. This distinction is important for responsible digital transformation because organisations cannot support technology adoption effectively if all forms of technostress are addressed in the same way.

This paper focuses on techno-complexity as a specific and actionable barrier to developers'

acceptance of emerging development technologies. Techno-complexity refers to the extent to which users perceive technologies as difficult to understand, learn, or use. In development work, such complexity may reduce perceived ease of use and indirectly weaken intention to use new tools. At the same time, the growing role of intelligent and AI-assisted development environments makes this issue increasingly relevant for organisations that seek to introduce new technologies responsibly, without overlooking developers' competencies, workload, and need for human-centred support. This paper aims to examine the relationship between techno-complexity, perceived ease of use, and intention to use emerging development technologies among developers and programmers. The main research question is: How does techno-complexity influence developers' intention to use emerging development technologies, and is this relationship mediated by perceived ease of use? In addition, the paper considers techno-insecurity and techno-uncertainty as exploratory extensions to examine whether different technostress dimensions provide additional explanatory value beyond techno-complexity.

The novelty of this paper lies in the combination of a developer-specific context, the isolation of techno-complexity from other technostress dimensions, and the testing of perceived ease of use as the mechanism linking complexity to acceptance. Rather than treating technostress as a single aggregate barrier or presenting its dimensions only as parallel predictors, the study conceptualises techno-complexity as a distinct usability-related mechanism and contrasts it with techno-insecurity and techno-uncertainty through an exploratory extension. The contribution is therefore not the proposal of a new general technology-acceptance model, but a more precise explanation of how one specific technostress dimension operates among software-development professionals.

2 Theoretical Background and Research Model

Technology acceptance research explains technology adoption primarily through users' beliefs about whether a technology is useful and easy to use. In the Technology Acceptance Model, perceived ease of use refers to the degree to which a person believes that using a technology would be free of effort, while intention to use reflects behavioural readiness to adopt the technology (Davis, 1989; Venkatesh & Davis, 2000). In software development work, perceived ease of use is particularly relevant because new development tools often require continuous learning, adaptation to new workflows, and evaluation of complex or automatically generated outputs. Recent research on AI-assisted and generative development tools shows that developers do not interact with such technologies only as passive users. They evaluate suggestions, validate outputs, manage uncertainty, and adjust their programming strategies (Barke et al., 2023; Prather et al., 2024; Vaithilingam et al., 2022; Weisz et al., 2025). These tools may improve productivity and support complex development tasks, but they may also introduce new usability, security, and verification challenges (Pearce et al., 2025; Perry et al., 2023; Russo et al., 2024). Therefore, developers' acceptance of emerging development technologies should be examined not only through their expected benefits, but also through the perceived complexity of using them.

Technostress research provides a useful framework for understanding technology-related strain in digitally intensive work. Technostress creators include several dimensions, such as techno-overload, techno-invasion, techno-complexity, techno-insecurity, and techno-uncertainty (Ragu-Nathan et al., 2008; Tarafdar et al., 2019). However, recent work suggests that these dimensions should not be treated as identical barriers to technology adoption, especially in contexts shaped by artificial intelligence and rapidly changing digital tools (Chang et al., 2024; Högemann et al., 2025; Nastjuk et al., 2024). Among these dimensions, techno-complexity is especially relevant for software developers because it captures the perceived difficulty of understanding, learning, and using new technologies. In this paper, techno-complexity is conceptualised as a specific barrier to developers' acceptance of emerging development technologies. When developers perceive a technology as complex, they may also perceive it as less easy to use. Lower perceived ease of use can then reduce their intention to use the technology. This mechanism is important for responsible digital transformation because organisations introducing increasingly intelligent development environments need to support developers not only with access to tools, but also with training, integration support, and human-centred implementation practices (Gombar & Boban, 2026; Russo et al., 2024).

Based on this reasoning, the following hypotheses are proposed:

H1. Techno-complexity negatively influences perceived ease of use of emerging development technologies.

H2. Perceived ease of use positively influences intention to use emerging development technologies.

H3. Perceived ease of use mediates the relationship between techno-complexity and intention to use emerging development technologies.

In addition to the confirmatory model, the study considers two further technostress dimensions: techno-insecurity and techno-uncertainty. Techno-insecurity refers to the perceived threat that technological change may weaken one's professional position or competence, while techno-uncertainty reflects the pressure of frequent technological change. Since these dimensions may influence acceptance differently from

techno-complexity, they are included as exploratory extensions.

ERQ1. Do techno-insecurity and techno-uncertainty provide additional explanatory value beyond techno-complexity in explaining developers' intention to use emerging development technologies?

As shown in Fig. 1, the proposed research model centres on techno-complexity as the main predictor, perceived ease of use as the mediator, and intention to use as the outcome variable, with techno-insecurity and techno-uncertainty examined as exploratory technostress extensions.

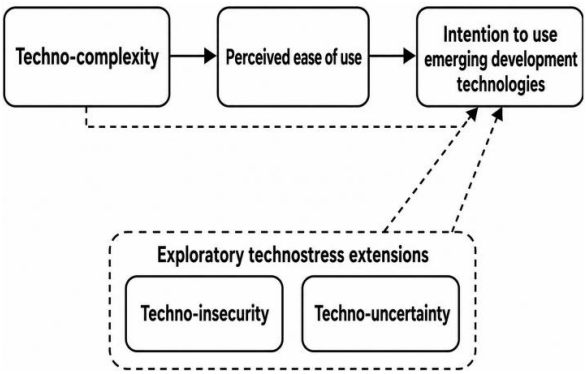


Figure 1. Proposed research model

Note. Solid arrows indicate hypothesised main relationships. Dashed arrows indicate direct or exploratory paths.

3 Methodology

3.1 Sample and Data Collection

The study is based on survey data collected from 237 software-development professionals in the Croatian IT sector. Data were collected during the first quarter of 2025 through an online questionnaire administered using Google Forms. The questionnaire link was distributed via email and messaging applications, specifically WhatsApp and Viber. Before accessing the questionnaire, all participants were informed about the purpose of the study, the voluntary nature of participation, the confidentiality and protection of their data, and that responses would be used exclusively for research purposes in aggregated form. Respondents were included in the analytical sample when their reported occupational title identified them as developers, programmers, or professionals performing an equivalent software-development role. Responses were anonymised, and the individual respondent was used as the unit of analysis. Regarding project roles, 219 respondents (92.4%) were developers, nine (3.8%) were software architects, six (2.5%) were project leaders, and three (1.3%) were test engineers. Most respondents were aged 21–30 years (68.4%), followed by 31–40 years (21.5%) and 41–50 years (8.0%). Men represented 69.6% of the sample, women 29.1%, while 1.3% preferred not to report gender. Regarding education, 59.9% held a graduate degree, 20.3% an undergraduate degree, 9.3% vocational secondary education, 6.8% general secondary education, and 3.4% a postgraduate degree. Information on respondents' years of professional experience was not available in the analytical dataset.

Respondents did not evaluate one named technology, platform, or AI tool. The questionnaire referred more broadly to new technologies encountered in respondents' own software development work, encompassing categories such as development tools, programming languages, databases, and primary operating systems. Accordingly, the term "emerging development technologies" is used as an umbrella category rather than as a reference to a single homogeneous technology. Implications concerning AI-assisted software engineering are therefore interpretative extensions of the findings and not direct empirical evidence about the use of specific AI systems.

3.2 Measures

The study included five constructs measured with 20 items: techno-complexity (five items), techno-insecurity (four items), techno-uncertainty (five items), perceived ease of use (three items), and intention to use (three items). Techno-complexity captures the perceived difficulty of understanding, learning, and using new technologies. Techno-insecurity reflects concerns that technological change may weaken one's professional competence or position, while techno-uncertainty captures the pressure associated with frequent technological changes. Perceived ease of use refers to the extent to which respondents perceive new technologies as easy to learn and use, while intention to use reflects their readiness to use such technologies. The questionnaire was developed by adapting established technostress and technology-acceptance scales to the context of emerging development technologies. The technostress items were based on established technostress-creator dimensions, particularly techno-complexity, techno-insecurity, and techno-uncertainty (Ragu-Nathan et al., 2008; Tarafdar et al., 2019). The perceived ease-of-use and intention-to-use items were grounded in the Technology Acceptance Model (Davis, 1989; Venkatesh & Davis, 2000). Respondents indicated their agreement on a seven-point Likert scale ranging from 1 (strongly disagree) to 7 (strongly agree). Composite scores were calculated as the mean of the respective items. Table 1 presents the English wording of the items included in the analysis.

Table 1. Measurement items

Construct	Items
Techno-complexity (TC)	**TC1:** I do not know enough about new technologies in my area of work to perform my job satisfactorily. **TC2:** It takes me a long time to understand and use new technologies in my area of work. **TC3:** I cannot find enough time to learn and improve my technological skills. **TC4:** I believe that new employees in this organisation know more about computer technology than I do. **TC5:** I often find it too complex to understand and use new technologies in my area of work.
Techno-insecurity (TINS)	**TINS1:** I feel a constant threat to my job security because of new technology in my area of work. **TINS2:** I have to continually upgrade my technological skills to avoid being replaced. **TINS3:** I feel threatened by colleagues with advanced knowledge of new technologies in my area of work. **TINS4:** I feel that less knowledge about new technologies in my area of work is shared among colleagues because of fear of replacement.
Techno-uncertainty (TU)	**TU1:** There are always new developments in the technologies used in our organisation. **TU2:** There are constant changes in the programming languages used in our organisation. **TU3:** There are constant changes in the development tools used in our organisation. **TU4:** There are constant changes in the databases used in our organisation. **TU5:** There are constant changes in the primary operating systems used in our organisation.
Perceived ease of use (PEOU)	**PEOU1:** I find it easy to learn and master new technologies in my area of work. **PEOU2:** I easily find the information I need about new technologies in my area of work. **PEOU3:** I consider new technologies in my area of work easy to use.
Intention to use (ITU)	**ITU1:** I am inclined to use new technologies in my area of work. **ITU2:** Assuming that I have access to new technologies in my area of work, I intend to use them. **ITU3:** Given that I have access to new technologies in my area of work, I expect that I will use them.

Note. All items were measured on a seven-point Likert scale ranging from 1 (*strongly disagree*) to 7 (*strongly agree*). TC = techno-complexity; TINS = techno-insecurity; TU = techno-uncertainty; PEOU = perceived ease of use; ITU = intention to use.

3.3 Data Analysis

The analysis was conducted in several steps. First, the analytical dataset was screened for missing values, and descriptive statistics were calculated for all constructs. Second, internal consistency was assessed using Cronbach's alpha and McDonald's omega. Third, Pearson correlations were used to examine bivariate relationships among the constructs. To test the research model, robust observed-variable path analysis was conducted using standardized composite scores. The main model examined the effect of techno-complexity on perceived ease of use and the effect of perceived ease of use on intention to use. The mediating role of perceived ease of use was assessed using bootstrap confidence intervals with 5,000 resamples (Hayes, 2022). Techno-insecurity and techno-uncertainty were analysed as exploratory extensions to assess whether they provided additional explanatory value beyond techno-complexity. These exploratory paths were not treated as confirmatory hypothesis tests but as diagnostic extensions used to examine whether other technostress dimensions changed the interpretation of techno-complexity. Discriminant validity was additionally assessed using the heterotrait–monotrait ratio of correlations (HTMT), calculated from the item-level correlations among the five constructs. HTMT values below 0.90 were considered evidence that the constructs were empirically distinguishable. The results of this assessment are reported together with the descriptive statistics and reliability indicators.

4 Results

4.1 Descriptive Statistics, Reliability, and Discriminant Validity

Table 2 presents descriptive statistics and reliability indicators for the constructs included in the analysis. All scales showed acceptable internal consistency, with Cronbach's alpha values ranging from 0.730 to 0.879. The highest mean value was observed for intention to use, indicating that respondents generally expressed a relatively high readiness to use emerging development technologies. Techno-complexity, techno-insecurity, and techno-uncertainty showed lower mean values, suggesting that technostress-related perceptions were present but not dominant in the sample.

Table 2. Descriptive statistics and reliability of constructs

Construct	k	M	SD	α	ω
TC	5	2.775	1.184	0.785	0.860
TINS	4	2.310	1.123	0.733	0.846
TU	5	2.430	1.000	0.730	0.829
PEOU	3	4.602	1.322	0.832	0.902
ITU	3	5.167	1.387	0.879	0.926

Note. Constructs were measured using multi-item Likert-type scales. Composite scores were calculated as mean values of their respective items.

Discriminant validity was supported by the HTMT assessment. All ten pairwise HTMT values were below the 0.90 threshold, ranging from 0.106 to 0.863. The highest value was observed between techno-complexity and techno-insecurity (HTMT = 0.863), followed by techno-insecurity and techno-uncertainty (HTMT = 0.819) and perceived ease of use and intention to use (HTMT = 0.787). These results indicate that the five constructs were empirically distinguishable. The correlation analysis showed a moderate negative relationship between techno-complexity and perceived ease of use (r = -0.341), and a strong positive relationship between perceived ease of use and intention to use (r = 0.675). The bivariate relationship between techno-complexity and intention to use was weak and not statistically significant (r = -0.072), indicating that the role of techno-complexity should be interpreted primarily through the proposed usability pathway rather than as a simple direct effect.

4.2 Test of the Main Model

The main model tested the relationship between techno-complexity, perceived ease of use, and intention to use. Robust path analysis was conducted using standardized composite scores. The results are presented in Table 3.

Table 3. Main path and mediation results

Path	β	Evidence/decision
H1: TC → PEOU	-0.341	p < .001; supported
H2: PEOU → ITU	0.735	p < .001; supported
Direct: TC → ITU	0.179	p = .002; mediation test
H3: TC → PEOU → ITU	-0.249	95% CI [-0.345, -0.150]; supported

Note. Standardized coefficients are reported. The direct path was included to test mediation and should not be interpreted as evidence that techno-complexity has a simple negative total effect on intention to use. The indirect effect was estimated using 5,000 bootstrap resamples.

Fig. 2 visualises the tested mediation mechanism. The results show that techno-complexity significantly reduces perceived ease of use, while perceived ease of use significantly increases intention to use emerging development technologies. The indirect effect is negative and statistically significant, indicating that techno-complexity weakens intention to use specifically through the perceived ease-of-use pathway.

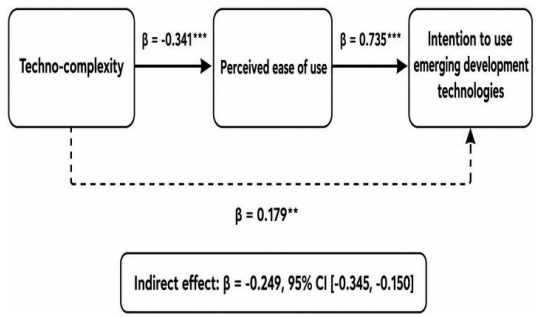


Figure 2. Empirical model with standardised coefficients

Note. **p < .01; ***p < .001.

The results support H1. Techno-complexity had a significant negative effect on perceived ease of use (β = -0.341, p < .001), indicating that developers who perceive emerging development technologies as more complex also perceive them as less easy to use. The model explained 11.6% of the variance in perceived ease of use. H2 was also supported. Perceived ease of use had a strong positive effect on intention to use (β = 0.735, p < .001) when techno-complexity was included in the model. Together, techno-complexity and perceived ease of use explained 48.3% of the variance in intention to use. The mediation analysis supported H3. The indirect effect of techno-complexity on intention to use through perceived ease of use was negative and statistically significant (β = -0.249, 95% CI [-0.345, -0.150]). The results suggest that techno-complexity weakens developers' intention to use emerging development technologies indirectly, through its negative effect on perceived ease of use.

The total bivariate relationship between techno-complexity and intention to use was weak and not statistically significant, while the direct path became positive after perceived ease of use was controlled. This pattern indicates that techno-complexity should not be interpreted as a uniform negative predictor of intention to use. Rather, its negative role appears through the usability pathway: when complexity reduces perceived ease of use, it indirectly weakens acceptance readiness. The positive direct path may suggest a suppression pattern, or that developers who perceive technologies as complex may still recognise their professional relevance once ease of use is accounted for.

4.3 Exploratory Extension

The exploratory analysis examined whether techno-insecurity and techno-uncertainty provided additional explanatory value beyond techno-complexity. When techno-complexity, techno-insecurity, techno-uncertainty, and perceived ease of use were included as predictors of intention to use, perceived ease of use remained the strongest predictor (β = 0.735, p < .001). Techno-insecurity had a significant negative effect on

intention to use (β = -0.215, p = 0.035), while techno-uncertainty had a small positive effect (β = 0.116, p = 0.042). Techno-complexity also showed a positive direct effect in this extended model (β = 0.267, p = 0.003), confirming that its role is not adequately captured by a simple negative direct relationship.

5 Discussion

The results support the proposed mediation mechanism and clarify the role of techno-complexity in developers' acceptance of emerging development technologies. Techno-complexity had a significant negative effect on perceived ease of use, perceived ease of use had a significant positive effect on intention to use, and the indirect effect of techno-complexity on intention to use through perceived ease of use was statistically significant. These findings show that techno-complexity is not simply a general negative attitude toward technology, but a specific usability-related barrier. In other words, complexity becomes problematic when it makes technologies harder to understand, learn, or integrate into development work. At the same time, the positive direct path from techno-complexity to intention to use suggests a possible dual role of complexity in developer contexts. However, this interpretation should be treated cautiously. For developers, complex technologies may also signal advanced functionality, professional relevance, or opportunities for competence development. The negative indirect effect shows that complexity weakens acceptance when it reduces perceived ease of use. The main contribution of the study is therefore not the claim that techno-complexity uniformly decreases intention to use, but that it operates through a specific usability pathway.

This finding is relevant in the context of increasingly AI-assisted software engineering. Recent work shows that generative AI is becoming embedded in programming, requirements engineering, and broader software engineering tasks, but also that its value depends on developers' ability to understand, evaluate, and integrate AI-generated outputs into existing work practices (Cheng et al., 2025; Feuerriegel et al., 2024; Russo, 2024; Sengar et al., 2025). The present study does not directly measure generative AI adoption, and its empirical focus is broader than a specific AI coding assistant or generative AI system. Nevertheless, the results are relevant for such environments because AI-assisted tools may intensify the same usability, learning, and verification demands that are captured by techno-complexity.

The findings also contribute to the discussion on organisational AI capability and responsible digital transformation. Organisations do not build digital capability only by acquiring advanced technologies; they also need human competencies, learning routines, and implementation practices that allow employees to use such technologies effectively and critically (Mikalef & Gupta, 2021). In software development, this is especially important because emerging tools may influence not only productivity, but also code quality, security, and professional responsibility. When automated outputs are used in knowledge-intensive tasks, issues such as reliability, interpretation, and responsible use become part of the adoption process, particularly in algorithmic environments where unreliable or incorrect outputs may affect user decisions and professional outcomes (Ji et al., 2023). From this perspective, techno-complexity should be understood as a practical signal: it indicates where developers may need clearer explanations, better integration support, and more structured learning before a tool can be responsibly adopted. The exploratory results further support the central claim that not all technostress dimensions are equal. Techno-insecurity and techno-uncertainty did not operate in the same way as techno-complexity. While techno-complexity primarily acted as a usability-related barrier, techno-insecurity appeared to reduce willingness to use emerging technologies, and techno-uncertainty showed a small positive effect that may reflect adaptation pressure in fast-changing technological environments rather than simple resistance. This distinction is important because generic technology adoption interventions may miss the real source of strain. If the problem is complexity, organisations should reduce cognitive and usability barriers. If the problem is insecurity, they should address professional confidence, role clarity, and competence development. If the problem is uncertainty, they should provide clearer communication about technological change and support employees in managing continuous adaptation.

The practical implication is that organisations should assess the "complexity budget" of emerging development technologies. This means evaluating how much new knowledge, workflow change, verification effort, and cognitive load a tool requires from developers. Such an assessment is particularly important when introducing highly automated or intelligent development tools, in which developers must not only operate the system but also assess the quality, relevance, and risks of its outputs. Responsible implementation should therefore include gradual introduction, targeted training, peer learning, feedback mechanisms, and explicit guidance on the limits of automated suggestions. In this sense, the complexity budget is not only a usability issue, but also part of responsible digital behaviour in algorithmic environments. When development tools become more automated, developers need not only technical skills but also awareness, trust calibration, and critical engagement with automated outputs. Prior work on algorithmic environments shows that user awareness, trust, and digital responsibility shape how individuals interpret personalised or automated content (Gombar & Boban, 2025; Gombar et al., 2026). Similarly,

concerns about generative AI implementation highlight that uncritical adoption may create organisational and cybersecurity risks when responsibility for automated outputs is not clearly understood (Humphreys et al., 2024). This aligns with human-centred approaches to generative AI in software engineering, which emphasise that productivity gains should not be separated from developer experience, accountability, and responsible professional practice (Russo et al., 2024). The study has several limitations. First, it relies on self-reported survey data, which capture developers' perceptions rather than observed tool use. Second, the cross-sectional design does not allow causal conclusions. Third, the study examines emerging development technologies broadly and does not directly measure the adoption of a specific AI coding assistant or generative AI system. The results should therefore not be interpreted as direct evidence about generative AI adoption, but they still identify techno-complexity as a specific usability-related barrier in developers' acceptance of emerging development technologies.

Future research should examine how this mechanism operates in concrete AI-assisted development settings, such as code generation, requirements engineering, testing, or software architecture. Longitudinal studies could investigate whether techno-complexity decreases as developers gain experience with a tool, or whether it remains a persistent barrier when tools evolve rapidly. Experimental or mixed-method research could also test whether training, transparency, peer support, or better workflow integration reduces the negative effect of techno-complexity on perceived ease of use. In this way, future studies can move beyond asking whether developers adopt emerging technologies and examine how organisations can make adoption more understandable, responsible, and sustainable.

6 Conclusion

This paper examined techno-complexity as a specific barrier to developers' acceptance of emerging development technologies. The findings show that higher techno-complexity is associated with lower perceived ease of use and that, through this usability pathway, it weakens intention to use. The study, therefore, supports the argument that technostress should not be treated as a single uniform obstacle to technology adoption.

The paper's contribution is its focused explanation of how complexity affects developer readiness. Rather than suggesting that emerging technologies are rejected simply because they are new or demanding, the findings indicate that complexity becomes problematic when it makes tools feel harder to understand and use. This distinction is important for organisations introducing increasingly intelligent development environments. For practice, the key message is that responsible technology adoption requires more than access to advanced tools. Organisations should manage the complexity budget of new development technologies by supporting developers with training, gradual implementation, clear communication, and human-centred integration into work practices. As software engineering environments become more automated and AI-assisted, developer readiness will depend not only on technological capability but also on whether complexity is translated into usable, understandable, and responsible professional practice.

References

Barke, S., James, M. B., & Polikarpova, N. (2023). Grounded Copilot: How programmers interact with code-generating models. *Proceedings of the ACM on Programming Languages, 7*(OOPSLA1), Article 78, 85–111. https://doi.org/10.1145/3586030

Chang, P.-C., Zhang, W., Cai, Q., & Guo, H. (2024). Does AI-driven technostress promote or hinder employees' artificial intelligence adoption intention? A moderated mediation model of affective reactions and technical self-efficacy. *Psychology Research and Behavior Management, 17,* 413–427. https://doi.org/10.2147/PRBM.S441444

Cheng, H., Husen, J. H., Lu, Y., Racharak, T., Yoshioka, N., Ubayashi, N., & Washizaki, H. (2025). Generative AI for requirements engineering: A systematic literature review. *Software: Practice and Experience*. https://doi.org/10.1002/spe.70029

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly, 13*(3), 319–340. https://doi.org/10.2307/249008

Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P. (2024). Generative AI. *Business & Information Systems Engineering, 66*(1), 111–126. https://doi.org/10.1007/s12599-023-00834-7

Gombar, M., & Boban, M. (2025). Research on the impact of algorithmic echo chambers on perceptions and attitudes of social network users in a digital society. In 2025 MIPRO 48th ICT and Electronics Convention (MIPRO) (pp. 1026–1033). IEEE. DOI: 10.1109/MIPRO65660.2025.11131918.

Gombar, M., & Boban, M. (2026). Algorithmic awareness and digital responsibility: The role of platform trust and digital literacy. *Business Systems Research*, 17(1), 179–203. https://doi.org/10.2478/bsrj-2026-0009

Gombar, M., Boban, M., & Pejić Bach, M. (2026). Digital trust and phygital responsibility: A user-centered model for sustainable consumer behavior in algorithmic environments. *World, 7,* Article 86. https://doi.org/10.3390/world7050086

Hayes, A. F. (2022). *Introduction to mediation, moderation, and conditional process analysis: A*

regression-based approach (3rd ed.). The Guilford Press.

Högemann, M., Hein, L., Britsche, J.-O., & Thomas, O. (2025). Technostress and generative AI in the workplace: A qualitative analysis of young professionals. *Frontiers in Artificial Intelligence, 8*, Article 1728881. https://doi.org/10.3389/frai.2025.1728881

Humphreys, D., Koay, A., Desmond, D., & Mealy, E. (2024). AI hype as a cyber security risk: The moral responsibility of implementing generative AI in business. *AI and Ethics, 4*, 791–804. https://doi.org/10.1007/s43681-024-00443-4

Ji, Z., Lee, N., Frieske, R. M., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), Article 248. https://doi.org/10.1145/3571730

Mikalef, P., & Gupta, M. (2021). Artificial intelligence capability: Conceptualization, measurement calibration, and empirical study on its impact on organizational creativity and firm performance. *Information & Management, 58*(3), Article 103434. https://doi.org/10.1016/j.im.2021.103434

Nastjuk, I., Trang, S., Grummeck-Braamt, J.-V., Adam, M. T. P., & Tarafdar, M. (2024). Integrating and synthesising technostress research: A meta-analysis on technostress creators, outcomes, and IS usage contexts. *European Journal of Information Systems, 33*(3), 361–382. https://doi.org/10.1080/0960085X.2022.2154712

Nguyen-Duc, A., Cabrero-Daniel, B., Przybylek, A., Arora, C., Khanna, D., Herda, T., Rafiq, U., Melegati, J., Guerra, E., Kemell, K.-K., Saari, M., Zhang, Z., Le, H., Quan, T., & Abrahamsson, P. (2025). Generative artificial intelligence for software engineering—A research agenda. *Software: Practice and Experience*. https://doi.org/10.1002/spe.70005

Pearce, H., Ahmad, B., Tan, B., Dolan-Gavitt, B., & Karri, R. (2025). Asleep at the keyboard? Assessing the security of GitHub Copilot's code contributions. *Communications of the ACM, 68*(2), 96–105. https://doi.org/10.1145/3610721

Perry, N., Srivastava, M., Kumar, D., & Boneh, D. (2023). Do users write more insecure code with AI assistants? In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security* (pp. 2785–2799). Association for Computing Machinery. https://doi.org/10.1145/3576915.3623157

Prather, J., Reeves, B. N., Denny, P., Becker, B. A., Leinonen, J., Luxton-Reilly, A., Powell, G., Finnie-Ansley, J., & Santos, E. A. (2024). "It's weird that it knows what I want": Usability and interactions with Copilot for novice programmers. *ACM Transactions on Computer-Human Interaction, 31*(1), Article 4. https://doi.org/10.1145/3617367

Ragu-Nathan, T. S., Tarafdar, M., Ragu-Nathan, B. S., & Tu, Q. (2008). The consequences of technostress for end users in organizations: Conceptual development and empirical validation. *Information Systems Research, 19*(4), 417–433. https://doi.org/10.1287/isre.1070.0165

Russo, D. (2024). Navigating the complexity of generative AI adoption in software engineering. *ACM Transactions on Software Engineering and Methodology, 33*(5), Article 135, 1–50. https://doi.org/10.1145/3652154

Russo, D., Baltes, S., van Berkel, N., Avgeriou, P., Calefato, F., Cabrero-Daniel, B., Catolino, G., Cito, J., Ernst, N., Fritz, T., Hata, H., Holmes, R., Izadi, M., Khomh, F., Kjærgaard, M. B., Liebel, G., Lafuente, A. L., Lambiase, S., Maalej, W., ... Vasilescu, B. (2024). Generative AI in software engineering must be human-centered: The Copenhagen Manifesto. *Journal of Systems and Software, 216*, Article 112115. https://doi.org/10.1016/j.jss.2024.112115

Sengar, S. S., Hasan, A. B., Kumar, S., & Carroll, F. (2025). Generative artificial intelligence: A systematic review and applications. *Multimedia Tools and Applications, 84*, 23661–23700. https://doi.org/10.1007/s11042-024-20016-1

Tarafdar, M., Cooper, C. L., & Stich, J. F. (2019). The technostress trifecta—Techno eustress, techno distress and design: Theoretical directions and an agenda for research. *Information Systems Journal, 29*(1), 6–42. https://doi.org/10.1111/isj.12169

Vaithilingam, P., Zhang, T., & Glassman, E. L. (2022). Expectation vs. experience: Evaluating the usability of code generation tools powered by large language models. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. Association for Computing Machinery. https://doi.org/10.1145/3491101.3519665

Venkatesh, V., & Davis, F. D. (2000). A theoretical extension of the Technology Acceptance Model: Four longitudinal field studies. *Management Science, 46*(2), 186–204. https://doi.org/10.1287/mnsc.46.2.186.11926

Weisz, J. D., Kumar, S. V., Muller, M., He, J., Houde, S., Ross, S. I., Martinez, F., Agarwal, M., Talamadupula, K., & Richards, J. T. (2025). Examining the use and impact of an AI code assistant on developer productivity and experience in the enterprise. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery. https://doi.org/10.1145/3706599.3706670

GIS Support for Managerial Decision-Making in Enhancing Cycling Tourism Routes

Peter Márton, Romana Vaščáková
University of Žilina
Faculty of Management Science and Informatics
Department of Mathematical Methods and Operations Research
Univerzitná 8215/1, 01026 Žilina, Slovakia
{peter.marton@uniza.sk, vascakova1}@stud.uniza.sk

Abstract. *Cycling tourism is a form of recreational travel in which individuals explore destinations by bicycle. Interest in this mode of transport has been growing due to environmental awareness, healthy lifestyles, and continuous investments in cycling infrastructure. This article presents a comprehensive literature review focusing on the environmental, health, and economic importance of cycling, while highlighting the critical role of infrastructure in its development. It synthesizes selected case studies on the application of geographic information systems (GIS) and multi-criteria decision-making (MCDM) methods for planning and evaluating cycling routes through spatial analyses. Finally, the paper outlines an ongoing research framework utilizing Strava Metro data and GIS-MCDM modelling on the Váh Cycle Route, establishing the baseline for future spatial and cost-benefit evaluations.*

Keywords. cycling tourism, cycling infrastructure, geographic information systems, sustainable mobility

1 Definition of Cycling Tourism

According to Greene, cycling tourism is a form of recreational travel in which individuals explore destinations by bicycle. This definition applies to all types of cycling tourism, whether short, independent trips or organised multi-day group cycling tours through urban or rural areas, regardless of whether a conventional or electric bicycle is used. (Greene, 2024)

Weston et al. from the European Parliament define cycling tourism as travelling between different places by bicycle for recreational purposes, with cycling itself forming an integral part of the tourist experience. (Weston et al., 2012)

Interest in cycling tourism is constantly growing, mainly due to increasing awareness of environmental issues, healthy lifestyle and investments in cycling infrastructure. (Greene, 2024; Grand View Research, 2026)

1.1 Cycling as an environmentally friendly mode of transport

Cycling and cycling tourism represent environmentally friendly modes of transport that contribute to reducing carbon emissions and supporting sustainable mobility. (Functional Areas in the EU, 2023) Within the so-called sustainable transport hierarchy, walking and cycling are among the most environmentally friendly modes of movement, while individual car and air transport are at the opposite end of the spectrum in terms of environmental impact, as can be seen in Fig. 1. (Functional Areas in the EU, 2023; Philips et al., 2020)

Figure 1. Sustainable transport hierarchy (Philips et al., 2020)

According to estimates by CREDS (Centre for Research into Energy Demand Solutions), CO_2 emissions over the life cycle of an electric bicycle are significantly lower compared to other modes of transport, reaching approximately 22 g/km. In contrast, emissions from electric cars are approximately 104 g/km, from hybrid cars 168 g/km and from petrol cars 258 g/km. (Philips et al., 2020)

Another advantage of cycling is its low spatial demand. In one hour, approximately 14,000 cyclists can pass through a 3.5-metre-wide section in a city, whereas in the case of car transport, this figure is only

around 2,000 people. Cycling transport is therefore more efficient in terms of space utilisation. In addition, up to 15 bicycles can be parked in a single car parking space. (European Cyclists' Federation, 2018; Dekoster & Schollaert, 1999)

1.2 Cycling as a tool for health promotion

In addition to their positive impact on the environment, cycling and cycling tourism also provide health benefits. Regular physical activity plays a key role in the prevention of chronic non-communicable diseases, such as cardiovascular diseases, diabetes, hypertension, obesity and certain types of cancer. At the same time, they have a positive effect on mental health. (Functional Areas in the EU, 2023; European Cyclists' Federation, 2018)

In 2014, Kelly et al. published a comprehensive systematic review and meta-analysis summarising the results of studies examining the relationship between cycling and the risk of all-cause mortality (ACM) among healthy adults. Based on a sample of approximately 200,000 adults aged 20–93 years, they found that cycling at a physical activity level of 11.25 MET-hours per week, corresponding to approximately 116 minutes of cycling per week at a speed of 15.1 km/h, was associated with a 10% reduction in mortality risk. (Logan et al., 2023; Kelly et al., 2014; Ainsworth et al., 2011) Based on these studies, it can be concluded that the promotion of cycling and cycle tourism is also important from the perspective of public health.

Hendriksen et al., in a study conducted on a sample of 1,236 Dutch employees, found that regular commuting to work by bicycle was associated with a lower rate of sickness absence. They found that, during the observed year, the average length of work absence among cyclists was more than one day shorter compared with non-cyclists. Cyclists were absent from work for an average of 7.4 days in the observed year, whereas non-cyclists were absent for 8.7 days. At the same time, the more frequently and the longer employees cycled, the fewer days they were absent from work. (European Cyclists' Federation, 2018; Hendriksen et al., 2010)

1.3 Investments in cycling infrastructure

The systematic development and modernisation of cycling infrastructure also have a significant impact on the growing interest in cycling tourism. These processes are closely related to the objectives of the European Green Deal, which aims to achieve a 90% reduction in transport-related greenhouse gas emissions by 2050. The development of cycling infrastructure plays an important role in achieving this objective. (Functional Areas in the EU, 2023; European Commission, n.d.)

1.4 Economic benefits of cycling

Cycling and cycle tourism also bring economic benefits, among others. In 2025, the global cycling tourism market reached a value of USD 146.19 billion, with Europe representing the largest market. The global cycle tourism market is expected to reach a value of approximately USD 332.48 billion by 2033, with an estimated annual growth rate of 11.1% from 2026 to 2033. (Grand View Research, 2026)

The EU Cycle project, implemented under the INTERREG programme, has identified and quantified the annual economic benefits associated with cycling. These benefits include a reductions in CO_2 emissions, a reduction of noise and air pollution, fuel savings, longer and healthier lives, a reduction in the number of employee sickness absence days, the expansion of the cycling market, the development of cycle tourism, a reduction in traffic congestion and lower costs for the maintenance and construction of road infrastructure for motor vehicles.

The total annual economic benefit of cycling in the European Union is estimated at between EUR 150 and EUR 155 billion, with the most significant share coming from health benefits and the development of cycle tourism, as shown in Fig. 2. The calculation of these benefits is based on an annual cycling volume of 146 billion kilometres in the EU-28 countries. (Functional Areas in the EU, 2023; European Cyclists' Federation, 2018)

Benefit	Estimated Value (billion euros)
CO2 emissions savings	0.6 – 5.6
Reduction of air pollution	0.435
Reduction of noise pollution	0.3
Fuel savings	4.0
Longer and healthier lives	73
Less sickness absence at the workplace	5
Bicycle market	13,2
Cycle tourism	44
Easing of road congestion	6,8
Saving on construction and maintenance costs for road infrastructure for motorised vehicles	2,9
Total annual benefits	150 - 155 bn euros

Figure 2. Estimate of the annual economic benefits of cycling in the European Union (Functional Areas in the EU, 2023)

An example of the economic benefits of cycling is the 750 km network of cycle superhighways in the Hovedstaden region of Denmark. This project, with a total investment of approximately EUR 29 million, generates a social benefit of up to EUR 76 million, while its internal rate of return reaches 11%. According to economic analyses of Danish infrastructure projects, the cycle superhighway network has also been ranked among the top five most profitable projects. (Garrett, 2019)

Another example is Cykelslangen, a 235-metre cycle bridge in Copenhagen, Denmark, which, with construction costs of EUR 4.2 million, generated benefits of EUR 5.9 million over the observed period, while the internal rate of return of this project reached 9%. The study also estimated that a shift from car transport to cycling brings savings of at least EUR 1.34 for every kilometre travelled. These data confirm that investments in cycling infrastructure represent not only an environmentally friendly solution, but also a highly economically efficient one. (Garrett, 2019)

In 2009, Mercat et al. published a study in which they analysed the average daily expenditure of cyclotourist and tourists in France. They found that cyclotourist spend, on average, EUR 10 more per day than regular tourists, while in the case of touring cyclotourist who travel from place to place, this difference is as much as EUR 21. Higher expenditure is mainly associated with accommodation, where cyclotourist spend an average of EUR 11 more, and to food, where the difference is approximately EUR 5. Expenditure related to visits to cultural monuments and other activities is also higher, by approximately EUR 1 per day. Cycloturist who travel from place to place have the highest transport costs among all groups. Compared with regular tourists, they spend an average of EUR 5 more. The study also found that this group of cyclotourist uses rail transport up to three times more frequently than regular tourists. (Atout France, 2009) A more detailed breakdown of individual expenditures by group is shown in Fig. 3.

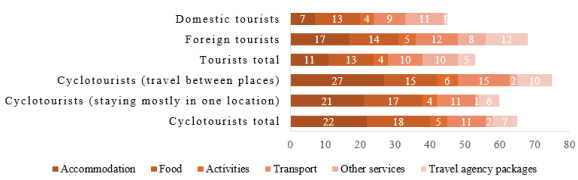


Figure 3. Comparison of daily expenditure by tourists with daily expenditure by cyclotourist (Atout France, 2009)

2 Tools and approaches for evaluating cycling infrastructure projects

Various analytical approaches and decision-making tools are used in the evaluation of projects, including those in the field of cycling infrastructure, among which are cost-benefit analysis (CBA) and multi-criteria decision-making (MCDM).

Cost-benefit analysis (CBA) is a systematic approach used to calculate and compare the benefits and costs of a specific project or decision. It examines whether the expected benefits and costs are in balance, while expressing them in terms of net present value. It can therefore be understood as a list of advantages and disadvantages that supports financial decision-making. The main advantages of CBA include its simplicity, clarity and the possibility of comparing different solution alternatives. (Al Moussawi et al., 2020) According to academics, another strength is its theoretical clarity, which makes the analysis a common language that is known and used worldwide. At the same time, it is a widely used and well-established tool for project evaluation. (Annema et al., 2015) A disadvantage is the need to use uniform units of measurement, which is impossible in the case of qualitative benefits. (Al Moussawi et al., 2020) Qualitative benefits may include impacts on equity and environmental impacts, such as the deterioration of nature and landscape caused by new infrastructure, which cannot be accurately expressed in monetary terms. (Annema et al., 2015)

Multi-criteria decision-making (MCDM), also known as multi-criteria decision analysis (MCDA) or multi-criteria analysis (MCA), is considered one of the most accurate decision-making methods and is regarded as a revolutionary approach in this field. (Annema et al., 2015; Taherdoost & Madanchian, 2023) This method considers various qualitative and quantitative criteria, such as cost, quality, or safety, which need to be defined to identify the best solution. Experts assign different weights to individual criteria based on their importance in a specific project. (Taherdoost & Madanchian, 2023) Within MCDM, there are several specific methods, such as the Analytic Hierarchy Process (AHP), Simple Additive Weighting (SAW), Weighted Sum Model (WSM), Technique for Order Preference by Similarity to Ideal Solution (TOPSIS), and the Best-Worst Method (BWM), each of which has its own advantages and disadvantages. (Taherdoost & Madanchian, 2023; Alsalem et al., 2018) The case studies included in this article used the BWM and TOPSIS methods. To process and visualize the spatial criteria of these models, the selected MCDM approaches were implemented within a GIS environment.

2.1 Application of CBA in the Netherlands

In 2000, the Dutch government introduced CBA as a mandatory tool to support decision-making on major transport infrastructure projects. At that time, MCDM was rejected for two main reasons. The first reason was the assignment of weights to individual criteria, which is not always clear. The second reason was the higher probability of double-counting project effects, as MCDM lacks strictly defined criteria for their inclusion compared to CBA. (Annema et al., 2015)

In 2015, J. A. Annema, N. Mouter, and J. Razaei published a study in which they examined the opinions of 21 politicians working in the field of transport on CBA. They found that politicians do use CBA, but not as a decisive tool, rather, they use it only as one of the

inputs in decision-making. According to the politicians, CBA is not sufficiently transparent, and the net present value indicator does not provide a complete picture. As an example, they mentioned the expansion of Amsterdam Schiphol Airport, where they considered it more important to openly discuss whether, for instance, the creation of new jobs is more important than increased noise for residents living nearby, rather than focusing only on one specific number. (Annema et al., 2015)

Based on interviews with politicians, the authors formulated three main considerations. The first is that CBA can be considered a useful tool, as it is used by most politicians. At the same time, however, it is not a decisive tool, politicians approach it critically and perceive it more as one of several inputs in the decision-making process. (Annema et al., 2015)

Second, the results do not indicate that MCDM is better than CBA. (Annema et al., 2015)

The third consideration is that a combination of CBA and MCDM is not an ideal solution. Combining these two different methods could further reduce the transparency and credibility of the results. (Annema et al., 2015)

The authors and politicians see an alternative in a simpler approach used in the United Kingdom and Switzerland, which is based on presenting politicians not with a single indicator, but with several criteria to which no weights are assigned. Politicians then decide on the individual criteria and seek compromises. (Annema et al., 2015)

From the perspective of the authors and politicians, an interesting solution is also the so-called “trade-off sheet” - a list of the most important trade-offs. In this case, CBA could provide information on economic efficiency, while MCDM would focus on economic or social aspects that cannot be expressed in monetary terms. The result of such a “trade-off sheet” would serve as a basis for political discussion. (Annema et al., 2015)

2.2 Application of CBA in Sweden

Similarly to the Netherlands, Sweden also uses CBA to support decision-making on major national transport investments. In doing so, it applies the ASEK guidelines (Arbetsgruppen för SamhällsEkonomiska Kalkyler), which are regularly updated, usually every three to five years. One of the main objectives of using the ASEK guidelines is to ensure the comparability of individual CBA analyses of transport projects, so that priorities can be effectively set among projects from different parts of the country. (Eliasson, 2013)

Since 2009, every proposal for a national transport investment must undergo a comprehensive assessment, which includes, in addition to the project proposal itself, the expected impacts, reasons for implementation, status, CBA results and effects expressed in non-monetary terms. (Eliasson, 2013)

Awareness that CBA will be used as a tool for prioritising projects has led to greater cost-effectiveness of investment proposals, which many consider to be the most significant benefit of using CBA. Planners therefore focus on the cost-effectiveness of a project already in its early stages and, to some extent, prevent inefficient and overestimated projects from progressing to further stages of preparation. Compared with Norway, which does not rely on CBA in project selection, Sweden has a lower share of socially unprofitable projects, i.e. projects whose benefit-cost ratio (BCR) is lower than 1. (Eliasson, 2013)

2.3 Different application of MCDM and GIS in cycling infrastructure planning

Geographic information systems (GIS) are also used, among other tools, in the planning and modernisation of cycling infrastructure. An example is a case study from the Italian city of Palermo, which focused on the use of GIS in assessing the potential of multimodal transport based on a combination of cycling and rail transport. The main objective of this case study was to reduce individual car transport in the city and support more environmentally friendly alternatives. In the case study, the authors focused on three residential districts on the outskirts of Palermo with a low frequency of public transport services and an absence of cycle routes and shared bicycles. They developed a GIS-based methodology that made it possible to assess the potential for introducing new cycle routes and bike-sharing stations in connection with the existing public transport infrastructure. When designing new cycle routes, the authors considered several factors, such as road width, gradient, traffic density, cycle crossings and distance from public transport stops. Attention was paid to the Go2School bike-sharing programme and to secondary school students. Based on a questionnaire survey, the authors found that the introduction of new cycle routes and bike-sharing stations would increase the percentage of students cycling on their own or shared bicycles from 1.5% to 10.1%, while reducing the percentage of students commuting to school by private car from 21% to 15.7%. An increase was also recorded in the use of public transport. (Capodici et al., 2021)

Another example is a case study focusing on the Veneto region in north-eastern Italy. Its aim was to support the development of cycling tourism as part of sustainable tourism in rural areas. In selecting and evaluating the area, the authors considered environmental, cultural and economic factors. These factors were subsequently analysed using GIS and MCDA, which made it possible to comprehensively assess the suitability of the area for the development of cycling tourism. Based on the analysis, areas with the greatest potential for the construction of cycle routes were identified, while not only the accessibility of the area was evaluated, but also the existing infrastructure,

cultural value and points of interest in the surrounding area. (Parlato & Pezzuolo, 2025)

In the Turkish city of Istanbul, the authors Guler and Yomralioglu used a combination of GIS and the BWM to create an integrated framework for the optimal placement of bike-sharing stations and cycling infrastructure. The case study focused on six districts of Istanbul and considered eight key criteria, such as proximity to public parks, shopping centres, educational institutions, existing cycling infrastructure, public transport stops, bus routes, population density and terrain slope. The BWM method made it possible to effectively determine the relative importance of individual criteria through pairwise comparisons, in which the best and worst factors are evaluated against all the others. The best factor was proximity to public transport stops, while the worst factor was proximity to shopping centres. The authors subsequently created raster layers for each criterion in a GIS environment and normalised them to a value range from 0 to 1, where 0 represented an unsuitable location and 1 represented a highly suitable location. Using the weighted linear combination (WLC) method, these layers were merged into a final suitability map. Based on this map, the most suitable locations for bike-sharing stations were identified, as well as new routes extending the cycling infrastructure in Istanbul. The proposed station locations were ranked according to their suitability using the TOPSIS. Finally, a sensitivity analysis was carried out, confirming the stability of the results even when the weights of individual criteria were changed. (Guler & Yomralioglu, 2021)

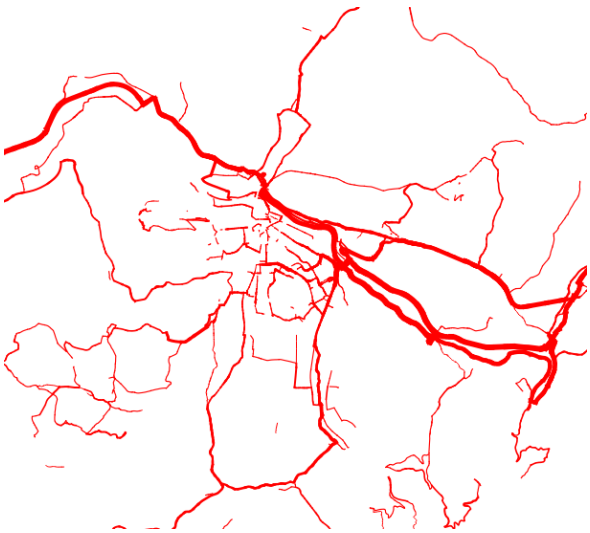


Figure 4. Bicycle traffic density around Žilina based on Strava Metro data; line thickness represents cyclist volume (Strava Metro data, 2025)

This picture shows aggregated and de-identified data from Strava Metro.

3 Spatial Decision Support Framework for Evaluating the Váh Cycle Route

Research focusing on the Váh Cycle Route in Slovakia utilising GIS and MCDM commenced in March 2026. This study centres on integrating crowdsourced Strava Metro data into a spatial analytical environment (Márton & Blažek, 2026; Márton et al., 2026) to evaluate cycling infrastructure and support evidence-based planning. By combining actual movement patterns with other spatial datasets, the framework aims to identify heavily used corridors, assess infrastructure performance, and serve as an empirical baseline for subsequent CBA and strategic managerial decision-making.

The Váh Cycle Route, which is the oldest and longest cycling tourism route in Slovakia, spans near 250 km along the Váh River across four Self-Governing Regions (SGR): Žilina (41.2 km), Trenčín (97.9 km), Trnava (43.2 km), and Nitra (65.7 km). While the route is fully signed in the field, its infrastructure has not yet been completed as a continuous, safe, and segregated cycleway. Construction is currently progressing in stages within individual SGRs. Upon completion across all four regions, it will become the longest continuous cycleway in Slovakia, impacting millions of users—as evidenced by over 3.5 million cyclists recorded on its completed sections in the Trenčín SGR since 2018 (Mazreku, 2025; SITA, 2026).

3.1 Definition of the Research Problem

Strategic planning and managerial decision making in cycling infrastructure development often suffer from a lack of objective, continuous, and long-term empirical data on actual user behavior. Traditional data collection methods, such as manual counts or localised automatic counters, provide geographically isolated snapshots that are insufficient for evaluating multi-jurisdictional long-distance corridors like the Váh Cycle Route.

While crowdsourced Big Data platforms (e.g., Strava Metro) offer high-resolution mobility insights, utilising them directly in planning presents a key analytical challenge: raw spatial data require a standardized, long-term analytical baseline to ensure consistent multi-year monitoring (2021–2026 in our case) across evolving territorial conditions. Consequently, regional decision-makers lack a unified spatial decision-support framework to objectively evaluate multi-year mobility shifts, identify infrastructural bottlenecks, and prioritise investments along strategic cycling corridors.

3.2 Research Methodology

To address this decision making challenge, we propose an integrated spatial decision support system combining GIS, crowdsourced Big Data, and spatial MCDM. The pilot application focuses on the Váh Cycle Route, with the objective of validating a methodology that can subsequently be scaled to other national and international cycling routes. The methodology consists of three main phases:

1. Topology-Independent Spatial Aggregation: To enable consistent longitudinal monitoring (2021–2026), crowdsourced movement data from Strava Metro will be aggregated into the discrete global grid system Uber H3 (Brodsky et al., 2018).
2. Integration of Behavioral and Spatial Demographics: The aggregated spatial units will be enriched with multi-dimensional mobility attributes, including directional flow volumes, commuting vs. recreational travel purpose, e-bike share, age cohort distributions, and time-of-day temporal profiles.
3. Spatial MCDM and CBA Decision Framework: The spatial mobility profiles derived from the GIS analysis will serve as empirical baseline criteria within a broader managerially oriented framework. These empirical indicators (actual demand, spatial connectivity, user volume trends) will be combined with qualitative criteria (traffic safety, service proximity) using spatial MCDM techniques (Best-Worst Method - BWM and TOPSIS) and subsequent Cost-Benefit Analysis (CBA). This integrated approach translates complex spatial Big Data into prioritised strategic recommendations for local and regional authorities.

3.3 Scientific Contribution

The scientific contribution of our study lies in bridging the gap between crowdsourced spatial Big Data and strategic managerial decision making in mobility planning:

- It establishes a reproducible spatial assessment methodology that enables us to conduct reliable multi-year longitudinal monitoring (2021–2026) of cycling demand along the Váh Cycle Route and other international, national and regional cycle routes.
- It demonstrates how crowdsourced spatial mobility indicators can be effectively integrated as objective empirical weights into spatial MCDM and CBA models, replacing subjective estimates in transport planning.
- It delivers an evidence-based decision support model applied to the Váh Cycle Route, providing regional SGRs with actionable spatial insights to prioritise missing infrastructure segments, target safety interventions, and optimize the location of complementary cycling services.

Conclusion

This paper highlights the multifaceted importance of cycling tourism from environmental, health, and economic perspectives, demonstrating that sustainable mobility planning requires an integrated, data-driven approach. Through a systematic literature review, it confirmed that the effective development and evaluation of cycling infrastructure depend heavily on GIS combined with decision-making frameworks such as CBA and MCDM. While extensive national cycle route networks exist, such as in Slovakia, their completion as continuous, safe, and segregated cycleways remains a critical challenge for regional authorities.

To bridge the gap between theoretical frameworks and practical infrastructure appraisal, this paper shortly introduces an ongoing research framework centred on the Váh Cycle Route. We propose utilising the H3 grid system. By converting Strava Metro movement patterns (2021–2026) into H3 grid units and enriching them with behavioural and demographic attributes, the framework establishes a reliable spatial baseline.

The primary scientific contribution of our future research lies in demonstrating how crowdsourced spatial Big Data can be transformed into objective, empirical weights within spatial MCDM and CBA models. Rather than relying on subjective assumptions or isolated manual counts, this integrated approach enables to conduct consistent longitudinal monitoring, identify infrastructure bottlenecks, and optimise the spatial distribution of complementary services.

Finally, embedding crowdsourced spatial analytics into regional decision support systems provides local and regional managers with an evidence based tool for strategic planning. This approach ensures that future investments in cycling infrastructure along the Váh Cycle Route and similar international, national or regional cycle routes maximize public health, environmental sustainability, and economic returns.

References

Ainsworth, B. E., Haskell, W. L., Herrmann, S. D., Meckes, N., Bassett, D. R., Jr., Tudor-Locke, C., Greer, J. L., Vezina, J., Whitt-Glover, M. C., & Leon, A. S. (2011). 2011 compendium of physical activities: A second update of codes and MET values. *Medicine & Science in Sports & Exercise*, 43(8), 1575–1581. https://doi.org/10.1249/MSS.0b013e31821ece12

Al Moussawi, H., Fardoun, F., & Louahlia, H. (2020). Optimization and decision-making methods in realization of tri-generation systems. *Towards Sustainable Chemical Processes: Applications of Sustainability Assessment and Analysis, Design and Optimization, and Hybridization and Modularization* (pp. 43-67). Elsevier.

https://doi.org/10.1016/B978-0-12-818376-2.00002-8

Alsalem, M. A., Zaidan, A. A., Zaidan, B. B., Hashim, M., Albahri, O. S., Albahri, A. S., Mohsin, H. A. & Mohammed, K. I. (2018). Systematic review of an automated multiclass detection and classification system for acute leukaemia in terms of evaluation and benchmarking, open challenges, issues and methodological aspects. *Journal of Medical Systems*, 42, Article 204. https://doi.org/10.1007/s10916-018-1064-9

Annema, J. A., Mouter, N., & Razaei, J. (2015). Cost-benefit analysis (CBA), or multi-criteria decision-making (MCDM) or both: Politicians' perspective in transport policy appraisal. *Transportation Research Procedia*, 10, 788–797. https://doi.org/10.1016/j.trpro.2015.09.032

Atout France. (2009). *Spécial économie du vélo: Étude complète*. https://ekladata.com/iSZpvbgl7Dsa5xeFa4h4iYNf7B0/2009-PDF-economie-velo.pdf

Brodsky, I. (2018, June 27). H3: Uber's hexagonal hierarchical spatial index. *Uber Engineering Blog.* https://www.uber.com/blog/h3/

Capodici, A. E., D'Orso, G., & Migliore, M. (2021). A GIS-based methodology for evaluating the increase in multimodal transport between bicycle and rail transport systems: A case study in Palermo. *ISPRS International Journal of Geo-Information*, 10(5), Article 321. https://doi.org/10.3390/ijgi10050321

Dekoster, J., & Schollaert, U. (1999). Cycling: The way ahead for towns and cities. *Office for Official Publications of the European Communities.* https://aa.ecn.cz/img_upload/3a7e18c1249b899407e75e7f626db792/cycling_en.pdf

Eliasson, J. (2013). *International comparison of transport appraisal practice: Annex 4 Sweden country report.* Institute for Transport Studies. https://assets.publishing.service.gov.uk/media/5a7ba777ed915d1311060a4c/annex-4-sweden.pdf

European Commission. (n.d.). *Sustainable transport*. Mobility and Transport. https://transport.ec.europa.eu/transport-themes/sustainable-transport_en

European Cyclists' Federation. (2018). *The benefits of cycling: Unlocking their potential for Europe*. https://www.ecf.com/media/resources/TheBenefitsOfCycling_final-v2.pdf

Functional Areas in the EU. (2023). *Developing world-class cycling infrastructure at the functional area level*. https://functionalareas.eu/wp-content/uploads/2023/08/Cycling-Infrastructure-at-the-FA-level.pdf

Garrett, A. (2019). *Cost-benefit of cycling infrastructure*. Cycling Embassy of Denmark. https://cyclingsolutions.info/cost-benefit-of-cycling-infrastructure/

Grand View Research. (2026). *Cycle tourism market size, share & trends analysis report by group, by booking mode, by age group, by region, and segment forecasts, 2026–2033*. https://www.grandviewresearch.com/industry-analysis/cycle-tourism-market-report

Greene, J. (2024). *Bicycle tourism*. EBSCO Research Starters. https://www.ebsco.com/research-starters/history/bicycle-tourism

Guler, D., & Yomralioglu, T. (2021). Location evaluation of bicycle sharing system stations and cycling infrastructures with Best Worst Method using GIS. *The Professional Geographer*, 73(3), 535–552. https://doi.org/10.1080/00330124.2021.1883446

Hendriksen, I. J. M., Simons, M., Galindo Garre, F., & Hildebrandt, V. H. (2010). The association between commuter cycling and sickness absence. *Preventive Medicine*, 51(2), 132–135. https://doi.org/10.1016/j.ypmed.2010.05.007

Kelly, P., Kahlmeier, S., Götschi, T., Orsini, N., Richards, J., Roberts, N., Scarborough, P., & Foster, C. (2014). Systematic review and meta-analysis of reduction in all-cause mortality from walking and cycling and shape of dose response relationship. *International Journal of Behavioral Nutrition and Physical Activity*, 11, Article 132. https://doi.org/10.1186/s12966-014-0132-x

Logan, G., Somers, C., Baker, G., Connell, H., Gray, S., Kelly, P., McIntosh, E., Welsh, P., Gray, C. M., & Gill, J. M. R. (2023). Benefits, risks, barriers, and facilitators to cycling: A narrative review. *Frontiers in Sports and Active Living*, 5, Article 1168357. https://doi.org/10.3389/fspor.2023.1168357

Márton, P., & Blažek, B. (2026). Architektúra open-source infraštruktúry pre spracovanie a priestorovú optimalizáciu geografických big data. *Proceedings of the 14th Conference on Open Source Software in Education, Research and IT Solutions (OSSConf 2026)* (pp. 97–106). University of Žilina

Márton, P., Gogola, M., & Vaščáková, R. (2026). Využitie dát Strava pre rozvoj cyklodopravy a projektovú výučbu v oblasti geoinformatiky. *Proceedings of the 14th Conference on Open Source Software in Education, Research and IT Solutions (OSSConf 2026)* (pp. 106–116). University of Žilina

Mazreku, A. (2025, December 30). *Výstavba unikátnej cyklotrasy napreduje. Vážska magistrála spája tri kraje, pri Ilave stavajú kľúčový úsek.* TA3. https://www.ta3.com/clanok/1026771/vystavba-unikatnej-cyklotrasy-napreduje-vazska-magistrala-spaja-tri-kraje-pri-ilave-stavaju-klucovy-usek

Parlato, M. C. M., & Pezzuolo, A. (2025). Rural–urban linkages and development: A GIS-MCDA approach to sustainable cyclo-tourism and rural land protection. *Land Degradation & Development*, 36(14), 4809–4826. https://doi.org/10.1002/ldr.5669

Philips, I., Anable, J., & Chatterton, T. (2020). *E-bike carbon savings – How much and where?* CREDS. https://www.creds.ac.uk/publications/e-bike-carbon-savings-how-much-and-where/

Taherdoost, H., & Madanchian, M. (2023). *Multi-criteria decision making (MCDM) methods and concepts.* Encyclopedia, *3*(1), 77–87. https://doi.org/10.3390/encyclopedia3010006

TASR. (2026, March 11). *Na jeseň bude dokončených 90 % stokilometrovej Vážskej cyklomagistrály*. TERAZ.sk. https://www.teraz.sk/spravy/na-jesen-bude-dokoncenych-90-stokil/947355-clanok.html

Weston, R., Davies, N., Lumsdon, L., McGrath, P., Peeters, P., Eijgelaar, E., & Piket, P. (2012). *The European cycle route network EuroVelo*. European Parliament. https://www.researchgate.net/publication/262099932_The_European_Cycle_Route_Network_EuroVelo

Age Differences in the Effectiveness of Personality-Tailored Advertising on Social Networks: Evidence from Croatia

Krunoslav Matešić
Catholic University of Croatia, Zagreb, Croatia
Full Address
krunoslav.matesic@unicath.hr

Mirko Palić
Faculty of Economics & Business, Zagreb, Croatia
Trg J. F. Kennedyja 6, Zagreb, Croatia
mirkopalic@gmail.com

Sabina Sačer
Croatian Military Academy, Croatia
Ilica 256b, Zagreb, Croatia
sabina.sacer@gmail.com

Abstract. *This study examines whether the effectiveness of personality-tailored social media advertising differs by age among Facebook users in Croatia. Two seven-day quasi-experimental Facebook campaigns promoted Shopaj24, a nationwide digital classified-advertising platform, and a dental clinic serving Zagreb. Adults aged 18 and older were segmented only using Facebook Likes associated with Big Five personality traits (Likes-user matrix) and shown either trait-congruent or non-congruent visual advertisements, with copy and nominal budget held constant. Across both campaigns, 156,127 impressions reached 102,377 users. Advertising effectiveness was assessed using click-through rates and click/non-click frequencies across three age groups (18–34, 35–54, and 55+). Age-related differences were most consistent for Extraversion and Neuroticism. The findings suggest that age can refine personality-based targeting, although effects vary by service context and personality trait.*

Keywords. Digital advertising, personality-tailored advertising, Big Five, age differences, Facebook, Croatia.

1 Introduction

Personalisation is a central advantage of digital advertising because persuasive messages tend to be more effective when they align with recipients' motives, identities, and psychological characteristics (Hirsh, Kang, and Bodenhausen, 2012; Matz et al., 2017). Age is also relevant to responses to advertising. Recent research indicates that advertising effectiveness, message preference, memory, and purchase-related judgments differ across age groups and that these differences depend on the type of appeal and product context (van der Goot, Bol and van Weert, 2021; Mecredy, Stocchi and Feetham, 2025; Chen et al., 2025). Earlier work similarly linked age to shifts in motivational priorities, information processing, and responses to emotional appeals (Williams & Drolet, 2005; Carpenter & Yoon, 2011). Age is also associated with systematic mean-level differences in Big Five traits across adulthood (Soto et al., 2011). Consequently, age may be associated with differences in responses to personality-congruent cues.

Personality has been associated with purchase intentions, browsing behaviour, and the likelihood of interacting with an advertisement (Hirsh, Kang and Bodenhausen, 2012; Matz et al., 2017; Azucar, Marengo and Settanni, 2018). However, psychological-targeting research has mainly examined whether trait-congruent messages outperform non-congruent messages, while paying less attention to whether this advantage varies across age groups. This gap is important because age-related motivational and processing differences may strengthen, weaken, or reverse the effect of the same personality-tailored cue across consumer contexts.

The primary aim of this study was to examine whether responses to personality-tailored Facebook advertisements differed across age groups among Croatian users. A second aim was to determine whether the direction of these age differences was consistent across two contrasting service contexts - a nationwide digital classified-advertising platform and a local dental clinic - and across the Big Five trait segments inferred from Facebook Likes. Accordingly, the study asked whether age was associated with the incremental effectiveness of trait-congruent versus non-congruent visual advertising.

2 Previous research

2.1. Personalisation in digital advertising

Even though users' data are available, advertising today still operates under the mass communication model, leading to an increasing demand for interactivity and personalisation (Jurković, 2021; Statista, 2025). Bašić and Đurđević (2017) point out that because the average person is exposed to 600-625 advertisements every day, things have to change. So companies have to increase their advertising budgets to adapt to changing market conditions and quickly target consumers. In her research, Sačer (2022) shows that the visual characteristics of images correlate with users' personality traits. This aligns with Matz et al. (2019) and underscores the importance of appearance and aesthetics in advertising campaigns. Utilising users' individual preferences for specific visual elements improves the visual effectiveness of these campaigns.

2.2. The influence of age and personality traits on advertising perception

Age is related to both consumer decision-making and personality development across adulthood. Studies of the Big Five show systematic age-related mean-level differences and changes in trait stability (Soto et al., 2011; Cobb-Clark & Schurer, 2012; Wagner, Lüdtke and Robitzsch, 2019; Shchebetenko et al., 2020). Advertising research likewise indicates that responses to persuasive appeals, advertising memory, and broader indicators of advertising effectiveness can differ across age groups (van der Goot, Bol and van Weert, 2021; Mecredy, Stocchi and Feetham, 2025; Chen et al., 2025).

Importantly, age effects are not uniform. Recent studies show that they depend on the type of appeal, the product or service category, personal relevance, and the decision context (van der Goot, Bol, and van Weert, 2021; Chen et al., 2025). This context dependence provides a rationale for testing age across multiple advertised services rather than assuming a single general age effect.

Personality is also relevant to the reception of advertising. Trait-congruent messages have been shown to produce higher engagement than non-congruent messages, particularly for Extraversion and Openness (Hirsh, Kang, and Bodenhausen, 2012; Matz et al., 2017). Sačer (2023) extended this approach in Croatia by identifying Facebook Likes associated with Big Five dimensions, while Sačer (2022) validated visual materials linked to these traits. Combining age- and personality-based segmentation therefore provides a theoretically plausible test of whether psychological targeting effects differ across adulthood.

2.3. Effectiveness of online advertising

Click-through rate (CTR) is one of the most commonly used measures of online advertising success. It is defined as the ratio of advertisement clicks to impressions (Ružić, Biloš, and Turkalj, 2014). The average CTR for Facebook campaigns is 1,11%, but it varies by sector (from 0,47% for employment to 1,61% for legal services) (Chaffey, 2024).

Global spending on social media advertising is expected to reach 480 billion USD by 2030 (Davies, 2026). A vast amount of users' data in the digital space enables marketers to segment messages with great precision at low cost, enabling almost personalised communication. According to Bivens and Haimson (2016), cookies can be used by marketers to deepen and broaden data collection, helping them track users' online behaviour and enabling more targeted advertising.

2.4. Facebook Likes as one of the most powerful digital footprints of users

Every search history, social interaction, purchase, and other Internet activity is logged and can be transformed into Big Data (i.e., large datasets). This amount of data (thousands or even millions of users) and information can help scientists observe phenomena globally, not just in small datasets. In this way, they can reduce or even overcome the "noise" problem in the samples (Kosinski et al., 2016; Stachl et al., 2020). Kosinski, Stillwell and Graepel (2013) made a breakthrough in research using Big Data - specifically Facebook Likes. Only 30 Facebook Likes are enough to predict age with 60% accuracy, gender with 75% accuracy and Openness to new experiences with 30% accuracy. These findings raise important privacy concerns because it is proven that the average user during their Facebook use selects about 227 Likes, and that algorithm can judge a person's personality traits more accurately (with 280 Likes) than a spouse or than a friend (using just 65 Likes) (Youyou, Kosinski and Stillwell, 2015). This is not an isolated case; in her research, Sačer (2023) confirmed the results of Kosinski, Stillwell, and Graepel (2013); Youyou, Kosinski, and Stillwell (2015); and Kosinski (2014) for Croatian users. This research used the 50-item International Personality Item Pool (IPIP-50). It is a widely used instrument for measuring personality traits (Matz et al., 2017; Sačer, 2022; Matz et al., 2019; Youyou, Kosinski and Stillwell, 2015; Kosinski, 2014). Users' IPIP-50 scores were matched to their Facebook Likes, resulting in a User-Like matrix specific to Croatia. The same matrix was used in this research to segment Croatian Facebook users into groups based on their OCEAN scores.

3 Research design and methodology

Previous studies have demonstrated that trait-congruent appeals can improve advertising responses, but little is known about whether this effect varies across age groups. The present research therefore used a two-stage field quasi-experimental design. First, congruent and non-congruent creatives were compared within each campaign and trait segment. Second, age-related differences were examined for the trait-campaign combinations retained from the first stage.

H1: The effectiveness of personality-congruent advertising differs across age groups.

Facebook was selected because its advertising platform permits real-world targeting by age, location, and interests, making it suitable for a field test of psychological targeting comparable to earlier digital-persuasion research (Matz et al., 2017). Campaigns were implemented in Meta Ads Manager. The design was quasi-experimental because exposure occurred in the natural platform environment and ad delivery was optimised by the platform rather than assigned under full researcher control.

Psychographic audience segments were defined using the previously validated Croatian User-Like matrix (Sačer, 2023), which links Facebook Likes and interests to the Big Five traits. Only Likes in the highest 10% of trait associations that were available in Meta Ads Manager were used. Segment membership therefore indicated an inferred affinity with a trait; it was not a direct personality score obtained from a self-report inventory. Because Meta combines multiple selected interests with OR logic, an individual could enter a segment through one qualifying interest, and audience overlap across trait segments was possible.

All five Big Five dimensions were considered in the initial congruence analysis. The age analysis focused on the trait-campaign combinations that showed a congruence effect in that screening stage: Conscientiousness, Extraversion, and Neuroticism for Shopaj24, and Openness, Extraversion, and Neuroticism for the dental clinic. Agreeableness was not retained for either campaign. Because this selection was data-dependent, the age analyses should be interpreted as exploratory.

Two service contexts were used to assess whether age-related patterns were consistent across contrasting field settings. Shopaj24 was a nationwide digital classified advertising service, whereas the dental clinic was a geographically restricted health service in Zagreb. The comparison was not intended to represent the retail and health sectors as a whole; rather, it served as an exploratory cross-context test. Visual materials were drawn from a previously validated Croatian image set, in which each image was empirically associated with a Big Five trait (Sačer, 2022).

Each campaign ran for seven days. Within each retained trait segment, users were exposed to either a trait-congruent image or a non-congruent control image. Written copy was identical across services, and each ad set received the same nominal budget of EUR 15. Meta's delivery algorithm distributed impressions within those constraints. Thus, the manipulated element was visual trait congruence, while copy and nominal budget were held constant.

In total, the campaigns reached 102,377 Facebook users in Croatia (Shopaj24: 64,951; dental clinic: 37,426) and generated 156,127 impressions (Shopaj24: 102,764; dental clinic: 53,363).

The target population comprised adults aged 18 years or older residing in Croatia; the dental-clinic campaign was restricted to residents of Zagreb. For each retained trait segment, the campaign included a congruent and a non-congruent creative condition. No self-report personality inventory was administered, and the audience definitions were based exclusively on Facebook Likes included in the Croatian User-Like matrix.

Figure 1 shows an example of a trait-congruent creative for users with an inferred affinity for Openness. Meta Ads Manager did not make every Like in the Croatian matrix available for targeting; therefore, only eligible Likes were used. For example, 42 of the 165 Likes most strongly associated with Neuroticism were available in the platform.

In contrast to Matz et al. (2017), only the image was tailored to the inferred personality trait; the written copy remained unchanged within each service:

- Shopaj24: Shopping? The best offers are waiting for you on the new Shopaj24 platform. Discover them and see for yourself.
- Dental clinic: [FREE EXAMINATION] Perfect your smile. Schedule a free initial examination at a dental clinic with over 30 years of experience.

Meta Ads Manager combined selected Likes with the OR operator rather than requiring users to match all interests. Consequently, inclusion in a segment could be based on a single qualifying Like, and trait segments were not mutually exclusive. Table 1 reports the estimated audience sizes and the number of Likes available for each segment.

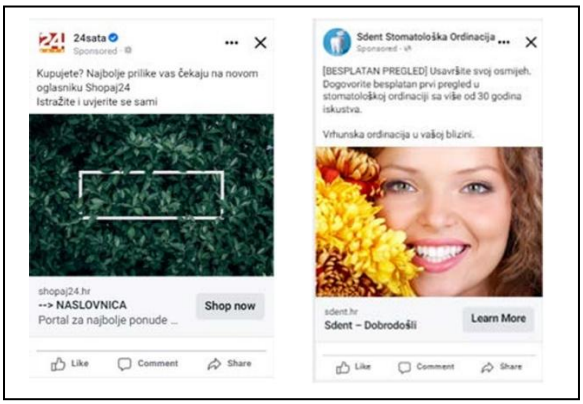


Figure 1. Example of a personality-trait-congruent advertisement for users with an inferred affinity for Openness

The methodological approach was chosen to extend prior field research on psychological targeting (Matz et al., 2017) while adapting the segmentation and visual stimuli to the Croatian context (Sačer, 2022; Sačer, 2023). Using two distinct service contexts provided an exploratory assessment of cross-context consistency. However, the absence of direct personality measurement and full random assignment limits claims to causal and construct validity.

Advertised service	Audience name	Estimated audience size	Num. of Likes
Shopaj24	Conscientiousness HR audience	1.900.000-2.200.000	46
	Extraversion HR audience	1.900.000-2.200.000	40
	Neuroticism HR audience	1.900.000-2.200.000	42
Dental clinic	Openness ZG audience	667.400-785.100	53
	Extraversion ZG audience	692.700-814.900	40
	Neuroticism ZG audience	686.100-807.200	42

Table 1. Audience segmentation based on Facebook Likes associated with Big Five traits

Note. Audience sizes are estimates generated by Meta Ads Manager from the specified target parameters.

Click-through rate (CTR; clicks divided by impressions multiplied by 100) was used as the descriptive indicator of campaign response. Inferential analyses were based on click and non-click frequencies rather than on CTR values themselves. For each retained trait-campaign combination, Pearson chi-square tests were used to compare response frequencies across the three age groups. Separate 2 x 2 chi-square tests then compared congruent and non-congruent creatives within each age group. Binary logistic regression was used to estimate differences in click probability across age groups, with ages 18-34 as the reference category. Models were fitted separately for each retained trait-campaign combination, and statistical significance was evaluated at $p < .05$.

Meta Ads Manager reports several narrower age bands (18-24, 25-34, 35-44, 45-54, 55-64, and 65+). For the present analyses, these categories were combined into three broader groups: younger adults (18-34), middle-aged adults (35-54), and older adults (55+). The grouping supported interpretable cross-campaign comparisons and reduced fragmentation of the response-frequency tables.

Age-related analyses were conducted separately for each retained campaign-trait combination. Because retention depended on the first-stage congruence results and several subgroup tests were conducted, the findings are considered exploratory.

4 Research results

CTR values are reported descriptively, whereas the chi-square tests were calculated from click and non-click frequencies. The analyses therefore assessed whether response frequencies differed across age groups and whether congruent creatives outperformed non-congruent creatives within each age group.

For the dental-clinic campaign, age was associated with response in the Neuroticism segment, $\chi^2(2) = 8.483$, $p < .05$, whereas the corresponding age-group tests were not significant for Extraversion or Openness. Among users exposed to Neuroticism-congruent advertising, CTR was 5.5% for ages 18-34, 3.7% for ages 35-54, and 3.0% for ages 55+.

Within each dental-clinic age group, Neuroticism-congruent and non-congruent creatives also differed significantly: ages 18-34, $\chi^2(1) = 9.412$, $p < .05$; ages 35-54, $\chi^2(1) = 15.698$, $p < .05$; and ages 55+, $\chi^2(1) = 10.673$, $p < .05$.

For Shopaj24, age-group differences were significant for Extraversion, $\chi^2(2) = 24.552$, $p < .05$, and Neuroticism, $\chi^2(2) = 55.154$, $p < .05$, but not for Conscientiousness, $\chi^2 = 3.009$, $p > .05$. Among users exposed to congruent creatives, CTR increased across age groups for Extraversion (3.0%, 3.5%, and 5.4%) and Neuroticism (4.5%, 5.2%, and 7.4%).

For Extraversion, congruent and non-congruent Shopaj24 creatives differed within ages 35-54, $\chi^2(1) = 14.674$, $p < .05$, and ages 55+, $\chi^2(1) = 24.326$, $p < .05$, but not within ages 18-34. For Neuroticism, the difference was significant within all three age groups: ages 18-34, $\chi^2(1) = 29.074$; ages 35-54, $\chi^2(1) = 73.425$; and ages 55+, $\chi^2(1) = 169.077$, all $p < .05$.

The logistic regression models yielded the same directional pattern. In the dental-clinic Neuroticism segment, users aged 35-54 (AME = −0.0113, SE = 0.0059, 95% CI [−0.0229, 0.0014], p = 0.053) and 55+ (AME = −0.0123, SE = 0.0055, 95% CI [−0.0231, −0.0015], p = 0.026) showed lower click probability than users aged 18-34. In Shopaj24, users aged 55+ showed higher click probability than users aged 18-34 in the Extraversion (AME = 0.0119, SE = 0.0030, 95% CI [0.0060, 0.0178], p = 0.000) and Neuroticism segments (AME = 0.0111, SE = 0.0027, 95% CI [0.0058, 0.0164], p = 0.000).

Overall, H1 received partial support. Age was associated with response to personality-congruent advertising in several, but not all, trait-campaign combinations, and the direction of the association differed between the two service contexts.

Across all delivered impressions, CTR was 2.25% for the dental clinic (2.78% for congruent and 1.72% for non-congruent advertisements) and 3.38% for Shopaj24 (3.92% for congruent and 2.85% for non-congruent advertisements). These descriptive values

were above the cited industry benchmarks of 0.83% for healthcare and 1.59% for retail (Irvine, 2025); however, benchmark comparisons are descriptive and do not establish the causal effect of tailoring.

5 Conclusions and Recommendations

The findings indicate that personality-congruent visual advertisements can generate higher engagement than non-congruent advertisements, but the age pattern was conditional rather than uniform. In the dental-clinic campaign, the strongest age-related pattern was for Neuroticism, which favoured younger adults. In the Shopaj24 campaign, Extraversion- and Neuroticism-congruent advertisements produced higher response rates among older adults. Age should therefore be interpreted as a contextual refinement of personality-based targeting, not as a universally acting moderator.

This study makes three scientific contributions. First, it extends research on psychological targeting by examining age differences in a field setting rather than treating age solely as a control variable. Second, it applies the same trait-congruence logic in two distinct service contexts and uses observed click behaviour rather than stated intentions. Third, it provides evidence from Croatia using a locally validated User-Like matrix. Together, the findings show that age-related differences in personality-tailored advertising are conditional on both the trait cue and the advertised service.

From a practical perspective, age can serve as an additional layer in psychographic targeting. For health-related services, assurance, efficacy, and control cues may be relevant when targeting younger users with an inferred affinity for Neuroticism. For digital classified advertising services, socially and emotionally reassuring cues may be useful when targeting older users with inferred affinities for Extraversion or Neuroticism. Because the effects were context-specific, these implications require replication before being generalised to other campaigns or sectors.

Several limitations qualify the findings. Personality was inferred from Facebook Likes and was not measured directly. Meta's OR targeting logic allowed entry into the segment via a single qualifying interest and permitted overlap across trait segments. Ad delivery was optimised by Meta, so exposure was not fully randomised, and users could receive unequal numbers of impressions. Only visual elements were tailored, campaigns lasted seven days, and outcomes were limited to clicks rather than conversions or retention.

The two services were contrasting field contexts, not representative samples of the retail and health sectors, and the findings may not generalise beyond Croatian Facebook users or to other platforms. In addition, trait-campaign combinations were selected based on first-stage congruence results, and numerous subgroup tests were conducted. The age analyses are therefore exploratory and may be affected by selection and multiplicity.

Future studies should preregister the primary contrasts, test age x congruence interactions directly, report regression coefficients and odds ratios with 95% confidence intervals, use direct personality measures where feasible, and include additional service categories and post-click outcomes. Such replications would clarify when age meaningfully enhances the effectiveness of personality-tailored digital advertising.

References

Azucar, D., Marengo, D., & Settanni, M. (2018). Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis. *Personality and Individual Differences*, 124, 150-159.

Bašić, M., and Đurđević, D. (2017). Utjecaj komparativnog oglašavanja na stratešku adaptivnost marki stranih proizvođača u Republici Hrvatskoj. *Zbornik Ekonomskog fakulteta u Zagrebu*, 15(1), 75-92

Bivens, R., & Haimson, O.L. (2016). Baking Gender Into Social Media Design: How Platforms Shape Categories for Users and Advertisers. *Social Media + Society*, 2(4), 1-12.

Bratić, D., Palić, M., & Miljković, P. (2021). Sentiment Analysis of the Infographics on Social Media. In Načinović Braje, I., Jaković, B., Ferjanić Hodak, D. (Eds.), *Proceedings of FEB Zagreb International Odyssey Conference on Economics and Business* (1073-1082). Zagreb: Ekonomski fakultet Sveučilišta u Zagrebu.

Carpenter, S. M., & Yoon, C. (2015). Aging and consumer decision making. In T. M. Hess, J. Strough, & C. E. Löckenhoff (Eds.), Aging and decision making: Empirical and applied perspectives (pp. 351–370). Academic Press. https://doi.org/10.1016/B978-0-12-417148-0.00017-0

Carpenter, S. M., & Yoon, C. (2011). Aging and consumer decision making. *Annals of the New York Academy of Sciences, 1235*, E1–E12.

Chen, W., Zhang, L., Yang, S., Shen, Z., Xie, F., & Benson, V. (2025). Age differences in preference and memory for advertisements: The roles of advertisement type and product type. Humanities and Social Sciences Communications, 12, Article 770. doi: 10.1057/s41599-025-05168-1

Chaffey, D. (2024). 2025 update: average ad click through rates (CTRs) for paid search, display and social media. Retrieved [2025-09-11] from

https://www.smartinsights.com/internet-advertising/internet-advertising-analytics/display-advertising-clickthrough-rates/

Cobb-Clark, D.A., i Schurer, S. (2012). The stability of the big-five personality traits. *Economics Letters*, 115(1), 11-15. doi: 10.1016/j.econlet.2011.11.015

Davies, E. (2026). Social Media Advertising Insights 2026: 35 Stats, Plus 2M Live Ad Pixels. Retrieved from https://technologychecker.io/blog/social-media-advertising-marketing-insights

Drolet, A., Jiang, L., Pour Mohammad, A., & Davis, C. (2019). The influence of aging on consumer decision-making. Consumer Psychology Review, 2(1), 3-16. doi: 10.1002/arcp.1046

Goldsmith, R. (2016). The Big Five, happiness, and shopping. *Journal of Retailing and Consumer Services*, 31, 52-61. doi: 10.1016/j.jretconser.2016.03.007

Hirsh, J.B., Kang, S.K., & Bodenhausen, G.V. (2012). Personalized persuasion tailoring: Persuasive appeals to recipients' personality traits. *Psychological Science*, 23, 578-581.

Irvine, M. (2025). Facebook Ad Benchmarks for YOUR Industry [Data]. Retrieved [2025-09-11] from https://www.wordstream.com/blog7ws/2017/02/28/facebook-advertising-benchmarks

Jurković, M. (2021). *Modeli oglašavanja na internetu* (Master Thesis). University North, Koprivnica.

Kosinski, M. (2014). *Measurement and prediction of individual and group differences in the digital environment* (Doctoral Thesis). Downing College, Cambridge, University of Cambridge.

Kosinski, M., Stillwell, D.J., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behaviour. *PNAS*, 110(15), 5802-5805. doi: 10.1073/pnas.1218772110

Kosinski, M., Wang, Y., Lakkaraju, H., & Leskovec, J. (2016). Mining Big Data to Extract Patterns and Predict Real-Life Outcomes. *Psychological Methods*, 21(4), 493-506.

Kozyreva, A., Lorenz-Spreen, P., Hertwig, R., Lewandowsky, S. & Herzog, S.M. (2021). Public attitudes towards algorithmic personalization and use of personal data online: Evidence from Germany, Great Britain, and the US. *PsyArXiv Preprints*. Retrieved from https://psyarxiv.com/3q4mg [2021-01-19]. doi:10.1057/s41599-021-00787-w

Liang, B., & Wang, Y. (2021). Using integrated marketing communications to promote country personality via government websites. *Place Branding and Public Diplomacy*. doi: 10.1057/s41254-021-00221-7

Matz, S.C., Kosinski, M., Nave, G., & Stillwell, D.J. (2017). Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the National Academy of Sciences*, 114(48), 12714-12719. doi:10.1073/pnas.1710966114

Matz, S.C., Segalin, C., Stillwell, D., Müller, S.R., & Bos, M.W. (2019). Predicting the Personal Appeal of Marketing Images Using Computational Methods. *Journal of Consumer Psychology*, 29(3), 370-390.

Mecredy, P., Stocchi, L., & Feetham, P. (2025). Differences in advertising's effectiveness across age groups. International Journal of Advertising, 44(2), 235-262. doi: 10.1080/02650487.2024.2370678

Miralam, M.S., Ali, N., & Jeet, V. (2020). Impact of Big Five Model on Leadership Initiation in Critical Business Environment Among Marketing Executives. *The Journal of Asian Finance, Economics and Business*, 7(11), 507-517. doi: 10.13106/jafeb.2020.vol7.no11.507

Ružić, D., Biloš, A., & Turkalj, D. (2014). *E-marketing* (3rd Edition). Osijek: Faculty of Economics Osijek, University Josip Juraj Strossmayer in Osijek.

Sačer, S. (2023). *Učinkovitost psihografskoga segmentiranja korisnika društvenih mreža pomoću njihovih digitalnih otisaka* (Doctoral Thesis). Faculty of Economics and Business, Zagreb, University of Zagreb.

Sačer, S. (2022). Povezanost vizualnih obilježja slika i osobina ličnosti korisnika pri promociji na društvenim mrežama. *Zbornik Ekonomskog fakulteta u Zagrebu*, 20(2), 47-60.

Shchebetenko, S., Kalugin, A.Y., Mishkevich, A.M., Soto, C.J., & John, O.P. (2020). Measurement Invariance and Sex and Age Differences of the Big Five Inventory-2: Evidence From the Russian Version. *Assessment*, 27(3), 472-486.

Specht, J., Egloff, B., & Schmukle, S.C.(2011). Stability and change of personality across the life course: The impact of age and major life events on mean-level and rank-order stability of the Big Five. *Journal of Personality and Social Psychology*, 101(4), 862–882. doi: 10.1037/a0024950

Soto, C. J., John, O. P., Gosling, S. D., & Potter, J. (2011). Age differences in personality traits from 10 to 65: Big Five domains and facets in a large cross-sectional sample. *Journal of Personality and Social Psychology, 100*(2), 330–348.

Stachl, C., Au, Q., Schoedel, R., Gosling, S.D., Harari, G.M., Buschek, D., …, and Bühner, M. (2020). Predicting personality from patterns of behavior

collected with smartphones. *Proceedings of the National Academy of Sciences*, 117(30), 17680-17687. doi: 10.1073/pnas.1920484117

Statista, (2025). Social Media Advertising. Retrieved [2025-09-07] from https://www.statista.com/outlook/amo/advertising/social-media-advertising/worldwide

van der Goot, M. J., Bol, N., & van Weert, J. C. M. (2021). Age differences in preferences for emotionally-meaningful versus knowledge-related appeals. Communications, 46(2), 205-228. doi: 10.1515/commun-2019-0108

Wagner, J., Lüdtke, O., i Robitzsch, A. (2019). Does personality become more stable with age? Disentangling state and trait effects for the big five across the life span using local structural equation modeling. *Journal of Personality and Social Psychology*, 116(4), 666-680.

Williams, P., & Drolet, A. (2005). Age-related differences in responses to emotional advertisements. *Journal of Consumer Research, 32*(3), 343–354.

Yoon, C. (2020). *The Aging Consumer: Perspectives from Psychology and Marketing* (2. izd.). Engleska: Routledge/Taylor Francis.

Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *PNAS*, 112(4), 1036-1040.

Circular Migration without Return Infrastructure: Policy and Social-Partner Sensemaking of Labour-Market Governance in a Developing Context

Violeta Cvetkoska, Igor Ivanovski, Filip Peovski, Bojan Kitanovikj, Ivona Serafimovska
Ss. Cyril and Methodius University in Skopje
Faculty of Economics - Skopje
bul. Goce Delcev 9V, 1000 Skopje, North Macedonia
{violeta.cvetkoska, igor.ivanovski, peovski, bojan.kitanovikj, ivona.serafimovska}@eccf.ukim.edu.mk

***Abstract**. The study examines how Macedonian policymakers, social partners, and labour-market intermediaries understand links among emigration, skill shortages, and circular migration. Reflexive thematic analysis of expert interviews, supplemented by a review of national policy documents, shows that migration is framed as a governance challenge. Although circular migration may enable knowledge transfer, remittances, diaspora networks, and return entrepreneurship, participants question its feasibility under institutional conditions. This circularity paradox reflects a return infrastructure deficit encompassing data, coordination, returnee support, skill recognition, employer matching, and trust. This context-bound sensitising concept clarifies when mobility may become developmental in migrant-origin and EU-candidate economies.*

Keywords. circular migration; skill shortages; labour-market governance

1 Introduction

Migration is often framed as a loss for small and peripheral labour markets. Yet the key issue is not only that people leave, but whether institutions can transform mobility into return, circulation and development. This question is especially important for North Macedonia, where outward migration, skill shortages and labour-market mismatch increasingly intersect with the country's EU-oriented reform agenda. Thus, the challenge gains a new dimension despite its demographic nature, looking into how the labour-market system can create credible pathways for return, reintegration and skill circulation.

Migration-development research has long debated whether emigration produces brain drain or whether it can generate remittances, knowledge transfer and brain circulation (De Haas, 2010; Batista et al., 2025). Return migration can contribute to development when migrants accumulate human capital abroad and are able to use it productively after returning (Dustmann et al., 2011). However, this positive scenario depends on more than individual willingness to return, thus requiring institutions able to recognise foreign-acquired skills, connect returnees with suitable opportunities, involve employers and social partners, and build trust in domestic labour-market prospects.

Terminologically, we use circular migration for repeated cross-border movement, return migration for relocation to the country of origin, skilled return for return with transferable human capital, retention for remaining in the domestic labour market, reintegration for post-return incorporation into work and institutions, and repatriation only for state-facilitated return.

This institutional dimension remains insufficiently explored in the context of EU-candidate and Western Balkan labour markets. Existing studies have examined emigration, youth migration aspirations, remittances, skills mismatch and brain drain in the region (King & Oruc, 2020; Peovski et al., 2026). However, less is known about how policymakers, social partners and labour-market intermediaries understand the feasibility of circular migration in a context marked by low wages, fragmented coordination, weak institutional trust and limited evidence on returnees.

Considering this, our research addresses that gap through a qualitative study of policymakers, social partners and labour-market intermediaries in North Macedonia. Thus, the study posits the following research questions: (RQ1) How do policymakers, social partners and labour-market intermediaries in North Macedonia frame the relationship among emigration, skill shortages and labour-market change? (RQ2) Which institutional conditions do policymakers, social partners and labour-market intermediaries in North Macedonia perceive as enabling or constraining return, reintegration and circular migration?

At the core of the study is a provisional interpretation of what we describe as a circularity paradox. Stakeholders recognise the potential value of

migration through foreign work experience, remittances, diaspora networks and knowledge transfer, while simultaneously questioning whether circular migration is realistic under current institutional and labour-market conditions. Their accounts point to both shortages of workers as well as perceived gaps in data, coordination, trust, qualification recognition and return-oriented policy design. We use the term return infrastructure deficit to organise these interconnected gaps and to describe the distance between the aspiration for circular migration and the institutional arrangements available to support return, reintegration and skill circulation.

This formulation is offered as a context-bound sensitising concept, whose contribution is to direct analytical attention from brain drain as human-capital loss toward the arrangements through which foreign-acquired skills, resources and networks might be recognised and absorbed after return. Its potential transferability is most plausible in small migrant-origin economies experiencing outward mobility, skill shortages, EU-linked skill portability and fragmented institutional coordination.

2 Theoretical Background

We build our study on three strands of migration and labour-market theory. The first concerns the conditions under which return migration becomes sustainable. Return is not simply the opposite of emigration, nor does it automatically follow from the accumulation of savings or skills abroad. Namely, earlier literature distinguishes between migrants who leave permanently to maximise lifetime earnings and those who migrate temporarily as part of household strategies to overcome constraints in the origin country (Constant & Massey, 2002). More recent perspectives add that return depends on both individual intentions and preparedness, or, in other words, the extent to which migrants and origin-country institutions are ready for reintegration (Engbersen & Snel, 2025). Thus, the modern literature views the migrants' return as a more structured social and institutional process.

The second literature strand concerns "what" migrants bring back. The value of migration is often assessed through remittances and formal human capital, but migrants may also transfer ideas, expectations, norms, work practices and professional standards. We apply the concept of social remittances here because it shows that migration can reshape origin societies through the circulation of knowledge and behavioural repertoires, not only through money (Levitt & Lamba-Nieves, 2011). However, these transfers are not self-executing as skills and ideas become developmental when local labour markets, employers and institutions are able to recognise, absorb and use them.

Additionally, the study considers the literature on mobility regimes and labour-market governance. Mobility is not equally available to all workers, nor is it governed only by states. Instead, it is shaped by employers, employment agencies, social partners, international organisations, qualification systems and intermediaries. In this sense, mobility-regime theory highlights that movement is enabled, constrained and valued differently across groups and contexts (Schiller & Salazar, 2016). Similarly, research on the governance of return shows that return policies are embedded in broader institutional and political arrangements and do not function as isolated policy measures (Koch, 2017). We argue that in the context of labour markets in Central and Eastern Europe, social partners and intermediaries are particularly relevant because temporary and cross-border labour mobility can reconfigure their role in worker protection, skills matching and policy coordination (Danaj & Meszmann, 2024).

Xiang and Lindquist (2014) define migration infrastructure as systematically interlinked technologies, institutions and actors that facilitate and condition mobility. Adapting their commercial, regulatory, technological, humanitarian and social dimensions, we define return infrastructure as a return-facing configuration of employer matching, recognition rules, data systems, reintegration support and diaspora networks. Therefore, a return infrastructure deficit is relational, not reflecting an absence of all policy, but weak linkage among functions needed to foster post-return labour-market incorporation.

3 Methodology

We adopted an abductive qualitative research design based on nine semi-structured expert interviews with Macedonian institutional actors. The unit of interest was institutional sensemaking rather than the prevalence of individual attitudes or migrant behaviour. A role-based purposive sampling strategy was therefore used to recruit participants with direct responsibility for, or operational knowledge of, employment policy, skills development, employer representation, labour-market matching, qualification recognition or migration governance (Palinkas et al., 2015). Eligibility was based on the participant's institutional position and access to relevant policy or implementation knowledge rather than on organisational tenure alone.

The sample comprised four policymaking or policy-facing actors (PM1-PM4) and five social partners or labour-market intermediaries (SP1-SP5). Together, they represented government ministries and bodies, public agencies, a chamber or employer-facing institution, an employers' organisation, a regional migration initiative, a development NGO and a university-based professional training and career centre. Participants mainly occupied senior or middle-level positions and had up to 20 years of organisational

experience. This composition (Table 1 in the Appendix) was intended to capture perspectives across policy formulation, implementation, employer demand, skills provision and labour-market intermediation.

Potential participants were approached by email between May and October 2025. Interviews, with an average duration of 60 minutes, were conducted in person in the Macedonian language. A common interview guide covered labour-market changes, emigration pressures, skill shortages, employment measures, the roles of social partners and intermediaries, qualification recognition, and mechanisms for matching labour supply with demand. The semi-structured format ensured coverage of these common topics while allowing follow-up questions. Participants received information about the study and provided written informed consent. With their permission, interviews were audio-recorded and transcribed verbatim, and all materials were anonymised. All interviews were conducted by three members of the research team, who afterward translated the transcripts into English. Quoted passages were checked against the Macedonian originals by a separate research team member. Ethical approval was granted by the host university (No. 05-3157/1, 21 May 2025).

The adequacy of the nine-interview sample was assessed in relation to the study's bounded analytical purpose and the information power of the cases (Malterud et al., 2016). The study addresses a focused set of research questions, so the participants were selected for highly specific institutional knowledge, and the interviews covered complementary positions across policy design and labour-market implementation. These characteristics enabled an intensive comparison of recurring themes and points of disagreement. Consistent with reflexive thematic analysis (RTA), we do not treat saturation as a fixed or fully objective endpoint because themes are interpretively constructed rather than simply discovered in the data (Braun & Clarke, 2021). Instead, we regard the material as thematically sufficient for developing a provisional account of institutional sensemaking in this bounded setting. As such, the sample is not intended to exhaust all institutional perspectives, establish the prevalence of particular views or represent the experiences of migrants and returnees.

The interviews were analysed using Braun and Clarke's (2019) RTA, following a recursive analysis approach. First, the research team read the complete interview material repeatedly and prepared brief case summaries and analytical memos. Initial coding remained close to respondents' accounts and captured statements concerning emigration, labour shortages, policy measures, return, qualification recognition, institutional coordination and trust. In a second cycle, related codes were compared across cases and stakeholder groups and organised into candidate themes. Candidate themes were then reviewed against the full dataset for internal coherence, distinctiveness, relevance to the research questions, and ability to account for divergent or qualifying accounts. Analysis first proceeded inductively from participants' accounts. Emerging themes were then iteratively compared with return-migration and migration-infrastructure theory, as the mismatch between desired circularity and perceived implementation gaps prompted the interpretation of a return infrastructure deficit. We returned to the full transcripts, including divergent cases, to test and refine this explanation. This movement from observations to a theory-informed explanation constitutes the abductive element of the methodology. Eventually, the concept of return infrastructure deficit was developed during this later interpretive stage to connect several empirically identified gaps.

Coding was undertaken manually, as the primary analyst conducted the initial coding, while the co-authors reviewed the thematic structure and challenged alternative interpretations. Analytical discussions were used to refine the interpretation rather than to calculate inter-coder agreement, which would be inconsistent with the reflexive orientation of the analysis. Credibility was further supported through memo-writing, comparison across stakeholder groups, attention to divergent accounts, tracing each theme to the underlying interview material, and the use of illustrative quotations in the findings.

To distinguish perceptions from documented arrangements, we conducted a focused contextual review of the Migration Policy Resolution 2021-2025, National Employment Strategy 2021-2027 and National Youth Strategy 2023-2027 (Government of the Republic of North Macedonia, 2021; Ministry of Labour and Social Policy, 2021; Agency for Youth and Sport, 2022). The documents were used to verify formal provisions and stated data gaps, not implementation effectiveness, and were not coded as a second dataset.

4 Findings and Results

Based on the data analysis, we create several themes of policy and social-partner sensemaking related to skill shortages, brain drain and labour-market governance. They report participants' perceptions, whereas documentary observations are identified separately, and return infrastructure deficit denotes our theoretical interpretation.

1. From unemployment to scarcity: the labour market now lacks both workers and job-ready skills. A dominant theme across interviews is that North Macedonia's labour-market problem is no longer reflecting issues of unemployment. On the contrary, stakeholders describe a change toward labour scarcity, especially in sectors that require practical, technical or occupation-specific skills. The shortage is framed as

both quantitative in terms of not having enough people, and qualitative, or not enough workers with immediately usable competences.

"The demand for qualified labour is increasing, while the supply is decreasing." (PM2)

"[There is a shortage in the] number of people and …, skills of people who need to be in the labour market. A very large number of people who are needed in the local labour market are no longer here." (SP1)

"There is a lack of employees in certain sectors where there are gaps in terms of labour supply and demand… construction, agriculture, textile industry, tourism and hospitality… and the ICT sector." (PM4)

From this, we can deduce that these actors in the ecosystem perceive emigration as a direct pressure on labour shortages with potential implications for domestic production and service delivery.

2. Emigration is increasingly framed as permanent exit. Stakeholders repeatedly distinguish between temporary labour migration and what they see as a more serious pattern of permanent emigration. The language used by respondents is strong: "exodus", "epidemic" and "entire families leaving". This suggests that the circular-migration model is weakened by the perception that many migrants no longer plan to return.

"If once workers left the country with a low level of qualifications and with the possibility of ever returning, today entire families are leaving, especially young highly educated personnel." (PM3)

"What has been happening in recent years is a literal exodus, not just brain drain migration, but migration in a broader sense… we are facing permanent eviction from the country." (SP3)

"An 'epidemic' of migration is happening. Everyone goes abroad, regardless of educational profile. Even people and families who are financially well-established move out permanently." (SP5)

Subsequently, respondents also interpreted emigration as reflecting reduced confidence in domestic prospects, not wages alone.

3. The circularity paradox of views where circular migration is desirable, but weakly evidenced and institutionally uncertain. Respondents recognise that migration could produce benefits if migrants return with savings, knowledge, networks and entrepreneurial ideas. However, this positive interpretation is consistently conditional. Respondents presented circular migration as desirable as desirable but not yet supported by strong evidence or institutional mechanisms.

"People who go out don't all stay there. Most of them come back here and come back with ideas… They worked, saw how well it worked, made a good network there and came back here." (SP1)

"Migration is good if young people migrate for a certain period, get further education and acquire new knowledge and skills… and when they return, they contribute to the development of the national economy." (SP2)

"If young people would return to their country and take with them the experience and knowledge gained in other countries, they would certainly have benefits, but we have no data on such a trend." (PM1)

"It is still early for such processes." (SP5)

In this sense, this theme creates the central paradox of the study. On one hand, stakeholders value circulation in principle, while, on the other, they do not yet see a credible system that makes return and skill circulation likely. In other words, respondents recognise the potential benefits of migration in the form of remittances, foreign experience, new standards, networks, business ideas and possible diaspora investment. However, they also repeatedly state that circular migration is weak, rare or "too early."

The tension between accounts of increasingly permanent exit (Theme 2) and claims that many migrants return (Theme 3) is itself informative. Hence, the disagreement indicates that institutional actors lack a shared evidence base for assessing circularity, which is an informational component of the proposed deficit.

4. Active labour-market measures exist, but they are not sufficiently targeted toward return and retention. Interviewees frequently mention active employment measures, training schemes and the Operational Plan. Some respondents evaluate these measures positively, especially for youth employment and employability. However, others argue that the measures are unevenly evaluated, poorly targeted or insufficient for employers who primarily need workers rather than financial incentives.

"I think an evaluation should be done to see the effectiveness of these measures. Which of these measures specifically give greater results, and which ones give less results…" (SP2)

"Not very successful because the trainings are redirected to cities where there is no need for the corresponding occupations, or there are very few people who want to participate in retraining." (PM3)

"We don't need financial benefits, we need workers." (SP5)

"In the previous period, we had weak measures, and therefore it is necessary to work on their promotion and introduction of new measures that will be considered to act on several spheres in society." (PM4)

From this, we can conclude that policy activity should not be confused with policy capacity. Existing measures may support employability, but they do not yet amount to a coherent migration-retention-return strategy.

5. Qualification frameworks are seen as useful, but also incomplete and potentially double-edged. The National Qualifications Framework and European qualification instruments are viewed as important for recognition, transparency and labour-market matching. Yet respondents also identify implementation weaknesses, poor communication with users and the risk that internationally recognised qualifications may make outward mobility easier without improving return.

"The National Qualifications Framework [is] a great thing, but if you ask a young person in an employment office how much the National Qualifications Framework has helped you, they won't be able to say. What we lack in the country is the contact and sharing of information." (SP1)

"There were attempts to create a laboratory for the labour market… but I don't think it took off… [there is no] observatory in our country that would follow the trends and skills needs and would harmonize all this in a framework." (SP2)

"The EQF has helped increase the mobility of skilled workers… However, it also contributes even more to an increase in brain drain, as workers with EU-recognised qualifications often choose to migrate to higher-paid jobs in other EU countries." (SP4)

"It enables and facilitates the comparison and recognition of acquired qualifications… [and] creating programs according to the needs of the labour market." (PM3)

This theme shows the ambivalence of skills policy in migrant-origin countries where recognition systems can support matching and return, but respondents perceived skills portability as potentially facilitating outward mobility.

6. Perceived gaps in return infrastructure. Across the interviews, participants described a fragmented set of policies for facilitating return and reintegration. Their accounts identified weak data, unclear return measures, limited returnee-specific support, low institutional trust and insufficient attention to broader quality-of-life conditions.

"We have no data on whether there are measures for repatriation." (PM1)

"We have no knowledge of strategies and measures for the repatriation of qualified emigrants, only in the Youth Strategy 2023-2027, which is late in its implementation." (PM3)

"We do not have them yet, and we cannot know if they are effective." (SP2)

"I don't even see the measures at all. I have not heard that some people saw a benefit from them, so they returned to the country. Much more needs to be done so that these people feel safe to return home." (SP5)

"The return will depend on the economic policies, but also on the quality of life in the country… These policies are economic, cultural, social, legal and labour to provide opportunities for them." (PM4)

The document review qualifies participants' claims. The Migration Policy Resolution supports circular and return migration and reintegration, while acknowledging sparse data on circular and return flows, while the employment and youth strategies contain broader matching and qualification-recognition measures. Thus, participants' accounts evidence limited awareness and perceived implementation gaps, so subsequently, return infrastructure deficit is our interpretation of weak return-facing linkage across existing instruments.

In this vein, these accounts suggest that weaknesses in data, coordination, qualification recognition, employer matching, returnee-oriented support and institutional trust may combine to make return and skill circulation less credible. We use return infrastructure deficit as an interpretive label for this recurring pattern, meaning the gap between the aspiration for circular migration and the institutional capacity to make return, reintegration and skill circulation credible. At this stage, the concept summarises how the interviewed stakeholders understand these interconnected gaps; it should not be read as an independently measured institutional index or as evidence that the same configuration applies across all migrant-origin economies. Additionally, it is worth noting that stakeholders do not reject circular migration as an idea, but suggest that it cannot function without data, trust, coordination, recognition, employer matching and returnee-oriented policy design.

Moreover, respondents support cooperation with the EU, Erasmus+, IPA, ILO, EU qualification frameworks and bilateral or sectoral partnerships. However, the policymakers strongly warned against copying foreign models without adapting them to the Macedonian economic structure, institutional capacity and long-term development strategy, calling for "Europeanisation as learning, not imitation".

5 Discussion

The interviews suggest that institutional actors increasingly frame migration as a labour-market governance problem characterised by worker scarcity, practical skill shortages and weak retention. This framing is consistent with research on labour mobility and institutional change in Central and Eastern Europe (Danaj & Meszmann, 2024). In our case, however, this mobility is experienced asymmetrically. So, while destination economies may benefit from available labour, the situation in origin-country labour markets points to a reduced capacity to fill vacancies and retain skilled workers.

The findings reinforce the relevance of the brain-drain debate but also shows its limits. Brain drain captures the loss of human capital, yet it does not fully explain why emigration fails to become circulation. Stakeholders recognise that migration could support development through savings, experience, diaspora networks and knowledge transfer. This corresponds to work on social remittances, which shows that migrants may transfer ideas, practices, expectations and professional norms, not only money (Levitt & Lamba-Nieves, 2011).

Recent research also suggests that return can contribute to entrepreneurship, labour-market reintegration and institutional learning, but only when returnees are able to use their accumulated resources productively (Wahba, 2026). In our analysis, participants perceived this condition as weak,

considering that return is imagined as beneficial, but stakeholders repeatedly point to limited data, fragmented coordination, unclear repatriation measures, weak employer matching and low trust in domestic opportunities.

Namely, the study's central contribution lies in showing that North Macedonia, as a developing country in a resource-constrained context, illustrates a circularity paradox. Stakeholders support the idea of circular migration yet still perceive current arrangements as insufficient. Return-migration scholarship increasingly emphasises that return is not a single event, but a process shaped by expectations, reintegration conditions, institutional support and the possibility of re-migration (Meļķe, 2024; Schreier, 2024; Tjaden et al., 2025).

Our study extends this argument to an EU-candidate origin-country labour market by showing that the problem encompasses institutional preparedness as well. Existing employment measures, qualification frameworks and social-partner mechanisms are present, but they are not sufficiently connected into a coherent strategy for return, reintegration and skill circulation.

The interviews also highlight the ambivalent role of skills policy. Qualification frameworks, recognition systems and skills-development measures can improve transparency and labour-market matching. Yet for migrant-origin countries, they may also increase the portability of labour without guaranteeing domestic absorption. If wages, working conditions, career prospects and institutional trust remain weak, better-recognised skills may make outward mobility easier rather than making return more attractive. This supports wider arguments that migration systems are shaped by mobility regimes that distribute opportunities unequally across places, actors and labour-market positions (Schiller & Salazar, 2016).

Also, we highlight the constrained but important role of social partners and labour-market intermediaries. These actors are expected to bridge employers, workers, education providers and the state, yet their bridging capacity is limited by weak data, fragmented coordination and uneven influence over implementation. Danaj and Meszmann (2024) show that labour-market intermediaries can reconfigure the regulatory role of social partners in Central and Eastern Europe. We add that, in an origin-country setting, social partners and intermediaries are also central to whether migration becomes extractive or developmental. Without stronger institutionalised cooperation, respondents associated weak coordination with greater risks of permanent emigration, fragile retention and underused diaspora potential.

In this sense, Figure 1 presents a provisional analytical framework derived from the stakeholder accounts. It organises the relationships that participants perceived among external labour-market pressures, domestic institutional capacity, gaps in support for return, and plausible mobility outcomes. The arrows indicate a proposed sequence for future investigation rather than causal relationships tested in this study. Moreover, the interviews suggest that low wages, demographic decline, EU labour demand, emigration and sectoral skill shortages are interpreted through domestic arrangements involving ministries, employment agencies, social partners, qualification systems and training providers. Where these arrangements are perceived as fragmented, data-poor or weakly connected to credible domestic opportunities, stakeholders anticipate less circulation and greater risks of permanent emigration, fragile retention and underused diaspora potential.

Theoretically, the paper contributes with a bounded concept development by moving beyond brain drain as a human-capital loss and toward the institutional conditions under which mobility may become circulation. Thus, conceptually, the contribution is directional and relational. Whereas migration infrastructure explains how mobility is mediated, return infrastructure examines how mobility-acquired skills and resources are identified, recognised, matched and absorbed after return. The proposed deficit signals weak connections among existing functions, not a generic absence of institutions. As such, it is advanced as a context-bound sensitising concept whose dimensions and generalisability require comparative testing. This becomes especially relevant for small EU-candidate economies, where European alignment can increase skill portability without automatically improving domestic retention or return.

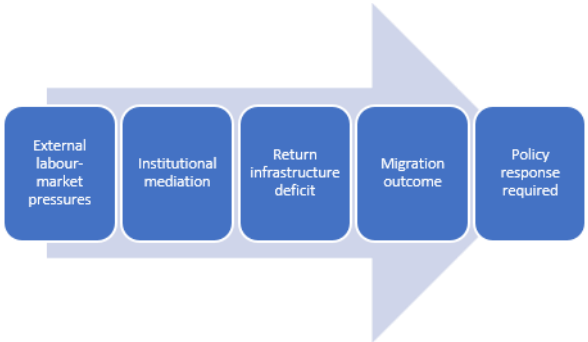


Figure 1. Provisional analytical framework of perceived return-infrastructure gaps

Policy-wise, the findings support a unified governance response instead of typical stand-alone programmes. First, the state statistical bodies and migration authorities together with relevant ministries should create an annual dashboard linking emigration, return and vacancy data. Second, the government employment agency and employer organisations should establish a voluntary, data-protection-compliant returnee skills register and vacancy-matching service. Third, the role of education and qualification-recognition bodies would be to publish processing standards and accelerated pathways for foreign qualifications and prior learning. Fourth, a

permanent interministerial authority with social partners should assign a lead institution, budget, deadline and indicator to each action. Pilots in shortage sectors should track recognition time, job-skill match, employment quality and 12- and 24-month retention. Respondents' accounts suggest that return depends not only on available jobs but also on a predictable and dignified future.

6 Limitations and Future Research

This study has several limitations. First, it is based on nine purposively selected institutional stakeholders from one national setting. Considering that, the findings should be interpreted as a context-bound account of expert sensemaking rather than as a representative description of all Macedonian institutions or as evidence of how frequently particular views are held. Although the sample incorporates perspectives from policy design, employment services, skills development and migration governance, its aim is not to capture the full diversity of all relevant sectors, regions or institutional positions. It also excludes the first-hand perspectives of emigrants, returnees, diaspora organisations and most individual employers, per the proposed research question.

The transferability of the findings consequently rests on analytical generalisation. The proposed interpretation may be informative for settings that share key contextual characteristics with North Macedonia, including a small migrant-origin labour market, strong outward mobility, sectoral skill shortages, EU-linked skill portability and fragmented governance capacity. With that in mind, we treat the return infrastructure deficit as a provisional sensitising concept whose dimensions and boundary conditions require corroboration through additional broader, comparative and multi-actor research.

Future research could extend the concept of return infrastructure deficit in several directions. Comparative studies across Western Balkan or EU-candidate countries would help examine whether similar institutional gaps exist in other migrant-origin economies facing skill shortages and demographic decline. Further research could also include interviews with returnees, skilled emigrants, employers and diaspora networks to explore whether the institutional deficits identified by stakeholders are confirmed by those directly involved in migration and reintegration. Finally, longitudinal studies would be valuable for assessing whether employment measures, qualification recognition systems, social-partner initiatives and returnee-support programmes influence reintegration outcomes and skill circulation over time.

7 Conclusion

The study examined how policymakers, social partners and labour-market intermediaries in North Macedonia frame the relationship among emigration, skill shortages and the possibility of circular migration in a resource-constrained, national context.

Moreover, the findings show that migration is generally perceived as a broader labour-market governance challenge by policymakers and social intermediaries in the labour market. Illustratively, stakeholders describe a context in which unemployment-centred policy concerns are increasingly replaced by worker scarcity, practical skill shortages and difficulties in retaining young and qualified people. At the same time, they recognise that migration could generate development benefits through knowledge, experience, remittances, networks and return entrepreneurship.

The central conclusion we draw from the data points to a circularity paradox in North Macedonia. Circular migration is widely regarded as desirable, yet participants perceived the institutional conditions needed to make it credible as weak. Existing employment measures, qualification frameworks and social-partner mechanisms provide important foundations, but they are not viewed as sufficiently coordinated into a coherent system for return, reintegration and skill circulation. The study provisionally conceptualises this gap as a return infrastructure deficit, encompassing weaknesses in data systems, institutional coordination, returnee-specific support, skill-recognition pathways, employer matching and trust in domestic opportunities.

By developing this context-bound concept, the study contributes to migration and labour-market scholarship by shifting attention to the brain drain phenomenon in a new light, meaning the institutional conditions under which mobility may become developmental. The findings suggest that the key policy challenge is deeper than just reducing emigration but building a labour-market system in which return becomes meaningful, feasible and attractive. For North Macedonia and similar EU-candidate, migrant-origin economies, the findings suggest that circular migration may remain more aspiration than reality unless skills and employment policy, social dialogue, and broader institutional reforms are integrated into a credible migration-skills-return strategy.

Acknowledgments

This research was funded by the European Union under HORIZON-CL2-2023-TRANSFORMATIONS-01-03 grant number 101132435 - SKILLS4JUSTICE. The APC was funded by the European Union under

HORIZON-CL2-2023-TRANSFORMATIONS-01-03 grant number 101132435 - SKILLS4JUSTICE.

References

Agency for Youth and Sport. (2022). *National Youth Strategy 2023–2027*. Government of the Republic of North Macedonia.

Batista, C., Han, D., Haushofer, J., Khanna, G., McKenzie, D., Mobarak, A. M., ... & Yang, D. (2025). Brain drain or brain gain? Effects of high-skilled international emigration on origin countries. *Science*, *388*(6749), eadr8861.

Braun, V., & Clarke, V. (2019). Reflecting on reflexive thematic analysis. *Qualitative research in sport, exercise and health*, *11*(4), 589-597.

Braun, V., & Clarke, V. (2021). To saturate or not to saturate? Questioning data saturation as a useful concept for thematic analysis and sample-size rationales. *Qualitative research in sport, exercise and health*, *13*(2), 201-216.

Constant, A., & Massey, D. S. (2002). Return migration by German guestworkers: Neoclassical versus new economic theories. *International migration*, *40*(4), 5-38.

Danaj, S., & Meszmann, T. T. (2024). Weathering intermediated temporary labour mobility: social partners in Central and Eastern Europe after EU enlargement. *Transfer: European Review of Labour and Research*, *30*(1), 67-85.

De Haas, H. (2010). Migration and development: A theoretical perspective. *International migration review*, *44*(1), 227-264.

Dustmann, C., Fadlon, I., & Weiss, Y. (2011). Return migration, human capital accumulation and the brain drain. *Journal of Development Economics*, *95*(1), 58-67.

Engbersen, G., & Snel, E. (2025). Liquid migration: Dynamic and fluid patterns of post-accession migration flows. In *Mobility in transition* (pp. 21-40). Routledge.

Government of the Republic of North Macedonia. (2021). *Resolution on Migration Policy of the Republic of North Macedonia 2021–2025 with Action Plan*.

King, R., & Oruc, N. (2020). Migration in the Western Balkans–trends and challenges. *Migration in the Western Balkans*, 1-10.

Koch, A. (2017). The politics and discourse of migrant return: The role of UNHCR and IOM in the governance of return. In *International organisations and the politics of migration* (pp. 41-59). Routledge.

Levitt, P., & Lamba-Nieves, D. (2011). Social remittances revisited. *Journal of ethnic and migration studies*, *37*(1), 1-22.

Malterud, K., Siersma, V. D., & Guassora, A. D. (2016). Sample size in qualitative interview studies: guided by information power. *Qualitative health research*, *26*(13), 1753-1760.

Meļķe, Z. (2024). Why It Did Not Work: Structural Problems Behind Unsuccessful Return Migration in Latvia. *Central and Eastern European Migration Review*, *13*(2), 161-177.

Ministry of Labour and Social Policy. (2021). *National Employment Strategy 2021–2027 with Employment Action Plan 2021–2023*. Government of the Republic of North Macedonia.

Palinkas, L. A., Horwitz, S. M., Green, C. A., Wisdom, J. P., Duan, N., & Hoagwood, K. (2015). Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. *Administration and policy in mental health and mental health services research*, *42*(5), 533-544.

Peovski, F., Kitanovikj, B., & Cvetkoska, V. (2026). 5.2 Towards sustainable migration pathways: Investigating migrant skill formation, mismatches and justice in North Macedonia. *SKILLS FOR FAIR, EFFICIENT LABOUR MIGRATION*, 279.

Schiller, N. G., & Salazar, N. B. (2016). Introduction: Regimes of Mobility Across the Globe. In *Regimes of Mobility* (pp. 1-18). Routledge.

Schreier, S. (2024). From policy to reality: Examining the rippling effects of return migration governance in Nigeria. *International Migration Review*, 01979183241293388.

Tjaden, J., Liebe, U., & Bruscoli, D. (2025). Explaining re-migration preferences among assisted returnees: evidence from Discrete Choice Experiments in Sudan, Somalia, and Ethiopia. *Journal of Ethnic and Migration Studies*, *51*(17), 4362-4386.

Wahba, J. (2026). Return migration and economic development: opportunities and challenges. *World Development*, *198*, 107225.

Xiang, B., & Lindquist, J. (2014). Migration infrastructure. *International migration review*, *48*(1_suppl), 122-148.

Appendix

Table 1. Respondent profile of the qualitative sample

Stakeholder group	Code	Institutional type	Main institutional role in the labour market	Role seniority	Years of organisational experience
Policymaker and policy-facing actor	PM1	Government ministry / public sector	Digital transformation and public-sector employment policy	Senior/middle	3–10
Policymaker and policy-facing actor	PM2	Government ministry / public sector	Economy, employment measures and labour-market policy	Senior/middle	0–3
Policymaker and policy-facing actor	PM3	Chamber / employer-facing policy institution	Employer needs, skill demand and business-sector coordination	Senior/middle	3–10
Policymaker and policy-facing actor	PM4	Central government body	Strategic labour-market and migration-policy coordination	Senior/middle	3–10
Social partner / labour-market intermediary	SP1	NGO / development organisation	Skills development, employability and workforce support	Senior/middle	10–20
Social partner / labour-market intermediary	SP2	Regional migration initiative	Migration governance and regional cooperation	Senior/middle	3–10
Social partner / labour-market intermediary	SP3	Public employment agency	Labour-market matching, employment services and active measures	Senior/middle	0–3
Social partner / labour-market intermediary	SP4	University-based professional training/career centre	Career development, professional training and skills support	Senior/middle	3–10
Social partner / labour-market intermediary	SP5	Employers' organisation	Employer representation, social dialogue and labour-demand articulation	Senior/middle	3–10

Career Paths of IT Project Managers: Exploring Generic Competencies and Progression Patterns

Katarina Pažur Aničić, Vjeran Bušelić
University of Zagreb Faculty of Organization and Informatics
Pavlinska 2, 42 000 Varaždin, Croatia
{kpazur, vbuselic}@foi.unizg.hr

***Abstract**. This paper examines the career paths of IT project managers (PM) and the generic skills required. Based on a survey of 98 professionals in Croatia, the analysis shows that fewer respondents began their careers as junior PMs. In comparison, a larger number entered the field from either software development roles or entry-level non-technical positions. Once in project management, professionals typically remain in the role or progress to higher managerial positions. For junior PMs, the most important skills are rapid learning, teamwork, problem-solving, and adaptability. For experienced PMs, the key competencies are problem-solving, motivating others and guiding teams toward shared goals.*

Keywords. IT project manager; Career path; Generic skills

1 Introduction

With the rapid advancement of technology, career development within the IT field has expanded to include numerous new professional roles. For instance, the European framework of ICT professional profiles, introduced in 2018, identified 30 distinct roles (CEN-CENELEC, 2018), while the latest CEN initiative proposes 37 AI professional role profiles, including AI Project Manager and AI Product Manager (CEN, 2026). Generative AI is significantly influencing labour market changes. World Economic Forum (WEF, 2025) Future of Jobs Report points out that in the next 5 years, more than 92 mil. jobs will be displaced, 170 mil. jobs created, and that technological advancements urge permanent reskilling and upskilling. Project managers (PMs) are an occupation seeing growth and hold the 12th position among the largest growing jobs by 2030 (WEF, 2025). According to the U.S. Bureau of Labor Statistics, employment for Project Management occupations in United States is projected to grow by 6% from 2024 to 2034, faster than the average for all occupations (Bureau of Labor Statistics, 2025). Within this category, information technology project managers are among the core subfields, driven by continuous digital transformation and large-scale IT investments.

As evident from the literature, some previous research indicates that limited attention has been given to the careers and career pathways of project managers (Bredin and Söderlund, 2013), and this situation hasn't changed significantly recently (Alkhudary and Gardiner, 2021). A systematic literature review conducted by the last authors identified 42 relevant papers related to the query (career AND "project manag*") OR (career AND "program* manag*") OR (career AND "portfolio manag*") in Scopus and Business Source Ultimate (BSU). To see if the situation has changed in the last few years, we conducted the query TI=(career AND "project manag*") OR TI=(career AND "program* manag*") OR TI=(career AND "portfolio manag*") OR TI=(career AND "product* manag*") in Web of Science database in May 2026, which resulted in only 49 papers, 21 of which in the period from (2020-2026). The exact search in Scopus yields 56 documents, 26 of which were published between 2020 and 2026. Relevant papers are referenced further within the literature review sections. Based on the literature screening, it can be concluded that researchers have devoted limited focus to the careers and career frameworks of project managers, particularly in the information technology (IT) sector.

The present study, based on responses from 98 project managers, aims to enhance the understanding of project management careers, with a particular emphasis on the IT sector. Prior to presenting the empirical findings of this study, the introductory section defines the concepts of career and career path, followed by a review of prior research on project management careers, with particular attention to studies situated within the ICT context.

2 State of the Art

2.1 Definition of career and career path

As can be found in Alkhudary and Gardiner (2021) career research began in the 1950s, initially in psychology and later expanded to organizational studies. Arthur et al. (1989) define the concept of a career as "the evolving sequence of a person's work experience over time". Baruch (2004) highlights a shift from traditional linear career paths within one organization to shorter, multidirectional moves across multiple organizations, reflecting increased dynamism. Another key distinction is that in the traditional model, the primary responsibility for career development lies with the organization, while in the transformative models, it shifts to the individual (Baruch, 2004).

It is important to distinguish between career paths, which refer to the objective sequence of work experiences, and the subjective perceptions individuals have about their "career progress, personal development, status, or satisfaction" (Walker, 1976). A person's career path may evolve through changes in organization as well as through transitions in the workplace or professional role. This paper focuses on career paths in the context of professional roles.

The present study analyses formalized career paths of project managers within the ICT sector. Accordingly, the focus is placed on careers in the objective sense, rather than on project managers' subjective perceptions of their careers.

2.2 Research on project manager careers

Project Management Institute (PMI) defines a project manager as „the person assigned by the performing organization to lead the team that is responsible for achieving the project objectives" (PMI, 2013, p. 555). The role of a project manager is highly complex, as it requires managing ten knowledge areas simultaneously - ranging from scope, time, and cost to quality, risk, resources, communications, procurement, integration, and stakeholder management (PMI, 2013). Moreover, the project manager acts like a maestro - coordinating diverse team members much like a conductor guiding musicians to deliver a harmonious performance (Abramovici, 2001).

Although the role of project manager is clearly defined, previous research (Alkhudary and Gardiner, 2021; Marion et al., 2014; Richardson et al., 2015) indicates that many individuals enter it informally and unintentionally, often advancing from technical positions and reliant on on-the-job experience, soft skill development or simply showing the right attitude - leading to the characterization of project management as an "accidental profession" (Richardson et al., 2015). Moreover, this situation highlights significant gaps in formal career structuring and the education of future project managers.

Based on a systematic literature review followed by focus groups with experienced project managers, Alkhudary and Gardiner (2021) identified that typical project manager career paths often start "by accident" or "by chance", often as a result of demonstrating the right attitude (i.e., the needed competencies for project managers). Mid-career project managers often stay engaged in projects, sometimes shifting to internal consulting between assignments. This reflects organizational reliance and limited career mobility outside project management. They usually end their careers in project management, consultancy, coaching, or their own businesses, rather than advancing to senior roles (Alkhudary and Gardiner, 2021).

According to the research by Chen et al. (2019), the levels of the PM career model are junior staff, junior supervisor, senior supervisor, project manager, and project/program director. They found that communication, teamwork, and organizing emerge as the key competencies across all levels of project management, while at the junior supervisor level, building relationships was ranked highest. Not surprisingly, the research results indicate that senior supervisors value a broader range of skills than junior supervisors, suggesting that more competencies are required at higher levels of project management.

The research conducted by Heilmann (2006) in 2002, through interviews with 30 managers from three Finnish ICT companies and three paper industry firms, supplemented by a short career anchor questionnaire, revealed three main career paths: general manager, specialist, and project manager, all of which start from a specialist role.

Hölzle (2010) recognized four stages in project managers career path, starting from project manager, then senior project manager, project director and program director, and aligned those with the leadership roles. However, the research results show that most of the studied companies do not differentiate between the leadership career path and the project managers' career path and the research advocates for organizations to design career paths for project managers or integrate those paths into the existing career path system.

As claimed by Alkhudary and Gardiner (2021), based on a systematic literature review, a gap exists between career path research and project management, which affects project managers' careers and well-being, as well as talent development in organizations, underscoring the need for future research on project managers' career development. Prior studies on project management career paths and progression rely primarily on respondents' reflections about how their careers developed (Alkhudary and Gardiner, 2021; Marion et al., 2014; Richardson et al., 2015), and not on concrete information about the professional roles that lead to the PM.

Therefore, this study examines the actual unfolding of project managers' career trajectories over the course of their professional lives – from entry-level positions

to their current roles, including all career changes and advancements.

2.3 Project manager competencies

As the „maestro“ orchestrating the team to deliver project outcomes, a project manager is required to have a wide range of competencies that can generally be divided into three categories: technical, managerial/ organization and people management skills (Schwalbe, 2019; El-Sabaa, 2001). Many of these skills are referred to as generic skills, which are frequently studied in the context of employability. Generic skills are not only recognized as important for a successful transition from higher education to the world of work (Jackson & Chapman, 2012; Pažur Aničić et al., 2023; Peng et al., 2016), but also for career progress in the workplace, especially within roles such as project managers (Chen et al., 2019; El-Sabba, 2001).

Research conducted by Hölzle (2010), which involved a qualitative study of 20 companies that have implemented career paths for project managers, revealed that 80% of these companies recognize social competence (including communication and cooperation skills) as the most important skill for a project manager. This was closely followed by leadership competence, valued by 50% of the companies surveyed.

According to El-Sabaa (2001), human skills (85.3%) emerged as the most critical for project managers for all sectors (including information system projects), including communication, mobilization, delegation, coping, political sensitivity, self-confidence, and enthusiasm. Conceptual and organizational skills (79.6%) ranked second, covering planning, organizing, goal orientation, holistic project vision, understanding industry context, and problem-solving. By contrast, technical skills (50.46%) were viewed as least essential, involving specialized knowledge of tools, techniques, methods, processes, required technologies, and computer use.

Chen et al. (2019) found that higher competency scores indicate greater importance for effective job performance. At the junior supervisor level, building relationships, communication, and teamwork ranked highest, while at the senior supervisor, project manager, and project/program director levels, communication, teamwork, and organizing were most critical. Overall, communication, teamwork, and organizing emerge as the key competencies across all levels of project management. Not surprisingly, their research as well indicates that senior supervisors value a broader range of skills than junior supervisors, showing that more competencies are required at higher levels of project management.

Borg and Scott-Young (2020) examined the work readiness of junior project managers from the perspective of employers in the Australian construction industry, identifying three key dimensions of readiness: empathetic communication, passion, and construction knowledge. However, graduates were seen as strong in professional presentation, meeting conduct, and use of technology, but weak in seeking help, handling confrontation, applying basic construction knowledge, and professional writing.

Given the focus of this paper on the ICT industry, it is essential to consider the specific characteristics of IT projects. In some sectors, project management is regarded as a temporary managerial assignment (Bredin & Söderlund, 2013), reflecting the inherently temporary nature of projects as endeavours with a defined end. By contrast, the ICT industry is predominantly project-oriented, where project managers frequently oversee multiple projects simultaneously, and their role extends beyond the boundaries of individual projects. Furthermore, IT project management is closely intertwined with IT product management, as project outcomes often take the form of IT products such as software. As Demirel and Camcı (2022) note, project and product scopes frequently overlap, with product lifecycles typically extending well beyond the duration of individual projects.

Although the present study focuses on a single industry, prior research indicates that differences in project management research are not strongly industry-specific. For example, El-Sabaa (2001) reported no significant variation in the relative importance of project management skills across information systems, electricity, and agricultural projects. Similarly, Crawford et al. (2013), based on in-depth interviews with project managers in the engineering, construction, and IT sectors in Australia, concluded that generational differences among project managers were more pronounced than those attributable to industry context. Moreover, the analysis of previous research reveals that authors employ different classifications of project management skills (Borg and Scott-Young, 2020; Chen et al., 2019; El-Sabaa, 2001; Hölzle, 2010), particularly for soft or generic skills. In general, there is a lack of conceptual clarity regarding competencies that are often referred to as soft, generic, transversal, or employability skills/competencies. In this study, we applied a list of generic competencies previously used to examine the importance of students' and recent graduates' generic skills (Pažur Aničić & Bušelić, 2021; Pažur Aničić et al., 2023).

3 Methodology

The present study aims to answer three research questions: 1) From which entry positions do IT project managers come?, 2) What are the typical career paths of IT project managers? and 3) Which generic competences are needed for entry level and experienced IT project managers?

This study is based on data collected from a study conducted among IT project managers in Croatian companies, between the beginning of October and the

beginning of December 2024. The survey was distributed via the authors' contacts and social networks, with the support of Project Management Institute Croatia (PMI Hrvatska), and relevant IT associations like Croatian Independent Software Exporters (CISEx), Croatian Employers' Association - Information and Communication Industry Association (HUP-ICT), IT Association at Croatian Chamber of Economy (HGK–IKT), and Agilist IT.

3.1 Instrument

The survey was prepared in Google Docs, consisting of four parts: 1) Respondents' professional experience in IT project management, 2) Using PM methodologies, methods and tools, 3) Project manager competencies and 4) AI in project management. The first part of the study, which explores project management practices among IT project managers in Croatia, focusing on their use of traditional and agile methodologies, prevalent tools, and generative AI integration within project management processes, was reported by Pažur Aničić & Bušelić (2025). The current study focuses on respondents' professional careers in IT project management and the generic competencies required for such careers. In particular, the study analyses answers to survey questions related to years of working experience and years of experience in managing IT projects, career paths from early career stages to the present, generic competencies required in their current position, and competencies required for junior project managers. The list of competencies assessed by respondents included 25 skills, identical to those used by Pažur Aničić & Bušelić (2021) in the study on the importance of generic skills of ICT graduates from the perspectives of employers, teaching staff, and students.

3.2 Sample

A total of 98 respondents participated in the research. However, not all of them provided detailed information about their career path. The sample for analysis includes 69 project managers from Croatia with experience in leading IT projects. Within the sample, 26 (38%) were females and 43 (62%) were males. Regarding years of experience, three-quarters of respondents (75%) had more than 11 years of work experience, with the largest group (N=26, 38%) having more than 20 years of experience. Regarding years of experience in IT projects, most respondents (N=45, 65%) have 3-15 years of experience. The sample generally covers both junior and senior positions. Another point supporting the representativeness of the sample is that 48 respondents hold certifications or education related to project management. Most of the respondents, 19 of them, hold certification in project management, 8 have a specialised education in PM, 17 have some kind of agile PM certification and 4 have other types of certifications.

3.3 Data analysis

Since respondents were allowed to provide information about their career path in any form, preparing the data for analysis involved coding and categorizing their responses. Although there are established theoretical classifications of roles in the ICT field, such as the e-CF professional profile (CEN-CENELEC, 2018) that divides roles into 30 categories, this complexity hinders analysis. Therefore, for this paper, we have developed a simplified classification that includes 10 ICT roles, as shown in Table 1.

These roles encompass positions across software development, non-technical areas, IT/technical roles, and management. Each role is further divided into two levels: entry/junior and senior. The roles of project managers and product managers, which are the primary focus of this research, are analyzed as separate roles.

Table 1. Categories of ICT roles

Category code	Category explanation	Example of roles
L1-swd	Level 1 - Software development	Software developer, web developer, software engineer, web admin, junior developer
L2-swd	Level 2 – Software development	Software developer team lead, solution architect, software implementation consultant, Quality Assurance (QA)
L1 -nT	Level 1- non-technical role	Call center agent, business analysis, project associate, specialist, customer service agent, sales representative...
L2-nT	Level 2 – non technical role	Team lead, senior/lead consultant, senior billing agent, key account manager, business development manager...
L1 IT	Level 1 – IT role	Network specialist, system engineer, IT support specialist, junior infrastructure specialist...
L2 IT	Level 2 – IT role	System engineer team lead, IT support team lead, technical lead, ITSM specialist
L1-M	Level 1 – management (manager)	Department manager, Operations manager, service delivery manager, QA manager, sales manager, head of department...
L2-M	Level 2 – management (director)	Director, QA director, CTO, consulting director, director of IT, director of operations
ProjectM	Project management	Project manager, Lead project coordinator, PMO manager, Senior project manager...
ProductM	Product management	Product manager, Product owner, Group product manager...

Another challenge in preparing the data for analysis was related to assigning years to specific positions. For example, some respondents indicated overlapping time frames for different roles (e.g., 2000–2002 for one position and 2002–2005 for another). To ensure that the total duration of a career path was accurately represented, each year was allocated to only one role.

4 Results

The study results are presented in two main sections – career paths of project managers and generic competencies needed for PM roles. It is important to emphasize that, at some point in their career, each respondent was in a role of either project or (in a smaller number of cases) product manager. At the time of the survey, 44 respondents were in a project manager position, while 10 were in a product manager position and 15 in other positions.

4.1 Project managers career paths

Fig. 1 shows visually the career paths for all 69 respondents, ordered based on the first appearance of a PM role. The vertical axis shows each respondent's career position, while the horizontal axis shows years of career. It is evident that only 10 respondents began their careers in project management positions, likely as junior PMs. Additionally, 9 respondents were not in a PM role in their careers but worked as product managers. Moreover, the visual representation in Fig. 1 shows that once reaching a PM position, a person usually remains there for a longer period.

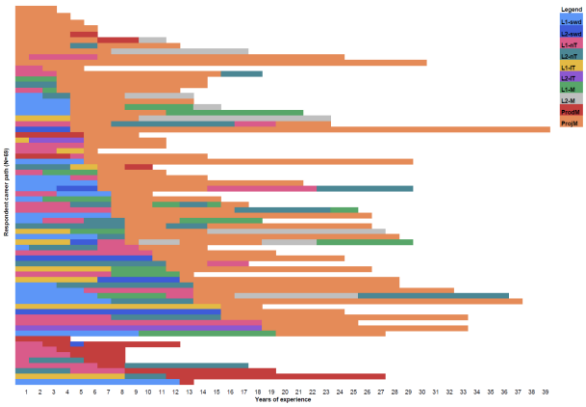


Figure 1. Career paths of IT project managers according to years, sorted according to the first appearance of PM role (N=69)

As evident from Table 2, 27,54% of future PMs start their careers as junior software developers (L1-swd) and 27,54% (N=19) in non-technical junior roles (L1-nt), which makes up more than 50% of the total sample. Within the total sample, only 14,49% (N=10) of respondents start their careers as PMs and 4,35% (N=3) as product managers. Even though in total more respondents started their career on level 1 (N=47), smaller number starts on level 2 positions (N=9). In Table 2, the total number of respondents decreases over the years because some of them have not yet reached 5 or 10 years of experience.

Table 2. Respondents' entry positions and positions in 5 and 10 years

Year / Role	Y1 (N)	Y1 (%)	Y5 (N)	Y5 (%)	Y10 (N)	Y10 (%)
L1-swd	19	27.54	10	15.15	1	1.75
L2-swd	3	4.35	5	7.58	3	5.26
L1 -nT	19	27.54	12	18.18	3	5.26
L2-nT	5	7.25	7	10.61	7	12.28
L1 IT	8	11.59	6	9.09	1	1.75
L2 IT	1	1.45	2	3.03	1	1.75
L1-M	1	1.45	3	4.55	5	8.77
L2-M	0	0	0	0	5	8.77
ProjM	10	14.49	17	25.76	28	49.12
ProductM	3	4.35	4	6.06	3	5.26
Sum	69		66		57	

Sankey diagram in Fig. 2 shows respondents' entry positions on the left side and their position in the fifth year of their careers on the right side. The term END in visualization reflects the respondents who have less than 5 years of working experience so only the information about their entry position is used. The percentage of PMs rose from 14,49% at career entry to 25,76% in five years. Moreover, it is evident that PMs come from all entry positions, except product management and L2-IT, and only one PM switched a role to product manager in a five-year time span.

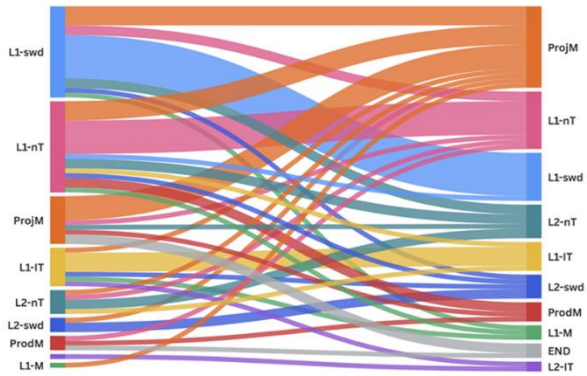


Figure 2. Sankey diagram for career change from year 1 to year 5 (created in Flourish, 2024)

Sankey diagram in Fig. 3 shows the situation after 10 years of career. There is an evident growth in number of PMs (49,12%), which accounts for almost 50% of all roles. Moreover, it is evident that the number of respondents working as L1-swd or L1-IT specialists is nearly negligible. Additionally, after 10 years, there are more respondents in managerial positions (L1-M and L2-M). After 5 years of career, there were no respondents with the L2-M role.

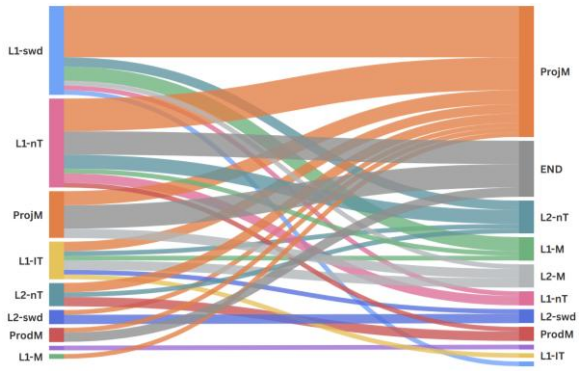


Figure 3. Sankey diagram for career change from year 1 to year 10 (created in Flourish, 2024)

Further, we analyse career paths of project managers after the PM role, which is evident from graphical representation in Fig. 1. There is a significant number of respondents who, once they enter the PM role (N=37), remain there, as well as those who become product managers (N=7). Three project and product managers (N=3) entered that role in the year of the survey, so their career path after that could not be observed. There is also a notable number of respondents who pursued other careers after project management (N=15) and did not return to the PM role. Approximately half of them (N=7) advanced to higher management positions, becoming directors (Level 2 – M), either directly or through a product management role or a Level 1 management position. Two respondents moved from PM to Level 1-M (N=2) and remained there, while a few continued their careers as product managers (N=2). There are also a few cases of project managers who transitioned to non-technical positions, either Level 1-nT or Level 2-nT, after their project management careers (N=4). Additionally, some respondents moved from PM to another position and then returned to the PM role (N=6).

4.2 Generic competencies needed

To analyse the importance of generic competencies, all responses were included (N=98), not only those from respondents who reported their career paths. During data preparation, however, responses indicating more than five generic competencies were excluded, as the questions were as follows:

Q1: On a scale of 25 generic competencies, please select the 5 that are personally most important to you when managing IT projects.

Q2: On a scale of 25 generic competencies (the same as in the previous question), please select the 5 most important generic competencies for individuals at the beginning of their careers who wish to start a career in IT project management (our students).

Consequently, 87 responses were analysed to identify generic skills required for entry-level PM positions, and 85 to determine the key skills for respondents' current positions. Table 3 presents the overall importance of the ten most relevant generic skills, both for entry-level PM positions and for respondents' current positions.

Table 3. The ten highest-rated generic competencies for project managers

Skills name	Entry position N=87	Current position N=85
05. Ability to rapidly acquire new knowledge	58.62 %	28.24 %
14. Ability to work in a team	51.72 %	30.59 %
07. Ability to identify and resolve problems	48.28 %	74.12 %
09. Ability to adapt and act in new situations	47.13 %	41.18 %
02. Ability to apply knowledge in practical situations	28.74 %	
20. Taking responsibility and assessing the quality of one's work	26.44 %	
08. Ability to perform well under pressure	25.29 %	40.00 %
12. Ability to manage projects	22.99 %	40.00 %
04. Analytical thinking that includes mathematical skills	21.84 %	22.35 %
17. Ability to argument and make reasoned decisions	19.54 %	24.71 %
15. Ability to motivate people and move toward common goals		56.47 %
13. Ability to use time effectively		28.24 %

The survey results revealed that, from the perspective of all 87 respondents, the most valued skills for entry-level project managers are the ability to rapidly acquire new knowledge (58.62%), work effectively in a team (51.72%), identify and resolve problems (48.28%), and adapt and act in new situations (47.13%), emphasizing the importance of agility, collaboration, and problem-solving in early project management careers. For respondents in their current positions, however, the most important skills shift toward the ability to identify and resolve problems (74.12%) and motivate people and move toward common goals (56.47%), followed by adapting and acting in new situations (41.18%), performing well under pressure (40.00%), and managing projects (40.00%). This result highlights the growing need for leadership, resilience, and advanced problem-solving as responsibilities increase with experience. Interestingly, eight of the top ten skills overlap between the entry-level PM positions and respondents' current positions. The only difference is that, at the entry level, the ability to apply knowledge in practical situations (28.74%) and taking responsibility while assessing the quality of one's work (26.44%) are emphasized, whereas later in the career, greater importance is placed on the ability to motivate people and move toward common goals (56.47%) and the ability to use time effectively (28.24%).

5 Discussion

This study contributes to a better understanding of project management career paths in the IT industry and generic competencies needed in project management position.

Three main patterns of project management career paths can be observed. The first comprises professionals who begin their careers as software developers and subsequently progress to project management roles. The second includes those who start in entry-level, non-technical positions such as business analyst, project associate, specialist, customer service agent, or sales representative. The third, though less common, involves professionals who enter the field directly as junior project managers. These findings are not unexpected and are consistent with previous research (Alkhudary & Gardiner, 2021; Marion et al., 2014; Richardson et al., 2015), which highlights that many individuals enter project management informally and unintentionally - often transitioning from technical positions and relying on on-the-job learning, soft skill development, or demonstrating the right attitude. However, in the context of the IT industry, it is difficult to speak of 'accidental project managers.' Due to the specific nature of IT projects (Schwalbe, 2019), particularly in the development of software solutions, it is desirable for project managers to possess certain knowledge of programming, client needs analysis, and related areas. Such competencies enable them to make informed decisions on time planning, budgeting, risk management, and other key project aspects. According to Pažur Aničić & Bušelić (2025), based on the analysis of the same sample of respondents as in the present study, the majority of projects typically span 6 to 12 months. The next most common are projects lasting 12 to 24 months. Accordingly, a certain amount of time is required for early-career professionals to experience all phases of the project life cycle in other roles (e.g., as software developers or business analysts) and to become familiar with its various aspects. In this context, it can be argued that, although perhaps unintentionally, one does not become a project manager in the IT industry by chance; rather, the necessary skills are acquired through experience in other entry-level positions.

The analysis of generic competencies required of project managers provides valuable insights into their career paths and progression. When these results are considered in relation to previous studies, certain patterns emerge. Pažur Aničić & Bušelić (2021), drawing on the same set of 25 generic competencies to examine their relevance for ICT graduates, found that the ability to rapidly acquire new knowledge was ranked highest by all three stakeholder groups - teachers, employers, and students. The next two competencies were the ability to identify and resolve problems and the ability to apply knowledge in practical situations. Both employers and students ranked ability to work in teams in fourth place. These findings suggest that acquiring new knowledge is consistently perceived as the most important generic competency, while other highly ranked competencies identified as essential for ICT graduates are equally applicable to future project managers. Little research can be found in the literature, particularly on IT project manager skills. Among the few earlier studies on project management competencies, El-Sabaa (2001) identified desirable traits of project managers across several sectors in Egypt, highlighting human skills as the most prominent, followed by organizational and technical skills. Napier et al. (2009), based on qualitative interviews with 19 professionals, identified Planning and Control as the most prominent skill category for IT project managers, followed by General Management, Leadership, Communication, Team Development, Client Management, System Development, Problem Solving, and Personal Integrity. Therefore, this study contributes to the understanding of important generic skills for project managers, both at the entry level and later as experienced professionals.

Furthermore, the paper demonstrates the changes in the importance of generic competencies in career progression. By linking project managers' career paths with the importance of specific generic competencies for entry PM roles and more mature roles (according to the respondent's current position), this research bridges career theory and competency-based perspectives, thereby extending previous studies on the development of generic skills. In particular, the findings resonate with earlier research on graduates' competencies (Jackson & Chapman, 2012; Pažur Aničić et al., 2023; Pažur Aničić & Bušelić, 2021; Peng et al., 2016), highlighting the relevance of these skills for students preparing for careers as IT project managers. This perspective adds to the literature on graduate employability and offers useful guidance for higher education, indicating key competencies needed for future project managers.

Additionally, the study builds upon and refines previous research (e.g., El-Sabaa, 2001; Napier et al., 2009) by situating competencies within the IT context and stressing the importance of particular generic skills for project managers. Most notably, those are the ability to acquire new knowledge, teamwork, problem-solving, and adaptability to new situations at entry-level positions, as well as problem-solving and motivating others at more senior levels.

6 Conclusion

This study enhances the understanding of IT project management careers, from early stages to maturity, by empirically examining career paths within the IT industry, where the concept of the "accidental project manager" may not be entirely applicable. The findings reveal distinct patterns of career entry, which include

transitions from software development roles, non-technical roles, and junior project management positions. This approach provides a more nuanced understanding of how IT project managers develop their professional identities and advance in their careers, while also adding a complementary perspective to previous research on project manager career development (Alkhudary and Gardiner, 2021; Bredin and Söderlund; 2013; Marion et al., 2014; Richardson et al., 2015). Moreover, the existing literature appears to provide limited insight into project management career paths in detail, particularly in terms of specific positions held and the time spent in each.

Despite the recognized theoretical and practical contributions of this study, several limitations should be noted. First, the research is based on a sample from a single country, which limits the generalizability of the findings. However, the IT industry is relatively consistent worldwide, and Croatian companies and project managers frequently participate in both national and international projects of varying size and complexity. Moreover, prior studies on project management careers have also been primarily qualitative, typically based on smaller samples and focused on single-country contexts (Bredin & Söderlund, 2013; Crawford et al., 2013).

A second limitation is the lack of a strong theoretical model on which to ground the study, as research on project managers' career paths, particularly within IT, remains scarce. As a result, the empirical component of this study followed an inductive approach, aiming to identify patterns of career progression that can inform future research. By contrast, the part of the study focused on the importance of generic competencies relied on established frameworks and methodologies already applied in previous research (Pažur Aničić & Bušelić, 2021; Pažur Aničić et al., 2023).

The results of the present study may serve as a basis for developing a project manager career model in the IT industry, grounded in the observed empirical patterns and aligned with the work of Bredin & Söderlund (2013), who conceptualize career models in terms of defined levels, their relation to project complexity, and corresponding competence requirements and training programs.

Moreover, future research could further examine the career paths of project managers entering the profession from different starting positions - for example, how those who began as software developers navigate their careers compared to those from non-technical backgrounds. Such studies could shed light on which knowledge and skills support their progression as project managers and which gaps may hinder their development.

References

Abramovici, A. (2001). Maestro, please! PM Network, 15(3), 55–57.

Alkhudary, R., & Gardiner, P. (2021). Stages in project managers' careers: Learning and growth opportunities. *International Journal of Project Management.* 39(5). 536–545. https://doi.org/10.1016/j.ijproman.2021.03.006

Arthur, M.B., Hall, D.T., Lawrence, B.S., (1989). Generating new directions in career theory: the case for a transdisciplinary approach. In: Arthur, M.B., Hall, D.T., Lawrence, B.S. (Eds.), *Handbook of Career Theory.* Cambridge University Press. New York

Baruch, Y. (2004). Transforming careers: From linear to multidirectional career paths: Organizational and individual perspectives. *Career Development International.* 9(1). 58–73. https://doi.org/10.1108/13620430410518147

Borg, J., & Scott-Young, C. M. (2020). Employers' perspectives on work readiness in construction: Are project management graduates hitting the ground running? *International Journal of Managing Projects in Business*. 13(6). 1363–1379. https://doi.org/10.1108/IJMPB-10-2019-0238

Bredin, K., & Söderlund, J. (2013). Project managers and career models: An exploratory comparative study. *International Journal of Project Management.* 31(6). 889–902. https://doi.org/10.1016/j.ijproman.2012.11.010

Bureau of Labor Statistics. U.S. Department of Labor. Occupational Outlook Handbook. Project Management Specialists. Retrieved from: https://www.bls.gov/ooh/business-and-financial/project-management-specialists.htm (visited September 24, 2025).

CEN-CENELEC. (2018). European ICT professional role profiles – Part 1: 30 ICT profiles (CWA 16458 1:2018). European Committee for Standardization. https://www.cencenelec.eu/media/CEN-CENELEC/AreasOfWork/CEN%20sectors/Digital%20Society/CWA%20Download%20Area/ICT_SkillsWS/16458-1.pdf

CEN (2026) CWA Workshop Agreement "AI Professional Role Profiles and Educational Profile". Available at: https://www.cencenelec.eu/media/CEN-CENELEC/CWAs/RI/2026/cwa18398_2026.pdf

Chen, T., Fu, M., Liu, R., Xu, X., Zhou, S., & Liu, B. (2019). How do project management competencies change within the project management career model in large Chinese construction companies? *International Journal of Project Management.*

37(3). 485–500. https://doi.org/10.1016/j.ijproman.2018.12.002

Crawford, L., French, E., & Lloyd-Walker, B. (2013). From outpost to outback: Project career paths in Australia. *International Journal of Project Management*. 31(8), 1175–1187. https://doi.org/10.1016/j.ijproman.2013.03.003

Demirel, Ş. and Camcı, A. (2022). Project management vs product management: A view on the journey from project to product in software development. In *Proceedings of the 12th Annual International Conference on Industrial Engineering and Operations Management.* Istanbul. Turkey. IEOM Society International. https://doi.org/10.46254/AN12.20220934

El-Sabaa, S. (2001). The skills and career path of an effective project manager. *International Journal of Project Management.* 19(1). 1–7. https://doi.org/10.1016/S0263-7863(99)00034-4

Flourish. (2024). Flourish data visualization platform. available at: https://flourish.studio/

Heilmann, P. (2006). Finnish managers' careers in ICT and paper business sectors. *Baltic Journal of Management.* 1(3). 339–351. https://doi.org/10.1108/17465260610690962

Hölzle, K. (2010). Designing and implementing a career path for project managers. *International Journal of Project Management.* 28(8). 779–786. https://doi.org/10.1016/j.ijproman.2010.05.004

Jackson, D., & Chapman, E. (2012). Non-technical competencies in undergraduate business degree programs: Australian and UK perspectives. *Studies in Higher Education.* 37(5). 541–567. https://doi.org/10.1080/03075079.2010.527935

Marion, J. W., Richardson, T., & Earnhardt, M. (2014). Project manager insights: An analysis of career progression. *Organisational Project Management.* 1(1). 53–73. https://doi.org/10.5130/opm.v1i1.3949

Napier, N. P., Keil, M., & Tan, F. B. (2009). IT project managers' construction of successful project management practice: A repertory grid investigation. *Information Systems Journal.* 19(3). 255–282. https://doi.org/10.1111/j.1365-2575.2007.00264.x

Pažur Aničić, K., & Bušelić, V. (2021). Importance of Generic Skills of ICT Graduates—Employers, Teaching Staff, and Students Perspective. *IEEE Transactions on Education,* 64(3), 245–252. https://doi.org/10.1109/TE.2020.3034958

Pažur Aničić, K., & Bušelić, V. (2025). From Methodology and Tools to AI: Insights into IT Project Management Practices in Croatia. *In 2025 MIPRO 48th ICT and Electronics Convention* (pp. 1829–1834). IEEE. https://doi.org/10.1109/MIPRO65660.2025.11131898

Pažur Aničić, K., Gusić Munđar, J. & Šimić, D. Generic and digital competences for employability — results of a Croatian national graduates survey. *Higher Education* 86, 407–427 (2023). https://doi.org/10.1007/s10734-022-00940-7

Peng, L., Zhang, S., & Gu, J. (2016). Evaluating the competency mismatch between master of engineering graduates and industry needs in China. *Studies in Higher Education*. 41(3). 445–461. https://doi.org/10.1080/03075079.2014.942268

Project Management Institute [PMI]. (2013). A guide to the Project Management Body of Knowledge (PMBOK® guide) (5th ed.). Newtown Square, PA: Project Management Institute.

Richardson, T. M., Earnhardt, M. P., & Marion, J. W. (2015). Is project management still an accidental profession? A qualitative study of career trajectory. *SAGE Open.* 5(1). https://doi.org/10.1177/2158244015572098

Schwalbe, K. (2019). Information Technology Project Management. 9th Edition. USA: Cengage.

Walker, J. W. (1976). Let's get realistic about career paths. *Human Resource Management*. 15(3), 20–25. https://doi.org/10.1002/hrm.3930150302

World Economic Forum (WEF) (2025). The Future of Jobs Report 2025. Insight report. https://reports.weforum.org/docs/WEF_Future_of_Jobs_Report_2025.pdf

A Conceptual Framework for AI-Supported Decision Making in Pre-Incubation Programmes

Petra Vondra, Nina Begičević Ređep, Tomislav Peharda
University of Zagreb
Faculty of Organization and Informatics
Varaždin, Croatia
{petra.vondra, nina.begicevic, tomislav.peharda}@foi.unizg.hr

Abstract. *Artificial intelligence (AI) offers significant opportunities to support the early-stage of entrepreneurial idea development as a part of pre-incubation programme. To examine current research in this area, a scoping literature review screened 289 records identified from Scopus and Web of Science, resulting in 20 studies for qualitative synthesis. Through a literature review, this paper addresses key research questions regarding the entrepreneurial stages, decision-making domains, AI technologies and approaches and system architectures related to the application of AI in decision-making within pre-incubation programs and early-stage idea development within higher education. Based on these insights, the paper develops a conceptual framework specifically tailored for AI-driven decision-making during these early phases of entrepreneurial idea development.*

Keywords. decision-making, pre-incubation, artificial Intelligence

1 Introduction

Universities are increasingly expected to contribute not only to education and research but also to innovation and regional economic development by fostering entrepreneurial competencies among students. Consequently, entrepreneurial idea development has become an important component of universities' third mission, encouraging higher education (HE) institutions to create environments that support entrepreneurial thinking, idea development and venture creation (Mele et al., 2022; Bielicki & Weinert, 2021; Wright et al., 2017).

Pre-incubation programmes have emerged as an important mechanism for supporting students during the earliest stages of entrepreneurial idea development. However, this phase is characterized by high uncertainty and numerous strategic decisions, including opportunity recognition, idea validation, business model development, partner selection and resource allocation. These decisions substantially influence subsequent venture development and entrepreneurial success (Kalisty, 2023; Bielicki, 2023). Consequently, improving the quality of entrepreneurial decision-making has become an increasingly important objective of university pre-incubation programmes.

Recent advances in AI create new opportunities to support entrepreneurial decision-making by providing data-driven insights, idea generation, information analysis and personalized recommendations (Shi et al., 2026; Qiu, 2024; Chen, 2026; Park, 2025). Nevertheless, existing research provides valuable but fragmented evidence on AI in entrepreneurship and HE, with relatively little attention devoted to a coherent conceptualisation of AI-supported decision-making throughout the university pre-incubation process.

To address this gap, this study systematically analyses existing literature on AI-supported decision-making in pre-incubation/entrepreneurship and HE through a scoping literature review and proposes a conceptual framework for AI-supported decision-making in university pre-incubation programmes.

2 Background

2.1 Pre-incubation as a Foundation for Entrepreneurial Development

Pre-incubation has become an increasingly important component of contemporary entrepreneurship education, serving as a bridge between entrepreneurial learning and formal business incubation. Unlike business incubators, which primarily support established ventures, pre-incubation focuses on the earliest stages of entrepreneurial development, where ideas are explored, evaluated, validated and gradually transformed into viable business opportunities (Mele et al., 2022).

Originally introduced as a preparatory stage for early-stage venture and student start-up development, pre-incubation has evolved beyond preparing ventures for incubation (Jones et al., 2013; Kalisty, 2023).

Contemporary programmes increasingly emphasise entrepreneurial learning, opportunity exploration, experimentation and the development of entrepreneurial capabilities that remain valuable regardless of whether participants ultimately establish a company (Bielicki, 2023).

Modern pre-incubation programmes typically combine multidisciplinary collaboration, experiential learning, mentoring, networking, challenge-based activities and increasingly AI-supported innovative teaching approaches. Participants work in interdisciplinary teams to develop and validate entrepreneurial ideas while receiving continuous feedback from mentors, academics, industry experts, and, increasingly, AI-supported learning environments that provide personalised guidance and decision support (Mele et al., 2022; Kadoić et al., 2026, Chen, 2025).

In addition, pre-incubation can be situated within a broader student start-up ecosystem. (Wright et al., 2017) argue that student entrepreneurship is supported not only through classroom-based entrepreneurship education, but also through experiential mechanisms, university support structures, pre-accelerators, accelerators, entrepreneurs, support actors, investors, funding mechanisms and the wider external context. This ecosystem perspective is relevant for pre-incubation because it highlights that student venture development depends on coordinated support from multiple actors rather than on isolated educational activities (Wright et al., 2017; Mele et al., 2022). Similarly, university business incubators facilitate collaboration between students, universities, companies, mentors and industry partners by providing networking opportunities, knowledge exchange, mentoring and access to entrepreneurial ecosystems (Kazhenov, 2023).

Accordingly, pre-incubation can be understood as a structured environment that supports entrepreneurial idea development while enabling participants to experiment, validate assumptions and prepare for subsequent incubation under conditions of uncertainty. Recent studies further suggest that embedding pre-incubation within formal university curricula provides students with a structured entrepreneurial pathway, guiding them through idea development, customer validation, business-model design, pitching and transition towards incubation or acceleration (Popov, 2022).

2.2 Decision-Making During Pre-incubation

The pre-incubation programme typically focuses on idea generation, problem-solution validation and initial market research to transform a raw concept into a structured business model. To facilitate this transformation, the program integrates structured training designed to elevate the teams' technological, business and entrepreneurial knowledge and skills, alongside personalized mentoring and coaching from industry experts to guide them through technical and market obstacles. Throughout all these stages and processes, generative artificial intelligence can be utilized as a valuable tool to support decision-making and enhance the learning experience (Kadoić et al., 2026; Shi et al., 2026).

During this stage, teams must decide precisely which core customer pain point they are solving and define their specific target market segment based on early validation data. They also need to determine the essential features required for their Minimum Viable Product (MVP) and evaluate whether the founding team possesses the right mix of skills to execute the vision. Ultimately, the final and most critical decision teams face at the end of pre-incubation is whether to refine the concept based on feedback, stop the project, or formally advance to the incubation and market-entry stage.

Unlike established organisations, early-stage entrepreneurial teams operate with limited information, knowledge and experience, evolving assumptions and constrained resources, requiring a continuous sequence of strategic and operational decisions throughout entrepreneurial idea development. These decisions are part of opportunity recognition, idea evaluation, business-model development, resource allocation, customer validation and readiness for incubation, all of which influence subsequent venture development (Kalisty, 2023; Mele et al., 2022).

The entrepreneurial process is inherently iterative. Consistent with lean startup and scientific approaches to entrepreneurship, entrepreneurs continuously formulate hypotheses, gather evidence, validate assumptions and refine or pivot their ideas as new information becomes available (Ries, 2011; Camuffo et al., 2024).

This perspective is consistent with decision-making theory, which conceptualises decision-making as a structured process involving problem identification, information gathering, evaluation of alternatives, implementation and continuous assessment (Taherdoost & Madanchian, 2023). In entrepreneurial settings, scientific approaches encourage entrepreneurs to reduce uncertainty by systematically formulating and testing hypotheses, gathering evidence and updating decisions before committing significant resources (Camuffo et al., 2024).

Despite these advances, the pre-incubation programme literature still predominantly describes educational activities, mentoring approaches and programme structures rather than conceptualising pre-incubation as an integrated entrepreneurial decision-making process. While individual activities such as idea generation, mentoring or business-model development are well represented, relatively little attention has been devoted to understanding how these activities collectively support interconnected

entrepreneurial decisions throughout the pre-incubation lifecycle (Kalisty, 2023; Mele et al., 2022).

2.3 Artificial Intelligence for Entrepreneurial Decision-Making

AI is increasingly transforming entrepreneurship by providing new opportunities to support entrepreneurial decision-making (Rong, 2024; Shi et al., 2026). Rather than replacing entrepreneurs, contemporary AI increasingly functions as an augmentation technology that supports human judgement through information retrieval, idea generation, predictive analytics, contextual recommendations and evidence synthesis while preserving human responsibility and strategic control (Shi et al., 2026; Orozova et al., 2024).

Recent advances in generative AI, large language models, machine learning, predictive analytics, recommendation systems, simulation and gamification (Zirek, 2024) have significantly expanded the range of AI applications available throughout the entrepreneurial process. These technologies support activities such as opportunity recognition, business-model development, market analysis, customer validation, risk assessment and entrepreneurial learning (Shi et al., 2026; Zirek, 2024). At the same time, entrepreneurs increasingly require competencies such as AI literacy, data interpretation, systems thinking and the ability to critically evaluate AI-generated outputs (Orozova et al., 2024).

Despite these advances, existing AI applications remain largely task-specific. Most studies focus on individual functions, such as idea generation, mentoring, personalised learning, business analytics, or simulation, rather than supporting entrepreneurial decision-making as a continuous process spanning the entire pre-incubation lifecycle (Chen, 2026; Park, 2025; Qiu, 2024; Zirek, 2024).

Consequently, the literature lacks an integrated framework for AI-supported decision-making throughout the pre-incubation programme.

3 Research Methodology and Results

3.1 Methodology

The aim of this study is to systematically analyse how AI-supported decision-making is addressed in preincubation/entrepreneurship/venture development and HE literature and to design requirements for an AI-supported conceptual framework for decision-making in university pre-incubation programmes.

To achieve this objective, the literature was analysed from three complementary analytical perspectives: functional, process and technology-architectural. Together, these perspectives provided the analytical foundation for deriving the design requirements of the proposed AI-supported decision-making framework.

The study is organized around the following research questions (RQ):

RQ1. Which entrepreneurial decision-making activities are currently supported by AI in entrepreneurship and HE literature?

RQ2. Which stages of entrepreneurial development are most frequently addressed by AI-based approaches?

RQ3.1. What AI approaches are used to support entrepreneurship and decision-making?

RQ3.2. What architectural components are described in existing AI-supported entrepreneurship systems?

In this study, architectural components refer to the conceptual functional building blocks of AI-supported entrepreneurship systems (e.g., AI services, knowledge resources, data sources, end-users applications) that enable decision support.

To answer these questions, a **scoping literature review (ScR)** was conducted using the **Scopus** and **Web of Science** databases. A scoping review was selected because it is particularly suitable for mapping emerging research fields, clarifying conceptual boundaries, identifying research trends and synthesising heterogeneous evidence without restricting the review to a specific study design (Peters et al., 2015). The review focused on publications published between 2024 and 2026 to capture the period of rapid expansion in generative AI research. Following the public release of ChatGPT in late 2022, the literature experienced an exponential increase from 2024 onwards, reflecting the growing adoption of generative AI and large language models in HE (Dai et al., 2026).

The search strategy combined keywords related to **pre-incubation/entrepreneurship/venture development, HE, decision support/decision-making and AI**. The final search query used in Scopus and Web of Science was: query: TITLE-ABS-KEY (("pre-incubation" OR preincubation OR "early-stage venture*" OR "early stage venture*" OR venture* OR entrepreneurship) AND ("decision support" OR "decision support system*" OR DSS OR "decision making" OR "decision-making") AND ("higher education" OR universit* OR college*) AND ("artificial intelligence" OR AI OR "intelligent system*")) AND PUBYEAR > 2023 AND PUBYEAR < 2027

The broader search terms "artificial intelligence" and "AI" were intentionally used to ensure comprehensive retrieval of relevant literature. More specific AI approaches were subsequently identified and analysed during the full-text coding process.

The database search identified **296 records**, including **39 publications from Scopus** and **257 publications from Web of Science**. After removing **7 duplicate records**, **289 papers** remained for title and abstract screening. During the initial screening phase, each paper was assessed based on four predefined screening questions (SQ) - inclusion criteria:

SQ1. Does the study address entrepreneurship or venture development?
SQ2. Does the study address decision making or decision support?
SQ3. Does the study focus on development of entrepreneurial ideas within or connected to higher education institutions?
SQ4. Does the study involve AI?

Each criterion was independently coded as Yes, Partially or No. Only papers receiving Yes for all four inclusion criteria were included in the subsequent full-text review. This screening excluded 269 papers, resulting in 20 studies for qualitative analysis. The study selection process was conducted and reported in accordance with the PRISMA 2020 reporting guidelines (Page et al., 2021). Following the PRISMA workflow, records were identified, screened against the predefined inclusion criteria, assessed for full-text eligibility and finally included in the qualitative synthesis, with the number of retained and excluded records documented at each stage, as illustrated in Fig. 1.

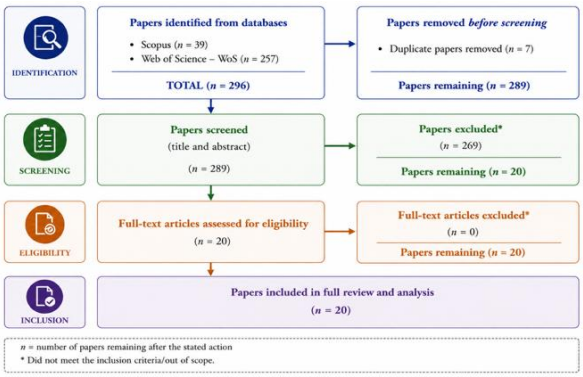


Figure 1. PRISMA Flow Diagram of the Study Selection Process

In the second phase, the **20 selected papers** were analysed in full. Data extraction focused on identifying the characteristics required to answer the research questions. Specifically, each study was coded according to analytical questions (AQ):
AQ1. What stage of entrepreneurship is addressed?
AQ2. Which decision-making domains are addressed?
AQ3.1 Which AI technologies and approaches are used?
AQ3.2 Does the paper describe a system architecture? If yes, which architectural components are mentioned?, and **its contribution to framework development**.

Finally, the extracted evidence was synthesised according to the three analytical perspectives. The resulting findings guided the development of the proposed AI-supported conceptual framework.

3.2 Scoping Literature Review Results

The qualitative analysis of the 20 selected studies identified recurring patterns across the three analytical perspectives defined in this study: functional, process and technology-architectural. In addition to addressing the RQs, each study was examined for complementary design characteristics relevant to the development of an AI-supported decision making in pre-incubation framework, including stakeholder roles, human oversight and governance aspects. **Table 1** summarises the main findings of the scoping review and illustrates how they guided the development of the proposed conceptual framework.

The final corpus consisted of 20 recent studies published between 2024 and 2026: six papers from 2024, seven from 2025 and seven from 2026. All included studies met the four first-phase inclusion criteria by connecting entrepreneurship, startup development or pre-incubation with decision making or decision support, HE and AI. The analysed studies were heterogeneous in scope and design. Some papers presented empirical studies of students, teachers, entrepreneurs or expert perceptions, some proposed conceptual or regulatory frameworks, and others described AI-enabled systems, platforms, simulations or recommendation mechanisms. Substantively, the corpus covered entrepreneurship education, competency development, entrepreneurial readiness, opportunity recognition, idea and business-model development, venture evaluation, risk and feasibility assessment, resource-allocation decisions, responsible AI and evaluator matching. The AI approaches discussed or applied across the corpus included generative AI and LLMs, machine and deep learning, predictive analytics, recommendation systems, knowledge graphs, learning analytics and simulation-based approaches. The diversity of the analysed studies required qualitative coding to systematically identify common patterns across entrepreneurial stages, decision domains, AI approaches and architectural components.

Despite this diversity, a consistent pattern emerged across the analysed studies. Existing research predominantly addresses isolated entrepreneurial activities or task-specific AI applications, whereas AI-supported decision-making across the complete pre-incubation lifecycle has received relatively limited attention. This identified gap represents the principal motivation for the proposed conceptual framework.

3.3 Conceptual Framework for AI-supported Decision-Making in Pre-incubation

The proposed conceptual framework (Fig. 2) was developed by synthesising insights from the background literature and the findings of the scoping review. Rather than directly translating the frequency of individual findings into framework elements, the review findings were conceptually synthesised. Frequently recurring concepts informed the core

framework dimensions, whereas less frequent but conceptually important findings were incorporated into broader functional or architectural components.

The framework was further informed by the authors' research and development experience in entrepreneurship support, learning design, learning analytics, AI-supported educational technologies and digital solutions for higher education.

Together, these inputs shaped the three complementary and interdependent dimensions of the framework: the Process Dimension, defining when entrepreneurial decisions occur; the Functional Dimension, specifying what decisions require support at each stage; and the Technology-Architectural Dimension, determining how AI capabilities, data resources, knowledge resources and end-user applications provide this support. Stakeholder roles and the Human-in-the-loop principle were incorporated as cross-cutting governance mechanisms spanning all three dimensions, ensuring that AI-generated insights are interpreted, validated and translated into final entrepreneurial decisions.

Design principle. AI is designed to augment rather than replace human judgement. Students, teachers, mentors, programme managers, evaluators and industry experts remain responsible for interpreting AI outputs, validating evidence, contextualising recommendations and making final decisions. This principle reflects the educational and entrepreneurial nature of pre-incubation, where AI supports evidence-based decision-making while preserving human responsibility, transparency and ethical oversight (Busch et al., 2025; Kadoić et al., 2026; Yordudom et al., 2026).

Process Dimension. The process dimension integrates the entrepreneurial stages identified across the reviewed studies with the process logic described in the foundational pre-incubation literature. Although the search strategy explicitly included the term pre-incubation, none of the selected studies conceptualised pre-incubation as a distinct AI-supported decision-making process. Instead, the reviewed literature predominantly addresses entrepreneurship education, entrepreneurial readiness, opportunity recognition, idea development, business-model development, validation and early venture development as separate educational or entrepreneurial activities (Park, 2025; Shi et al., 2026; Yordudom et al., 2026; Zirek, 2024).

Functional Dimension. The functional dimension synthesises the entrepreneurial decision domains identified across the reviewed studies. The literature predominantly addresses entrepreneurial readiness, opportunity recognition and evaluation, idea and business-model development, validation and risk- and feasibility-related decisions (Chen, 2026; Park, 2025; Rong, 2024; Sánchez Gonzales et al., 2026; Yordudom et al., 2026; Zirek, 2024). Building on these findings, the framework structures entrepreneurial decision-making into eight interconnected decision domains spanning the pre-incubation lifecycle: readiness assessment, problem framing, opportunity evaluation, idea selection, business-model evaluation, validation decision, pitch evaluation and incubation readiness assessment. Each domain represents a distinct entrepreneurial decision point at which AI supports evidence generation, recommendations, analytics and feedback, while final decisions remain under human responsibility.

Technology-Architectural Dimension. The technology-architectural dimension synthesises the AI technologies and approaches and architectural characteristics identified in the reviewed studies. The literature demonstrates a broad range of AI technologies, including generative AI, large language models, machine learning, deep learning, predictive analytics, recommendation systems, knowledge graphs, learning analytics and simulation-based approaches (Chen, 2026; Park, 2025; Qiu, 2024; Rong, 2024; Wang & Gu., 2026; Zirek, 2024). Based on this synthesis, the Technology-Architectural Dimension comprises four complementary architectural components.

AI-supported decision-support services represent an integrated layer that supports entrepreneurial decision-making throughout the pre-incubation lifecycle. Unlike individual AI applications identified in the reviewed studies (e.g., recommendation systems, tutoring tools or predictive models), these services combine multiple AI capabilities across successive entrepreneurial decision points. They also support the diverse information and decision-support needs of different stakeholder groups, including students, mentors, teachers, evaluators and programme managers. Their value lies in supporting continuity of decision-making rather than optimising isolated tasks. Accordingly, they provide capabilities such as recommendation, prediction, matching, analytics, explanation, summarisation, content generation and simulation.

Knowledge and content resources integrate entrepreneurial knowledge, business models, methods, best practices and case studies required to contextualise AI recommendations. **Data sources** combine educational, entrepreneurial and contextual information, including learning data, student and team profiles, entrepreneurial artefacts, mentor feedback and external ecosystem data.

Table 1. Summary of the ScR findings and their contribution to the proposed conceptual framework

Analytical question	Main findings from the literature review and no of studies	Contribution to the framework
AQ1. What stage of entrepreneurship is addressed?	• entrepreneurship education or competence (17) • opportunity recognition (6) • idea development (5) • early-stage venture development or business planning (19)	Defines the **Process Layer** by identifying stages that should be supported within pre-incubation lifecycle.
AQ2. Which decision-making domains are addressed?	• education or competencies decisions (16) • readiness decisions (8) • risk or feasibility decisions (10) • resource-allocation decisions (6) • governance decisions (7) • responsible-AI decisions (6) • evaluator-matching decisions (1) • opportunity, idea or business-model decisions (8)	Defines the **Functional Layer** by identifying the key entrepreneurial decision domains that should be supported throughout the pre-incubation process.
AQ3.1 Which AI technologies and approaches are used?	• generative AI and LLMs (10 studies) • machine/deep learning (11) • predictive analytics (9) • recommendation systems (4) • knowledge graphs (3) • learning analytics (4)	Defines the **Technology-Architectural Layer** by identifying the AI services and architectural components required to support an AI-supported decision-making framework for university pre-incubation programmes.
AQ3.2 Does the paper describe a system architecture? If yes, which architectural components are mentioned?	• explicit architectures (7) • partial/workflow architectures (5) • no technical architecture (4) Recurrent components: user applications/dashboards; role-specific interfaces; data sources; LMS, student/team and project data; knowledge/content repositories; recommendation engines; learning analytics; predictive models; simulation tools; feedback/evaluation mechanisms.	

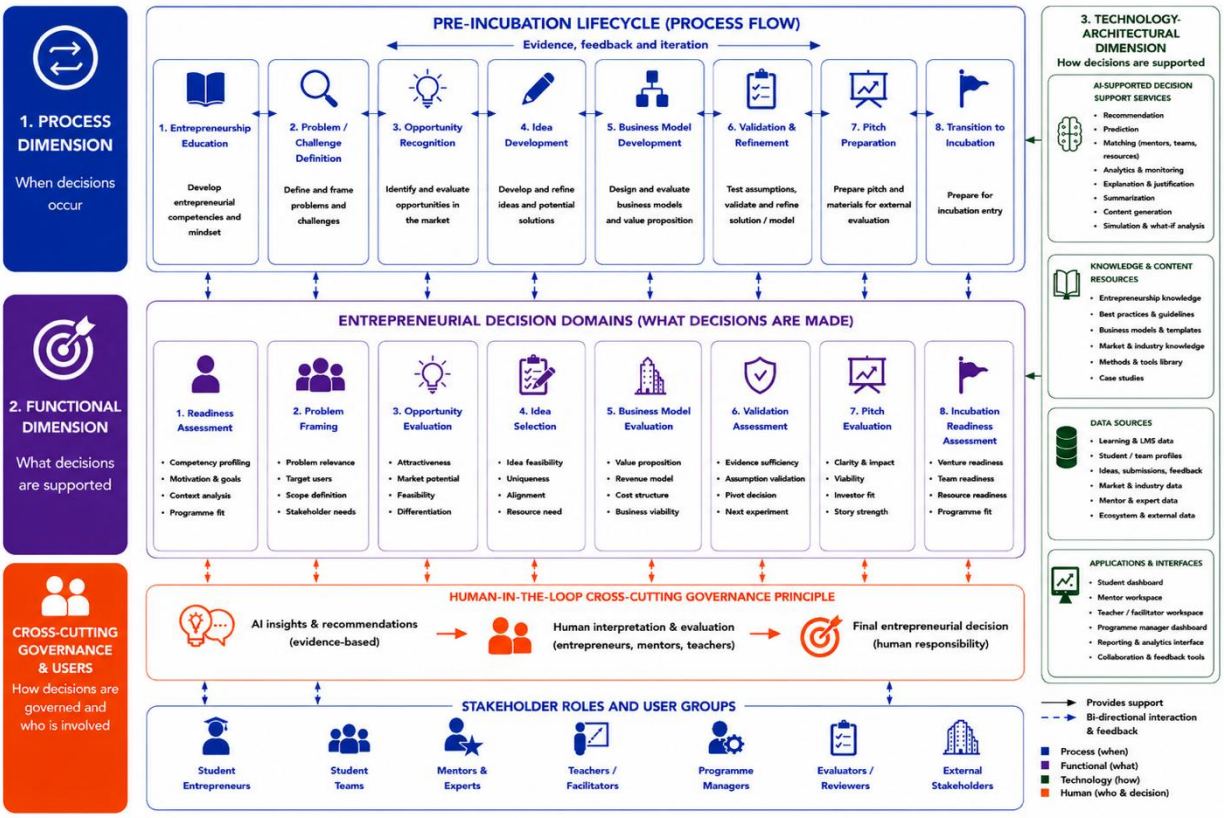


Figure 2. A three-dimensional framework for AI-Supported Decision-Making in Pre-Incubation Programmes

Finally, **applications and interfaces** deliver these capabilities to different stakeholder groups through role-specific dashboards and collaborative workspaces. Together, these components form a technology-agnostic architecture that supports entrepreneurial decision-making throughout the pre-incubation process.

User applications are reflected in student interfaces, dashboards, intelligent tutoring systems, Q&A tools, feedback mechanisms and recommendation environments (Chen, 2026; Wang & Gu, 2026; Park, 2025; Qiu, 2024).

Data sources include LMS data, student profiles, learning behaviour, project data, venture submissions, market information and evaluator or mentor data. Knowledge and content resources are represented through course materials, entrepreneurial cases, knowledge graphs, business-model templates and institutional knowledge bases (Shi et al., 2026; Chen, 2026; Wang & Gu, 2026; Xi et al., 2025; Qiu, 2024; Zirek, 2024).

User and Stakeholder Layer. The reviewed studies indicate that AI-supported entrepreneurial environments involve multiple stakeholder groups with different goals, responsibilities and information needs, including student entrepreneurs, teams, mentors, teachers, evaluators, programme managers and external ecosystem partners (Chen, 2026; Mele et al., 2022; Qiu, 2024; Wang et al., 2026; Xi et al., 2025; Zirek, 2024). Accordingly, the framework provides role-specific applications and interfaces tailored to the decision-support needs of each stakeholder group.

Human-in-the-loop as Cross-cutting Governance Principle. The reviewed studies consistently position AI as a decision-support technology that augments rather than replaces human judgement. While AI provides recommendations, predictions, analytics, matching, personalised feedback and early intervention signals, human stakeholders remain responsible for interpreting AI outputs, validating evidence, contextualising recommendations and making final decisions. This reflects the growing emphasis on transparency, ethical AI use, privacy, bias mitigation and human oversight in AI-supported entrepreneurship and education (Busch et al., 2025; Vargas, 2024; Romero-Untiveros & Melgarejo-Solis, 2024; Orozova et al., 2024; Chen, 2026; Wang & Gu, 2026; Xi et al., 2025).

4 Conclusion and Future Work

This study addressed the growing need for AI-supported decision-making in university pre-incubation programmes through a scoping literature review of recent research at the intersection of entrepreneurship, HE, decision support and AI. The review demonstrated that existing studies predominantly focus on isolated AI applications, educational support or individual entrepreneurial activities, while AI-supported decision-making throughout the pre-incubation process remains largely unexplored.

The review provides four main findings. First (RQ1)**,** the literature predominantly addresses the earliest stages of entrepreneurship, particularly entrepreneurship education, opportunity recognition, idea development and validation, while providing limited support for later stages of the pre-incubation lifecycle.

Second (RQ2)**,** AI is primarily applied to support entrepreneurial decision domains related to educational guidance, entrepreneurial readiness, opportunity recognition, idea development, business-model evaluation, validation and risk or feasibility assessment. Third (RQ3.1)**,** a broad range of complementary AI technologies and approaches has been adopted, including generative AI, large language models, machine learning, predictive analytics, recommendation systems, knowledge graphs, learning analytics and simulation. Finally (RQ3.2)**,** although several studies describe individual AI systems or partial architectures, comprehensive AI-supported architectures for university pre-incubation programmes remain largely absent.

Collectively, these findings provide the foundation for the proposed three-dimensional conceptual framework, which integrates the entrepreneurial process, entrepreneurial decision domains and technology-architectural capabilities into a unified model for university pre-incubation programmes. In addition to these three complementary dimensions, the framework incorporates stakeholder roles/user groups and the human-in-the-loop principle as cross-cutting elements spanning all dimensions, ensuring that AI augments rather than replaces human judgement.

From a theoretical perspective, the framework contributes to entrepreneurship and AI research by integrating concepts from entrepreneurial development, decision-making, AI-supported decision support and educational technologies into a coherent conceptual architecture. It extends current discussions by conceptualising pre-incubation as a continuous AI-supported entrepreneurial decision-making process rather than a collection of isolated AI applications.

From a practical perspective, the framework provides design guidance for universities, entrepreneurship educators, programme managers, mentors and system developers interested in developing AI-supported pre-incubation platforms based on the proposed framework. It identifies the key entrepreneurial decision domains, AI capabilities, data and knowledge resources, stakeholder roles and governance principles that should be considered when designing future AI-supported entrepreneurial ecosystems.

Although AI has considerable potential to support entrepreneurial decision-making, its practical implementation also introduces several challenges. The quality and reliability of AI recommendations

depend on the availability of accurate, complete and contextually relevant data, while biased or incomplete information may affect the quality of generated outputs. Furthermore, excessive reliance on AI recommendations may reduce critical reflection or entrepreneurial autonomy if recommendations are accepted without appropriate human evaluation. Consequently, AI-supported pre-incubation environments require appropriate governance mechanisms addressing data quality, transparency, ethical use, accountability and continuous human oversight to ensure that AI augments rather than replaces entrepreneurial judgement.

The study has two main limitations. First, the review was restricted to publications indexed in Scopus and Web of Science published between 2024 and 2026.

Second, the proposed framework represents a conceptual design derived from literature synthesis and the authors' research experience. Consequently, its usefulness and completeness have not yet been validated in authentic university pre-incubation settings.

Future work will therefore focus on implementing the framework within a real pre-incubation programme and evaluating its usefulness, usability and educational impact through studies.

Further studies should also examine how AI can provide context-aware, explainable and personalised decision support throughout the entrepreneurial idea development process.

Acknowledgments

This work was supported by the European Union through the NextGenerationEU programme, under the call for Institutional Research Projects financed from source 581 – Recovery and Resilience Facility (2025), within the project StartUp.Score: Development of a Model for Deciding on the Readiness of an Entrepreneurial Venture for Incubation.

References

**included in ScR*

Bielicki, M. (2023). How do we choose whom to support? Different approaches to selection criteria in business pre-incubation. Journal of Enterprising Communities: People and Places in the Global Economy, 19. https://doi.org/10.1108/JEC-09-2022-0133

Bielicki, M., & Weinert, A. (2021). Corporate pre-incubator: New platform for university-business collaboration. In F. Matos, I. Salavisa, Á. Rosa, & M. d. F. Ferreiro (Eds.), Proceedings of the 16th European Conference on Innovation and Entrepreneurship: A Virtual Conference. https://doi.org/10.34190/EIE.21.104

*Busch, R., Busch, M., May, C., & Nefen, E. (2025). Innovation and entrepreneurship within health education: International perspectives on multidisciplinary collaboration and digital skills. In Generative AI in international higher education: Challenges, applications and future directions (pp. 93–111). Taylor & Francis. https://doi.org/10.4324/9781003568650-9

Camuffo, A., Gambardella, A., Messinese, D., Novelli, E., Paolucci, E., & Spina, C. (2024). A scientific approach to entrepreneurial decision-making: Large-scale replication and extension. Strategic Management Journal, 45(6), 1209–1237. https://doi.org/10.1002/smj.3580

*Chen, X. (2026). Construction of an adaptive recommendation system for innovation and entrepreneurship courses oriented to artificial intelligence. In G. N. Nguyen, A. Swaroop, & P. Shukla (Eds.), Proceedings of the Fifth International Conference on Computing and Communication Networks (ICCCN 2025) (Lecture Notes in Networks and Systems, Vol. 1836). Springer. https://doi.org/10.1007/978-3-032-18907-3_39

Chen, M. (2025). Double Framed HyperSoft Set for Studies Factors That Influence and Ways to Improve Vocational College Instruction in Innovation and Entrepreneurship. Neutrosophic Sets and Systems. 85. 10.5281/zenodo.15265524.

Dai, K., Liu, Y., Zhang, X. (2026). Generative AI in higher education: A bibliometric review of emerging trends, power dynamics, and global research landscapes, Computers and Education: Artificial Intelligence, Volume 10, 2026, 100544, ISSN 2666-920X, https://doi.org/10.1016/j.caeai.2026.100544.

*Duong, C. D., Nguyen, T. H., Nguyen, M. H., Dang, N. S., Vu, A. T., & Do, N. D. (2025). Exploring the role of generative artificial intelligence (ChatGPT) adoption in digital social entrepreneurship: A serial mediation model. Social Enterprise Journal, 21(5), 910–936. https://doi.org/10.1108/SEJ-03-2024-0029

Jones, P., Voisey, P., & Thomas, B. (2013). The pre-incubator: A longitudinal study of 10 years of university pre-incubation in Wales. Industry & Higher Education, 27. https://doi.org/10.5367/ihe.2013.0168

Kadoić, N., Gusić Munđar, J., & Jagačić, T. (2026). Application of generative artificial intelligence for innovative teaching. Applied Sciences, 16(8), 3699. https://doi.org/10.3390/app16083699

Kalisty, K. (2023). The concept pre-incubation in theory and practice. Central European Review of

Economics & Finance, 42, 39–49. https://doi.org/10.24136/ceref.2023.003

Kazhenov, S. (2023). University business incubators and opportunities for collaboration with companies within the start-up 2. Global Journal of Business, Economics, and Management: Current Issues, 13(1), 106–114. https://doi.org/10.18844/gjbem.v13i1.8452

*Kurtovic, A., & Soric, K. (2025). AI and simulation in entrepreneurial education: Lessons from the RITup Ventures program. In N. B. Redep & M. Bubanic (Eds.), Proceedings of the 36th Central European Conference on Information and Intelligent Systems (CECIIS 2025) (pp. 269–276). University of Zagreb.

*Liu, Q., An, X., & Chen, W. (2025). College students' entrepreneurship education path and management strategy of start-up enterprises using causal attribution theory. Scientific Reports, 15. https://doi.org/10.1038/s41598-025-86797-z

*Margaretha, W., Ariesia, K., & Yustian, O. R. (2025). Artificial intelligence and entrepreneurship education: Increasing entrepreneurial intentions among future entrepreneurs in Bandung City. Journal of Theoretical and Applied Information Technology, 103(7). https://doi.org/10.5281/zenodo.17187749

Mele, G., Sansone, G., Secundo, G., & Paolucci, E. (2022). Speeding up student entrepreneurship: The role of university business idea incubators. IEEE Transactions on Engineering Management. https://doi.org/10.1109/TEM.2022.3175655

*Orozova, D., Angelova, N., & Zlatev, Z. (2024). New skills and knowledge for digital entrepreneurs in the age of artificial intelligence. European Conference on Innovation and Entrepreneurship, 19, 606–610. https://doi.org/10.34190/ecie.19.1.2447

Page, M. J., McKenzie, J., Bossuyt, P., Boutron, I., Hoffmann, T., Mulrow, C., Shamseer, L., Tetzlaff, J., Akl, E., Brennan, S., Chou, R., Glanville, J., Grimshaw, J., Hróbjartsson, A., Lalu, M., Li, T., Loder, E., Mayo-Wilson, E., McDonald, S., & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. Journal of Clinical Epidemiology, 134, 178–189. https://doi.org/10.1016/j.jclinepi.2021.03.001

*Park, T. (2025). The role of AI platforms in enhancing entrepreneurship education: A theoretical analysis. https://doi.org/10.1007/978-3-032-06982-5_8

Peters, M., Godfrey, C., Khalil, H., McInerney, P., Parker, D., & Soares, C. (2015). Guidance for conducting systematic scoping reviews. International Journal of Evidence-Based Healthcare, 13. https://doi.org/10.1097/XEB.0000000000000050

Popov, V. (2022). Shift in paradigm: Startup business-incubation at universities during the educational discipline inside of curriculum rather than extracurricular activities of students. Scientific Journal of Astana IT University, 48–60. https://doi.org/10.37943/WDNE3944

*Qiu, X. (2024). Construction of innovation and entrepreneurship information sharing system based on artificial intelligence. 368–373. https://doi.org/10.1109/AIARS63200.2024.00075

*Rahal, F., Farmanesh, P., Houmani, H., & Dehkordi, N. (2026). Does Gen-AI enhance the link between entrepreneurship education and student innovation behavior? Insights for quality and sustainable higher education. Sustainability, 18, 1258. https://doi.org/10.3390/su18031258

Ries, E. (2011). The lean startup: How today's entrepreneurs use continuous innovation to create radically successful businesses. Currency.

*Romero-Untiveros, L., & Melgarejo Solis, R. (2024). Regulation for the use of artificial intelligence in systems engineering education to foster educational innovation and entrepreneurship. https://doi.org/10.18687/LEIRD2024.1.1.782

*Rong, C. (2024). An AI-driven decision framework for promoting sustainable entrepreneurship in vocational colleges. Decision Making: Applications in Management and Engineering, 7(2), 748–769. https://doi.org/10.31181/dmame7220251481

*Sánchez Gonzales, H., Cartes-Barroso, M., & Ftah Ftah, K. (2026). Artificial intelligence as a strategic resource for journalism entrepreneurship: Perceptions of the entrepreneurial mindset and decision-making. Revista de Comunicación, 25, 363–388. https://doi.org/10.26441/RC25.1-2026-4347

*Shi, F., Peng, Q., & Yang, A. (2026). AI-empowered innovation and entrepreneurship education in universities: A development path and ecosystem framework. 211–216. https://doi.org/10.1145/3797552.3797588

*Skica, T., Moczek, T., & Döngül, E. S. (2026). Entrepreneurial competencies and the conceptual dimension of intentions: Insights from a hybrid machine learning approach. Journal of Entrepreneurship, Management and Innovation, 22(1), 36–56. https://doi.org/10.7341/20262213

Taherdoost, H., & Madanchian, M. (2023). Decision making: Models, processes, techniques. Cloud Computing and Data Science, 1–14. https://doi.org/10.37256/ccds.5120233284

*Yordudom, T., Imjai, N., Sangthong, M., Vongchavalitkul, B., Ru-Zhue, J., & Aujirapongpan, S. (2026). Strengthening strategizing and startup entrepreneurial readiness among Thai Gen Z management students: Do AI literacy, creativity, and seizing skills matter? The International Journal of Management Education, 24(2), 101416. https://doi.org/10.1016/j.ijme.2026.101416

*Vargas, M. (2024). Integración de la inteligencia artificial en decisiones curriculares: Análisis del plan curricular de la asignatura: Empresarialidad (Código 4121) de la Escuela de Ciencias de la Administración de la UNED, Costa Rica. REDU: Revista de Docencia Universitaria, 22(2).

*Wang, Z., & Gu, Y. (2026). Integrating entrepreneurship education into traditional curricula to promote economic growth and job creation based on deep learning. International Journal of e-Collaboration, 22(1). https://doi.org/10.4018/IJeC.407058

Wright, M., Siegel, D. S., & Mustar, P. (2017). An emerging ecosystem for student start-ups. The Journal of Technology Transfer, 42. https://doi.org/10.1007/s10961-017-9558-z

*Xi, S., Pi, O., Zhang, M., Xiong, B., Lane, J., & Shah, N. (2025). Who is a better matchmaker? Human vs. algorithmic judge assignment in a high-stakes startup competition. https://doi.org/10.48550/arXiv.2510.12692

*Zirek, M. (2024). A novel approach to university startup incubators by using AI powered simulation and gamification. In 2024 International Conference on Computing, Networking, Telecommunications & Engineering Sciences Applications (CoNTESA) (pp. 1–6). https://doi.org/10.1109/CoNTESA64738.2024.10891285

The Next Wave of Intelligent Systems

Christina Nilsson
When AI Targets Core Tasks: Human Actors and Role Repositioning in Intelligent Documentation Systems

Zahra Bibi, Nida Fatima, Rubina Khadim, Iftakhar Rasheed
Privacy Preserving Federated Learning withDifferential Privacy for DDoS Intrusion Detectionin Smart City IoT Networks

Lovro Posarić, Tara Nikolić
Towards Autonomous Resilience: Architecting Agentic Middleware for the Cloud-Edge Continuum

Štefan Horvat
Evaluating event influence on TGN performance in bike sharing station occupancy forecasting

Dino Vlahek, Sandra Lovrenčić, Dijana Oreški, Nina Begičević Ređep, Tena Jagačić, Renato Punčec, Hrvoje Aničić, Marko Mrkonjić
From industrial data to business decisions: Industrial case studies as teaching instruments in interdisciplinary analytics education

Hana Ivančić, Tomislav Jagušt
An AI Companion for Self-Regulated Learning: Integrating Conversational AI, Predictive Analytics, and Explainability

Claudia-Maria Miu, Delia-Elena Vranceanu, Andreea-Cristina Stoica
Artificial Intelligence Disclosure and Magical Perception:Effects on Consumer Trust in Chatbot Interactions

When AI Targets Core Tasks: Medical Secretaries and Role Repositioning in AI-Driven Documentation Systems

Christina Nilsson
University of Gothenburg
Department of Applied IT
Box 470, SE-405 30 Gothenburg, Sweden
christina.nilsson@ait.gu.se

Abstract. *AI-driven documentation systems such as ambient scribes are marketed with claims to perform clinical documentation. In many healthcare settings this work belongs to an established profession, the medical secretary. Drawing on 26 interviews across six healthcare regions, this interpretive qualitative study examines how medical secretaries respond to such claims. A replacement narrative first took hold but gave way as the secretaries learned more and engaged with the technology. They saw its limits, brought their expertise forward, and began to imagine a repositioned role, still relevant, focused on oversight and quality assurance. The article contributes the concept of anticipatory occupational sensemaking.*

Keywords. Human-AI collaboration, ambient scribes, automation, healthcare, medical secretaries

1 Introduction

AI systems are increasingly entering professional domains where core tasks have long depended on human judgement, situated knowledge, and occupational expertise. Ambient scribes offer a clear example of this development. Built on large language models, they capture clinical conversations in real time and generate structured medical records without manual transcription (Palm et al., 2025). In healthcare they are being introduced as an efficiency solution (Sanchez et al., 2026; Tierney et al., 2024), but the work they are entering is more complex than that framing suggests, and their consequences extend beyond the physicians they are designed to support.

In settings where doctors manage their own documentation, ambient scribes and advanced speech recognition tools promise a welcome reduction in workload. "*Medical notes that write themselves*" reads a tagline from a Swedish provider of ambient scribes (Tandem Health, n.d.).

However, clinical documentation is not organised this way everywhere. In Sweden, clinical documentation is supported by a distinct and formalised documentation profession, "medicinsk vårdadministratör" or "medicinsk sekreterare", translated here as medical secretary. Approximately 20,000 people are employed in this occupational category (Swedish National Agency for Higher Vocational Education, 2025) and many handle physicians' dictations across a broad range of healthcare settings, supported by their specialist expertise in medical terminology, clinical coding, and patient-facing readability.

For medical secretaries, therefore, the introduction of AI-driven documentation tools carries a fundamentally different meaning. The technology is positioned as capable of performing core tasks that have historically formed the backbone of their professional role, raising questions about what happens to professional roles when AI systems are introduced with claims to perform core occupational work.

This study focuses on this second documentation model at a critical moment, when organisations still have the opportunity to shape the transition before roles and responsibilities become stabilised.

What medical secretaries are facing is not unique. Learning algorithms increasingly enter work domains by producing responses, classifications, and summaries that resemble knowledge work with consequences for expertise, occupational boundaries and coordination (Faraj et al., 2018). Concerns of this kind have a long lineage in research on automation and deskilling (Thompson, 2024), where a central argument has been that the reorganisation of work around new technology can leave an occupational role with less skilled content and less room for professional judgement (Braverman, 1974).

This study examines the case of medical secretaries as an instance of this broader development and asks: *How do medical secretaries make sense of AI-driven documentation systems when these systems are positioned as capable of performing core parts of their work?*

The Swedish case of medical secretaries offers an unusually clear empirical window into a broader IS concern. It illustrates how human actors make sense of and respond to agentic IS artefacts introduced with

claims to perform core occupational work (Baird & Maruping, 2021; Faraj et al., 2018). Baird and Maruping (2021) argue that increasingly agentic IS artefacts require attention to how rights and responsibilities for task execution and outcomes are redistributed between humans and information systems. In this study, that concern is examined from the position of an affected occupational group in an early implementation phase. The case should therefore be read as *one* observable trajectory within a wider transformation in which AI systems enter cognitive work previously organised around human judgement, situated expertise, and established divisions of labour.

2 AI Documentation as Socio-technical Transformation

2.1 Clinical documentation as distributed, quality-critical work

Clinical documentation is often treated as a matter of producing clinical text, yet IS and health informatics research has shown that records form part of a wider organisational infrastructure. They support clinical reasoning, coordination, accountability, reimbursement, management, research, and other secondary uses of data (Berg, 1999; Bossen et al., 2019; Pine & Bossen, 2020). Documentation quality therefore depends on more than whether information from the clinical encounter has been captured. Records must also be complete, accurate, codeable, legally usable, and understandable across different organisational and patient-facing contexts (Kariotis et al., 2025; Pine & Bossen, 2020).

This broader view is important because clinical documentation is produced through situated work across roles. In many Swedish healthcare settings, physicians dictate or provide the clinical content, while medical secretaries prepare, structure, code, and quality assure the record. As introduced above, this occupation is supported by specialised higher vocational education for medical care administrators (occupational titles may vary across regions). During the period relevant to this study, MYHFS 2022:1, a national agency regulation issued by the Swedish National Agency for Higher Vocational Education, specified programme content in areas such as medical terminology, medical documentation, diagnosis classification, healthcare law, digital systems, confidentiality, and IT and e-health (Swedish National Agency for Higher Vocational Education, 2022, Appendices 1 and 2). This makes the Swedish case analytically useful: it shows a documentation arrangement in which quality, readability, coding, and accountability are already distributed before AI enters the setting.

For this reason, AI-driven documentation tools do not enter an empty space. They enter an existing division of documentation work. When such tools are framed primarily as text-generating technologies, important forms of human expertise risk becoming less visible. The question is therefore how such systems become connected to the social and organisational arrangements through which documentation quality is produced.

2.2 AI-driven documentation systems and human-AI role reconfiguration

Speech recognition systems and ambient AI scribes (hereafter ambient scribes) represent two related but distinct forms of AI-driven documentation technology. Speech recognition systems convert dictated speech into text and may include functions such as templates, structured fields, or language support. Ambient scribes go further by capturing clinician-patient conversations and generating draft clinical notes using machine learning and natural language processing (Tierney et al., 2024). They therefore shift the technical artefact from transcription support towards automated drafting, summarisation, and preliminary structuring of clinical documentation.

Recent research on these systems has mainly examined physician-centred outcomes, including documentation burden, time use, EHR efficiency, note quality, and technical performance (Oettl et al., 2026; Pearlman et al., 2025; Tierney et al., 2024). This literature helps explain why ambient scribes attract organisational interest. They promise relief from administrative workload and a more efficient clinical workflow. At the same time, emerging studies point to persistent risks. Comparative evaluations of clinical AI scribes have identified omissions and other errors in generated summaries, while studies of LLM-based medical summarisation have reported hallucinations and missing information (Asgari et al., 2025). Topaz et al. (2025) therefore warn that adoption is moving faster than validation and oversight.

These findings matter for human-AI collaboration. If AI-driven documentation systems generate drafts that require review, correction, and contextual adaptation, then implementation is not simply a question of replacing human writing with automated output. It becomes a question of how organisations configure responsibility around AI-generated documentation. This raises practical questions about who verifies the record, checks omissions, adapts language for patients, and ensures that coding, templates, and local documentation routines still work.

A socio-technical perspective is useful here because the consequences of AI documentation technologies depend on how technical capability is configured with human judgement, organisational routines, and professional responsibilities (Leonardi, 2011; Orlikowski, 1992). In healthcare, this is especially important because digital technologies affect coordination, accountability, and patient safety, not only task execution (Papoutsi et al., 2020; Sittig &

Singh, 2010). The present study therefore examines AI-driven documentation as an early socio-technical transition in which medical secretaries begin to interpret, evaluate, and reposition their role within emerging human-AI documentation arrangements.

3 Method

This study adopts an interpretive qualitative design (Klein & Myers, 1999) to examine how medical secretaries make sense of AI-driven documentation technologies in Swedish public healthcare. This design is suitable because the study focuses on an early stage of technological change, where roles and routines are still being negotiated and many consequences are not yet stabilised in everyday practice.

3.1 Research setting and participants

Swedish healthcare is a decentralised, publicly funded system in which 21 regions are responsible for medical care, delivered by both public and private providers. The study is based on empirical material from six of these regions. Participants had experience of speech recognition systems, ambient scribe demonstrations, pilot activities, or organisational discussions concerning AI-driven documentation. A separate survey of regional uptake was used to situate the interview material within broader developments across Swedish public healthcare.

The study includes 26 participants: 13 medical secretaries, 8 managers, 2 physicians, 1 project leader, and 2 vendors. Medical secretaries constituted the main participant group, while the additional interviews provided perspectives on implementation, clinical use, management, and vendor positioning. Most managers had a background as medical secretaries or equivalent documentation roles. Participants were recruited through snowball sampling, supplemented by direct contact with regional organisations. This enabled variation in how the technologies were introduced, discussed, and experienced across settings.

Participants' engagement with the technologies ranged from direct use of speech recognition systems in daily work and participation in ambient scribe pilots to familiarity based on demonstrations, internal communication, and colleagues' accounts, as summarised in Table 1. Quotations are labelled by role (MS for medical secretary, MAN for manager), and participants with hybrid backgrounds are labelled according to their current role. As snowball sampling relies on referrals, the sample may over-represent participants with an active interest in AI-driven documentation, a potential self-selection bias that should be kept in mind when interpreting the findings.

Table 1. Overview of data sources

Data source	n	Engagement and role in the study
Interviews: medical secretaries	13	Direct use, pilots, demonstrations; primary material for the thematic analysis
Interviews: managers	8	Oversight and direct use; pilots, demonstrations
Interviews: physicians	2	Direct use, pilots
Interviews: project leader	1	Head of implementation
Interviews: vendors	2	Product demonstrations and development
Regional uptake survey	17 of 21	Two questions on uptake of speech recognition and ambient scribes (Nov–Dec 2025); contextual mapping
Pilot documentation	3	Two evaluation reports, one slide deck (2024–2025); context and triangulation
Public web material	–	Vendor and regional communication; terminology and positioning

The 24 interviews with regionally based participants were distributed across the six regions as follows: Region A: 8; Region B: 7; Region C: 4; Region D: 3; Region E: 1; Region F: 1. The two vendor representatives operated across regions and were not affiliated with a single region.

3.2 Data collection and analysis

Data were collected through semi-structured interviews conducted in Swedish via Microsoft Teams in 2025, each lasting 30 to 50 minutes. The interview guide covered participants' experiences of speech recognition and ambient scribe technologies, their perceptions of how these systems might affect documentation work, and their expectations for future roles. Interviews were recorded with consent, transcribed using an AI-based transcription service, and checked against the recordings. The study does not fall under the approval requirement of the Swedish Ethical Review Act (SFS 2003:460), as it involves neither sensitive personal data nor any physical intervention or influence on participants. This research adheres to the standards stipulated by the European Code of Conduct for Research Integrity (ALLEA, 2023) and the Swedish Research Council's guidelines for good research practice (Vetenskapsrådet, 2024). All participants were informed about the purpose of the study and the handling of their data, that participation was voluntary, and that they could withdraw at any stage. Participants were anonymised through role-based labels and personal identifiers were removed from the transcripts. Interviews were recorded within

institutional Microsoft Teams environments; recordings were automatically deleted after transcription and verification in accordance with the applicable retention policies, and anonymised transcripts were retained for analysis. Interview quotations were translated from Swedish into English by the author and checked against the original transcripts.

The analysis followed a reflexive thematic approach (Braun & Clarke, 2021). All transcripts were imported into Atlas.ti, where coding was initially inductive and focused on how participants described the technologies, their own expertise, and the future of documentation work. In a subsequent interpretive phase, codes were iteratively grouped and regrouped into themes around replacement-oriented narratives, visible documentation expertise, and role repositioning towards oversight and human-AI collaboration, with earlier coding decisions revisited as the interpretation developed. The interviews were conducted and analysed by the author, whose position as an informed outsider, familiar with the organisational and digital context of Swedish public healthcare but without a background in the profession studied, provided contextual understanding while preserving analytical distance. The main themes were derived from the interviews with medical secretaries, while the interviews with managers, physicians, vendors, and the project leader were used to contextualise implementation, clinical use, management, and vendor positioning rather than to generate themes.

Contextual material was used to situate the findings but was not analysed with the same depth as the interviews. The regional uptake survey, conducted in November–December 2025, comprised two questions on the use of speech recognition systems and ambient scribes and was answered by 17 of Sweden's 21 regions. The responses showed an uneven and early-stage national landscape: six of the 17 responding regions reported broad use of AI-supported speech recognition while seven reported none (several noting that they used speech recognition without AI support), and no region reported ambient scribe use beyond the pilot stage in its own services, although broader use among private providers was noted. The pilot documentation consisted of two formal evaluation reports and one slide deck from regional pilots conducted in 2024–2025. Together with selected public web material, these sources were used to corroborate uptake and implementation status and to clarify how the systems were positioned in official communication, not to generate additional themes.

4 Findings

The findings are organised around three connected movements in how medical secretaries made sense of AI-driven documentation technologies. First, the technologies appeared through replacement-oriented narratives. Second, medical secretaries evaluated the systems against the practical demands of documentation work and made their own expertise visible. Third, they began to reimagine their role around oversight, quality assurance, and human-AI collaboration.

These three movements represent analytical patterns identified across the material rather than stages through which individual participants necessarily progressed. Participants moved between these positions, and several remained primarily concerned about replacement throughout. The movements should therefore be read as coexisting orientations in the material rather than as a uniform sequence.

4.1 AI first appeared through replacement-oriented narratives

Medical secretaries first encountered AI-driven documentation technologies through claims that the systems could take over tasks long associated with the core of their work. Speech recognition systems and ambient scribes were therefore not interpreted as neutral administrative tools. They were understood as technologies directed at writing, transcription, and documentation processes that had historically defined the medical secretary role.

This interpretation was reinforced by how the technologies were presented. One vendor tagline promised "Medical notes that write themselves" (Tandem Health, n.d.). One respondent recalled how a similar idea circulated around speech recognition: that no one should have to write anything and that medical secretaries would no longer be needed. She then laughed and added, "then some time passed…" (MAN-7). The remark captures the initial force of the replacement narrative, but also its later weakening as participants became more familiar with what the technologies could actually do.

Organisational signals made this interpretation more acute. One medical secretary described how the timing of AI discussions coincided with workforce uncertainty: "There was a redundancy notice and a hiring freeze at the same time. It made us think that management believed this [AI] would take over much of the work" (MS-9). Another participant stated the concern more directly: "Of course, you were worried that we might not have jobs left, that we would somehow be completely replaced" (MS-11).

In this first phase, AI was mainly encountered through a language of substitution. The concern was therefore shaped less by stable use of the systems and more by the expectations attached to them. Ambient scribes and speech recognition systems appeared as technologies that could change who was needed in documentation work, and on what terms.

4.2 Medical secretaries made documentation expertise visible

As participants engaged more closely with the technologies through demonstrations, pilots, colleagues' accounts, and direct observation of outputs, their assessment became more specific. The perceived threat did not disappear. Several participants still expected AI to reduce parts of the work and perhaps reduce the number of medical secretaries over time. Yet their evaluation increasingly focused on the gap between AI-generated documentation and what high-quality documentation requires in practice.

Readability was one area where this expertise became visible. Participants described how documentation must often be reshaped so that it works for patients, clinicians, and organisational use. One medical secretary explained: "We act as a filter between the clinician and the patient. We adapt the record so it becomes readable. Some doctors are very good, but often it would be unreadable if we just wrote things exactly as they are said" (MS-7). Another described the work of clarifying local or individual abbreviations: "There are both established and local abbreviations, sometimes even ones used by individual physicians. I therefore often write out the abbreviation and clarify it in parentheses" (MS-6).

Participants also pointed to errors that required situated judgement. Speech recognition outputs could misrecognise names, create long blocks of text without clear structure, omit information, or alter meaning. One respondent described how a clinically important distinction could be lost when "right" was rendered incorrectly. Another warned that if AI-generated text went directly into the record, "no one might notice" errors that should have been corrected (MS-3). These accounts show that participants evaluated AI documentation tools against the practical conditions of safe and usable documentation, not against text production alone.

Coding was another area where documentation expertise became visible. Participants described coding as closely connected to reporting, follow up, and financial flows in healthcare organisations. One respondent noted that "Documentation and coding are closely tied to reporting and how organisational performance is monitored, which in turn affects resources and finances" (MS-5). Another emphasised that coding is part of the profession's formal expertise: "Coding is an integral part of the medical secretary profession. During our training, we take courses in coding and in how to think about it. There is a strong focus on this, and there is a sense of pride in the fact that this responsibility rests with us" (MS-6). The stakes could be substantial: "If you miss a code, it can mean a difference of 30,000 kronor on an invoice" (MS-6).

This phase therefore revealed more than technical shortcomings. It showed how medical secretaries used practice-based expertise to evaluate the limits of AI in relation to documentation as it is actually done. Their accounts made visible forms of work that had been obscured in the earlier replacement-oriented narratives: interpretation, structuring, coding, readability, quality assurance, and accountability. The material therefore points to a qualified position. AI was seen as likely to affect the role, but important parts of documentation work were still understood as dependent on human judgement, adaptation, and responsibility.

4.3 The role was reimagined around oversight and human-AI collaboration

As the initial replacement narrative lost some of its force, many participants began to describe a different future role for medical secretaries. This did not mean that uncertainty disappeared. Several still expected writing tasks to decrease and some believed that fewer medical secretaries might be needed in the long run. Alongside this concern, however, the material shows a forward-looking movement in which the role was reimagined around oversight, coordination, quality assurance, and implementation support. This orientation coexisted with continued uncertainty. As one participant, quoted earlier on the fear of being completely replaced, concluded: "I think we are still needed, but perhaps not as many of us" (MS-11).

Participants repeatedly described a future role centred less on producing text and more on reviewing, controlling, and safeguarding documentation processes. One participant summarised this shift: "I do not think the profession will disappear, but that it will become more closely connected to operations and processes, with more control and less monotonous writing" (MS-8). Relief from transcription was also linked to freeing capacity for other tasks: "If we can get some relief, we can do other things better" (MS-8). AI was therefore interpreted as something that could move the human role towards broader oversight and contextual judgement.

This future role was closely linked to collaboration. Participants stressed that AI-generated output would still require human review and that clinicians should not be left to carry the entire documentation burden. One respondent put it clearly: "We cannot put everything on the doctors … they need to be doctors and do what they are good at, and we take care of the rest" (MS-1). The future envisioned here was a more differentiated division of labour in which AI, clinicians, and medical secretaries contribute in different ways.

Participants also connected this role shift to implementation and governance. Several argued that medical secretaries should be involved early when AI documentation tools are introduced, configured, and evaluated. One respondent stated: "I think it is important to include medical secretary expertise in the governance of AI. When is AI appropriate, and when is human involvement needed? Patient records are about patient safety" (MS-1). Another participant

stressed that medical secretaries should help lead implementation work so that documentation knowledge, legal requirements, and regional guidelines are built into how the systems are introduced.

The findings therefore show a shift from acute replacement anxiety towards a more qualified and proactive position. Medical secretaries did not simply defend an existing role. They began to articulate how their expertise could be used in a documentation process increasingly shaped by AI. In this sense, the findings point to an emerging model of human-AI collaboration in which the role of the medical secretary is not removed, but reconfigured.

5 Discussion

5.1 From affected employees to active transformation actors

The study shows that role repositioning can begin long before AI systems are fully implemented in everyday work. In this case, the movement was triggered by replacement-oriented narratives, vendor promises, organisational signals, and uncertainty about the future relevance of the medical secretary role.

However, the response that followed was not limited to anxiety or resistance. As medical secretaries encountered the technologies through demonstrations, pilot activities, colleagues' accounts, and organisational discussions, they began to assess what the systems could and could not do in relation to documentation work as they knew it. Their concern about replacement became a starting point for a more active form of sensemaking. They began to identify where human judgement remained necessary, where AI-generated output would need control, and how their own expertise might remain relevant in a documentation process increasingly shaped by AI systems.

This suggests that affected employees may themselves carry out part of the preparatory transformation work that organisations need, but may not yet have formally organised. By interpreting the technology, testing its implications against everyday work, and imagining possible future divisions of labour, medical secretaries began to translate an uncertain AI-driven change into more concrete organisational questions. This does not remove the need for organisational direction. It does, however, indicate that early employee sensemaking can provide a resource for shaping the transition before roles and routines have stabilised.

This matters for research on digital transformation because it shifts attention from affected employees as recipients of change to affected employees as possible contributors to role and practice transformation. Medical secretaries were not simply reacting to an already defined transformation. They were beginning to work on the transformation itself by exploring possible future scenarios, articulating alternative divisions of labour, and repositioning their expertise in relation to AI. Their response therefore points to an important source of agency in early AI implementation: the people whose work is most directly challenged may also actively be trying to make sense of how the organisation could move forward.

The findings do not support a general claim that all AI-affected groups will reposition themselves in the same way. However, the findings identify one possible trajectory. The broader contribution is to make this trajectory visible as a phenomenon that future IS research should examine across other domains where AI targets established cognitive work.

5.2 Situated expertise as a resource for human-AI collaboration

The case also shows that human-AI collaboration cannot be designed only from a general understanding of what the technology can automate. It depends on situated knowledge of the work into which the technology enters. In clinical documentation, this includes not only drafting text, but also recognising omissions, clarifying language, adapting records for patients, managing coding requirements, and understanding how documentation travels through clinical, administrative, legal, and organisational contexts. The findings suggest that medical secretaries are preparing for a transformation of their role, shifting from text producers to active agents within the AI system's lifecycle. They were beginning to contribute to how the transformation could be understood and organised in practice, where human oversight becomes a prerequisite for bridging the gap between the AI system's technical output and the actual requirements for clinical usability.

Medical secretaries' contribution lies in this documentation expertise. Their assessments of AI were grounded in knowledge of what a usable and accountable record requires.

This allowed them to see weaknesses in a future model built only around physicians and AI, especially where responsibility for checking, structuring, coding, and adapting the record remained unclear. The issue, then, is not whether medical secretaries see the technology better than other actors. It is that they see it from a position embedded in the documentation process. Research on invisible work has long noted that no work is inherently visible or invisible; whether the judgement and coordination embedded in a role are seen is a matter of context and negotiation (Star & Strauss, 1999). The replacement narratives, by reducing documentation to transcription, put exactly this negotiation in motion, and the medical secretaries responded by making the invisible work visible. Employee involvement therefore becomes an organisational resource for configuring workable human-AI collaboration.

Involving medical secretaries early is a way of bringing documentation knowledge into the configuration of human-AI collaboration. Their emerging future role, as described in the material, centres on oversight, quality assurance, coordination, and implementation support. This suggests that affected employees can become active contributors to the design of workable AI-supported practices, particularly when the technology targets tasks that have previously anchored their occupational role.

5.3 Early implementation as a window for proactive role shaping

The early phase of implementation is therefore consequential. Demonstrations, pilot projects, vendor narratives, procurement discussions, and informal organisational signals do not merely prepare the ground for later change. They already shape how employees understand the technology, how they imagine their future role, and what kinds of human-AI arrangements become thinkable. In this case, repositioning began before ambient scribes had become routine practice. That timing is central to the study's contribution.

For organisations, this creates a window for proactive role shaping. If early sensemaking is ignored, organisations risk missing the point at which employees are already trying to work out how their expertise can remain useful in an AI-supported future. If it is recognised, the organisation can use this moment to clarify responsibilities, involve documentation experts in procurement and evaluation, define appropriate human oversight, and design roles around the work that AI output still requires. The practical implication is therefore stronger than simply involving users early. Affected employees should be treated as transformation resources, not implementation obstacles.

This is particularly relevant in public healthcare, where digital transformation takes place under staffing pressure and high accountability demands. The findings suggest that AI-supported documentation may move medical secretaries towards roles centred on human oversight, verification, quality assurance, and control of AI-generated output. Such roles keep humans in the loop, but they may not require the same number of medical secretaries to perform traditional documentation work. A proactive approach should therefore also consider how this occupational expertise could be further developed or partly reskilled for other organisational needs. Medical secretaries' knowledge of terminology, coding, confidentiality, documentation systems, and patient-facing administration may be valuable beyond the specific task of producing records. This broader use of competence lies beyond the main empirical claim of the study, but the case suggests why the question should be raised early, while roles and routines are still open to shaping.

This repositioning converges with needs identified in the clinical AI safety literature. The auditing of AI-generated content, recognition of error patterns, and quality assurance that Topaz et al. (2025) call for as safeguards correspond in part to the role medical secretaries were beginning to articulate for themselves.

Taken together, the theoretical contribution of this study is to conceptualise role repositioning as anticipatory occupational sensemaking. This denotes an adaptive response through which an affected occupational group reinterprets its expertise in relation to an agentic system before that system is embedded in everyday practice. In terms of Baird and Maruping's (2021) delegation perspective, medical secretaries were beginning to specify which rights and responsibilities could be delegated to the system and which should remain with humans, before the organisation had formally addressed that question. This positions role repositioning as a potential precondition for workable human-AI collaboration rather than its outcome: whether the repositioned role is realised depends on whether organisations recognise and act on this early sensemaking. Whereas sociotechnical accounts have located the reconfiguration of roles and routines in people's ongoing interaction with technologies in use (Leonardi, 2011; Orlikowski, 1992), this case shows that it can be set in motion by anticipated delegation, by what the technology is claimed to do, before patterns of use have stabilised. The study thereby extends research on occupational change in the age of the learning algorithm (Faraj et al., 2018), showing that boundary work over expertise can begin in this anticipatory phase.

5.4 Limitations and future research

This study examines an early phase of AI-driven documentation change. It can therefore show how expectations, concerns, and role repositioning emerge before full implementation, but it cannot determine how these arrangements will stabilise once ambient scribes and related systems become embedded in everyday work. The Swedish setting is also distinctive because medical secretaries constitute a formalised documentation profession, which may limit direct transferability to healthcare systems where documentation work is organised differently. Future research should therefore follow AI-supported documentation longitudinally and compare settings where documentation work is distributed across roles in different ways.

This also raises a need for comparative studies of how occupational groups respond when AI systems are introduced into cognitive tasks that have previously anchored human expertise. Such studies could examine under what conditions anticipatory occupational sensemaking leads to role repositioning, occupational-level deskilling, resistance, or new forms of hybrid work.

5.5 Conclusion

This study has examined how medical secretaries make sense of and reposition their role when AI-driven documentation systems target core documentation work. The findings show that the introduction of AI is not understood only as a technical change, but as a challenge to occupational identity, expertise, and future relevance. In response, medical secretaries begin to reposition themselves around oversight, quality assurance, contextual interpretation, and human-AI collaboration. The study conceptualises this response as anticipatory occupational sensemaking, offering one empirically grounded example of how human actors may renegotiate their role when AI systems enter cognitive work domains previously organised around human expertise. The study does not claim that all occupational groups will respond to AI in the same way. This points to the need for further IS research on how AI reshapes work, expertise, and professional boundaries across organisational settings.

Acknowledgments

The author thanks the interview participants for generously sharing their time and experiences.

References

ALLEA. (2023). *The European Code of Conduct for Research Integrity – Revised Edition 2023*. All European Academies. https://doi.org/10.26356/ECOC

Asgari, E., Montaña-Brown, N., Dubois, M., Khalil, S., Balloch, J., Yeung, J. A., & Pimenta, D. (2025). A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. *Npj Digital Medicine*, *8*(1), 274. https://doi.org/10.1038/s41746-025-01670-7

Baird, A., & Maruping, L. M. (2021). The Next Generation of Research on IS Use: A Theoretical Framework of Delegation to and from Agentic IS Artifacts. *MIS Quarterly*, *45*(1), 315–341. https://doi.org/10.25300/misq/2021/15882

Berg, M. (1999). Accumulating and Coordinating: Occasions for Information Technologies in Medical Work. *Computer Supported Cooperative Work (CSCW)*, *8*(4), 373–401. https://doi.org/10.1023/a:1008757115404

Bossen, C., Pine, K. H., Cabitza, F., Ellingsen, G., & Piras, E. M. (2019). Data work in healthcare: An Introduction. *Health Informatics Journal*, *25*(3), 465–474. https://doi.org/10.1177/1460458219864730

Braun, V., & Clarke, V. (2021). *Thematic analysis: A practical guide*. SAGE Publications.

Braverman, H. (1974). *Labor and monopoly capital: The degradation of work in the twentieth century*. Monthly Review Press.

Faraj, S., Pachidi, S., & Sayegh, K. (2018). Working and organizing in the age of the learning algorithm. *Information and Organization*, *28*(1), 62–70. https://doi.org/10.1016/j.infoandorg.2018.02.005

Kariotis, T., Prictor, M., Gray, K., & Chang, S. (2025). Patient-Accessible Electronic Health Records and Information Practices in Mental Health Care Contexts: Scoping Review. *Journal of Medical Internet Research*, *27*, e54973. https://doi.org/10.2196/54973

Klein, H. K., & Myers, M. D. (1999). A Set of Principles for Conducting and Evaluating Interpretive Field Studies in Information Systems. *MIS Quarterly*, *23*(1), 67–93. https://doi.org/10.2307/249410

Leonardi, P. M. (2011). When Flexible Routines Meet Flexible Technologies: Affordance, Constraint, and the Imbrication of Human and Material Agencies. *MIS Quarterly*, *35*(1), 147–167. https://doi.org/10.2307/23043493

Oettl, F. C., Pruneski, J., Zsidai, B., Yu, Y., Fendrich, D., Tischer, T., Hirschmann, M. T., Zaffagnini, S., Samuelsson, K., & ESSKA Artificial Intelligence Working Group. (2026). From de-skilling to up-skilling: How artificial intelligence will augment the modern physician. *Journal of Experimental Orthopaedics*, *13*(1), e70677. https://doi.org/10.1002/jeo2.70677

Orlikowski, W. J. (1992). The Duality of Technology: Rethinking the Concept of Technology in Organizations. *Organization Science*, *3*(3), 398–427. https://doi.org/10.1287/orsc.3.3.398

Palm, E., Manikantan, A., Mahal, H., Belwadi, S. S., & Pepin, M. E. (2025). Assessing the quality of AI-generated clinical notes: validated evaluation of a large language model ambient scribe. *Frontiers in Artificial Intelligence*, *8*, 1691499. https://doi.org/10.3389/frai.2025.1691499

Papoutsi, C., Wherton, J., Shaw, S., Morrison, C., & Greenhalgh, T. (2020). Putting the social back into sociotechnical: Case studies of co-design in digital health. *Journal of the American Medical Informatics Association*, *28*(2), 284–293. https://doi.org/10.1093/jamia/ocaa197

Pearlman, K., Wan, W., Shah, S., & Laiteerapong, N. (2025). Use of an AI Scribe and Electronic Health Record Efficiency. *JAMA Network Open*, *8*(10), e2537000. https://doi.org/10.1001/jamanetworkopen.2025.37000

Pine, K. H., & Bossen, C. (2020). Good organizational reasons for better medical records: The data work of clinical documentation integrity specialists. *Big Data & Society*, *7*(2), 2053951720965616. https://doi.org/10.1177/2053951720965616

Sanchez, C. G., Goer, V., Kharko, A., Hägglund, M., Hagström, J., Schwarz, J., & Blease, C. R. (2026). Use of ambient AI scribe in physicians' clinical documentation: a protocol for a systematic review on effectiveness, efficiency, and satisfaction. *BMJ Open*, *16*(4), e115562. https://doi.org/10.1136/bmjopen-2025-115562

Sittig, D. F., & Singh, H. (2010). A new sociotechnical model for studying health information technology in complex adaptive healthcare systems. *Quality and Safety in Health Care*, *19*(Suppl 3), i68. https://doi.org/10.1136/qshc.2010.042085

Star, S. L., & Strauss, A. (1999). Layers of silence, arenas of voice: The ecology of visible and invisible work. *Computer Supported Cooperative Work (CSCW)*, *8*(1–2), 9–30. https://doi.org/10.1023/A:1008651105359

Swedish National Agency for Higher Vocational Education. (2022). Myndigheten för yrkeshögskolans föreskrifter om utbildningar inom yrkeshögskolan med inriktning medicinsk vårdadministratör [Regulations on higher vocational education programmes for medical care administrators] (MYHFS 2022:1). https://assets.myh.se/docs/lag-ratt/myhfs/myhfs-2022-1.pdf

Swedish National Agency for Higher Vocational Education. (2025). Områdesanalys 2025: Vårdadministration. https://assets.myh.se/docs/publikationer/omradesanalyser/omradesanalys-vardadministration-2025.pdf

Tandem Health. (n.d.). Tandem Health: AI medical scribe. Retrieved May 8, 2026, from https://tandem-english-v3.webflow.io/

Thompson, P. (2024). Doing labour process theory after Braverman. *Work in the Global Economy*, *4*(2), 139–144. https://doi.org/10.1332/27324176Y2024D000000028

Tierney, A. A., Gayre, G., Hoberman, B., Mattern, B., Ballesca, M., Kipnis, P., Liu, V., & Lee, K. (2024). Ambient Artificial Intelligence Scribes to Alleviate the Burden of Clinical Documentation. *NEJM Catalyst*, *5*(3). https://doi.org/10.1056/cat.23.0404

Topaz, M., Peltonen, L. M., & Zhang, Z. (2025). Beyond human ears: navigating the uncharted risks of AI scribes in clinical practice. *Npj Digital Medicine*, *8*(1), 569. https://doi.org/10.1038/s41746-025-01895-6

Vetenskapsrådet. (2024). *God forskningssed 2024 [Good research practice 2024]*. Swedish Research Council.

Privacy Preserving Federated Learning with Differential Privacy for DDoS Intrusion Detection in Smart City IoT Networks

Zahra Bibi, Nida Fatima, Rubina Khadim, Arooj Zahra
Department of Information and Communication Engineering
The Islamia University of Bahawalpur
Bahawalpur, Pakistan
{zarooch511, nida60271, rkhadim13, aroojz546}@gmail.com

Abstract. *Smart city IoT networks generate massive sensitive traffic data, and centralised intrusion detection systems (IDS) risk serious privacy violations and single points of failure. This paper proposes a privacy-preserving federated learning (FL) framework with differential privacy (DP) for detecting DDoS attacks in smart city IoT networks. Edge devices train local models via FedAvg combined with DP-SGD, sharing only private gradient updates. Using the CICIoT2023 dataset partitioned non-identically across ten clients (Dirichlet* $\alpha = 0.5$*) with heterogeneous sensor noise, plain FL achieves 93.9% F1, while FL+DP at* $\varepsilon \in [5.0, 10.0]$ *maintains F1 above 88%, showing a practical privacy-utility trade-off for edge deployment.*

Keywords. Federated Learning, Differential Privacy, Intrusion Detection System, Smart City IoT, DDoS Attack, CICIoT2023, DP-SGD, FedAvg, Privacy-Utility Trade-off, Opacus

1 Introduction

The rapid spread of Internet of Things (IoT) technologies has transformed urban environments into interconnected smart city ecosystems (Zanella et al., 2014). Thousands of distributed sensors, actuators, smart meters, traffic control systems, public surveillance cameras and emergency response networks form the digital nervous system of modern cities. These systems continuously generate and exchange vast streams of heterogeneous data, enabling real-time urban management, energy optimisation and public safety services. However, this massive connectivity simultaneously creates an expansive and highly attractive attack surface for cyber adversaries.

One of the most serious threats to IoT-based smart city infrastructure is the Distributed Denial of Service (DDoS) attack. For example, a volumetric DDoS attack on a traffic control server could disrupt the coordination of traffic signals across an entire city, causing significant delays and endangering the timely response of emergency vehicles. An application-layer DDoS attack against smart grid communication systems can trigger cascading power failures, and attacks on water utility sensor networks can disable automated contamination detection (Roman et al., 2018). The criticality of these systems demands robust, real-time intrusion detection capable of operating continuously at scale.

Existing centralised architectures for intrusion detection rely on gathering raw traffic from distributed devices into a single processing server where a global detection model is trained. While this architecture can achieve high detection accuracy, it carries serious privacy implications that are difficult to reconcile with smart city requirements. Aggregating raw IoT traffic data from thousands of distributed devices across a city produces highly detailed profiles of individual residents' activity, including but not limited to movement patterns, utility consumption behaviour and communication habits. Such centralisation is difficult to reconcile with data protection frameworks including the General Data Protection Regulation (GDPR) (European Parliament, 2016). Furthermore, the central server constitutes a single point of failure that an adversary can exploit to disable city-wide intrusion detection.

Federated Learning (FL) (T. Li et al., 2020) offers a compelling alternative architecture. Rather than transmitting raw data, each edge device independently trains a local detection model on its private data and transmits only model parameter updates. Since raw data never leaves the originating device, FL provides a strong baseline for privacy preservation. However, recent research has demonstrated that gradient updates can themselves leak sensitive information through gradient inversion and membership inference attacks (Zhu et al., 2019), motivating additional mathematical privacy guarantees.

Differential Privacy (DP) (Dwork & Roth, 2014) provides a rigorous formal framework for quantifying and bounding privacy leakage. Applied to federated learning via the DP-SGD algorithm (Abadi et al.,

2016), differential privacy injects calibrated Gaussian noise into local gradient updates before transmission. The privacy guarantee is controlled by a budget parameter ε: smaller values enforce stronger privacy at the cost of increased noise and reduced model accuracy. Balancing this privacy-utility trade-off in real-time smart city intrusion detection is the practical challenge investigated in this paper.

The primary contribution of this study is a systematic, empirically grounded characterisation of the privacy-utility trade-off in federated DDoS detection, obtained by combining an FL+DP intrusion detection architecture with a controlled sweep of the privacy budget ε on a realistic, heterogeneous smart city IoT setting. This central contribution is supported by three concrete elements:

1. *Architecture and initial evaluation.* A federated learning intrusion detection framework based on the FedAvg protocol, trained on a binary (DDoS-versus-benign) subset of five CICIoT2023 traffic files, non-identically partitioned across ten simulated smart city district clients using a Dirichlet distribution ($\alpha = 0.5$). This scope is deliberately limited (binary classification, five traffic files, single dataset) and is treated as an initial proof of concept rather than a comprehensive evaluation; broader coverage is discussed as future work in Section 6.
2. *Privacy-utility trade-off analysis.* A systematic empirical investigation of how the privacy budget ε affects detection utility, conducted across five representative values {0.5, 1.0, 5.0, 10.0, 50.0} spanning strong to weak privacy enforcement, under realistic heterogeneous per-client sensor noise ($\sigma \in [0.05, 0.50]$) that emulates measurement variability across real-world IoT sensors and gateways.
3. *Communication cost characterisation.* A preliminary feasibility analysis of the communication overhead of the framework, based on measured payload sizes of approximately 49 KB per client per round over 50 rounds; this is intended as a first-order indicator of bandwidth suitability rather than a validated deployment result on physical edge hardware (see Section 4.5).

2 Related Work

2.1 Federated Learning for Intrusion Detection in IoT

Federated learning has been applied extensively in detecting network intrusions. A federated intrusion detection system (IDS) for heterogeneous IoT was designed by T. Li et al. (2020), proving strong detection capability while maintaining privacy by avoiding the exchange of raw data. FL-based IDS has been implemented on the CIC-IDS2017 and N-BaIoT datasets, validating its effectiveness even when using non-IID distributions in practice (Abeshu & Chilamkurti, 2018; Sarhan et al., 2022). Rey et al. (2022) conducted a study on FL-based anomaly detection in simulated city district nodes, achieving over 90% detection rates on known attacks while having lower communication costs than conventional approaches.

2.2 Differential Privacy in Federated Learning

Differential privacy for use in federated learning systems was first introduced by Geyer et al. (2017) by way of client-level privacy via gradient perturbation. The DP-SGD method, developed by Abadi et al. (2016), is the most prominent means of gradient-level privacy for deep learning. Subsequently, DP-SGD was enhanced by McMahan et al. (2018) for user-level privacy in the federated context. Thakkar et al. (2021) introduced adaptive noise techniques and privacy amplification via sampling to mitigate accuracy loss under a strict privacy budget.

2.3 Smart City Security and Research Gap

Security-related studies that consider smart city IoT systems have primarily focused on specific elements of smart city IoT, such as lightweight authentication schemes for constrained devices like sensors (Kairouz et al., 2021), and anomaly detection in smart grids (Yu et al., 2019). Previous studies considering FL-based IDS techniques have generally considered either the case where differential privacy is completely ignored, or cases where differential privacy has been studied for a single fixed budget without investigating the effect of varying levels of privacy on performance degradation. This study specifically addresses the latter problem, as summarised in Table 1.

Table 1. Summary of related work

Reference	Focus Area	Key Contribution
T. Li et al. (2020)	FL for IoT IDS	Robust detection
Abeshu & Chilamkurti (2018); Sarhan et al. (2022)	FL for IDS	Confirmed efficiency
Rey et al. (2022)	FL anomaly detection	>90% accuracy
Geyer et al. (2017)	DP in FL	Efficient perturbation
Abadi et al. (2016)	DP in deep learning	DP-SGD
McMahan et al. (2018)	DP in FL	Extended DP

Thakkar et al. (2021)	DP in FL	Adaptive amplification
Kairouz et al. (2021)	Smart city security	Lightweight sensors
Yu et al. (2019)	Smart city security	Anomaly detection
This paper	Gap addressed	Systematic trade-off

3 Proposed Methodology

3.1 System Model

The proposed system models a smart city as a set of N = 10 geographically distributed edge nodes, each representing a city district or infrastructure zone (commercial, residential, industrial, transport hub or administrative centre). Each edge node i maintains a local private dataset Di containing network traffic records captured from local IoT device communications. A central aggregation server, implemented as a fog computing node, coordinates federated training without accessing raw data from any client.

The FL process proceeds over T = 50 global communication rounds. In each round, the server broadcasts the current global model parameters wt to all participating edge nodes. Each edge node performs E = 5 local training epochs using DP-SGD, clips gradients to bound ℓ2-sensitivity, adds calibrated Gaussian noise and transmits the resulting differentially private gradient update g̃ i to the central server. The server aggregates the received updates using FedAvg and produces the updated global model wt+1.

3.2 Dataset and Data Partitioning

The CICIoT2023 dataset (Neto et al., 2023), developed by the Canadian Institute for Cybersecurity, is selected as the primary evaluation benchmark. Five specific CSV files are used:

BenignTraffic3.pcap.csv,
DDoS-ICMP_Flood9.pcap.csv,
DDoS-TCP_Flood17.pcap.csv,
DDoS-UDP_Flood20.pcap.csv and
Mirai-udpplain.pcap.csv.

All files are merged, filtered to retain DDoS and benign records only, and labelled as binary classification targets. This choice of five files and a binary DDoS-versus-benign formulation is a deliberate scoping decision for this initial study, adopted to keep the federated simulation tractable while the privacy-utility trade-off is characterised; it is not intended as a full evaluation over the CICIoT2023 attack taxonomy, and is revisited as a limitation in Section 5 and as future work in Section 6. All 46 numerical flow features are retained after removing constant-value and all-NaN columns. StandardScaler normalisation is applied prior to a stratified 80:20 train-test split.

To simulate realistic federated deployment, the training set is partitioned across 10 clients using a Dirichlet distribution with concentration parameter α = 0.5, creating heterogeneous class distributions across clients. To further simulate real IoT sensor heterogeneity, per-client Gaussian noise with client-specific standard deviations σi ∈ {0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50} is injected into local training data. Evaluation is performed on a held-out noisy test set (σ = 0.30) to reflect measurement variability in real-world IoT gateway readings.

3.3 Justification of Experimental Design Choices

The experimental configuration was chosen to balance realism, computational tractability and comparability with prior work, as follows.

Number of clients (N = 10). Ten clients were chosen to represent a small but heterogeneous set of city district archetypes (commercial, residential, industrial, transport and administrative), consistent with the scale used in comparable simulated federated IDS studies (Rey et al., 2022). This value keeps per-round computation tractable on a single workstation (Section 4.1) while still requiring the aggregation server to reconcile non-trivially different local data distributions, which is the regime of interest for this study. Larger client counts are left for future work (Section 6) once the framework is validated on real hardware.

Privacy budget values. The five ε values {0.5, 1.0, 5.0, 10.0, 50.0} were selected to span, on a roughly logarithmic scale, the range from strong privacy enforcement to near-negligible privacy cost, consistent with the range of budgets commonly reported in the DP-SGD literature (Abadi et al., 2016; Kairouz et al., 2021). This spread allows the point at which utility begins to degrade sharply to be located empirically (Section 4.3) rather than assumed a priori, directly addressing the gap identified in Section 2.3, where prior FL-based IDS work generally evaluates a single fixed privacy budget.

Dirichlet partitioning (α = 0.5). Non-identical (non-IID) data partitioning is standard practice for simulating realistic federated deployments, since different city districts are not expected to observe the same mix of benign and attack traffic (Q. Li et al., 2022; Zhao et al., 2018). A concentration parameter of α = 0.5 produces moderate, rather than extreme, label skew across clients: it is low enough to create meaningfully heterogeneous local class distributions, while remaining within the range for which FedAvg is known to converge without additional non-IID mitigation mechanisms (Q. Li et al., 2022).

Heterogeneous per-client noise levels. The injected Gaussian noise range σ ∈ [0.05, 0.50] is intended to emulate differences in sensor calibration, measurement

precision and gateway signal quality that are expected across a real city's IoT infrastructure, rather than to model any specific deployed sensor. The range is deliberately wide, from near-clean ($\sigma = 0.05$) to substantially corrupted ($\sigma = 0.50$) readings, so that the resulting FL and FL+DP results reflect performance under a spread of plausible real-world data quality conditions rather than a single idealised setting. This is an approximation adopted for the initial study; calibrating the noise model against measured sensor error statistics from a real smart city deployment is identified as future work (Section 6).

3.4 Local Model Architecture

Each edge client trains a lightweight Multilayer Perceptron (MLP) classifier designed for resource-constrained IoT edge hardware. The architecture consists of three fully connected layers with 128, 64 and 32 units respectively, each followed by GroupNorm normalisation and ReLU activation. The output layer applies a sigmoid activation for binary DDoS classification. Total trainable parameters: ≈12,576.

GroupNorm is selected specifically in place of BatchNorm1d because BatchNorm1d makes each sample's normalised value depend on its batch peers, which conflicts with the per-sample gradient clipping required by Opacus DP-SGD (Yousefpour et al., 2021). GroupNorm normalises within each individual sample independently, making it fully compatible with differential privacy enforcement. Ghost clipping mode (grad_sample_mode='ghost') is employed to reduce per-sample gradient memory overhead.

3.5 Federated Averaging Protocol

Global model training follows the FedAvg protocol (McMahan et al., 2017). In each communication round, all N clients perform E = 5 local epochs with batch size 32, then return updated parameters to the server. Large clients exceeding 3,000 samples are randomly subsampled per round. The server computes the aggregated global update as:

$$\mathbf{w}_{t+1} = \Sigma_{i=1}^{N} \left(|D_i| \,/\, |D| \; \mathbf{w}_{t+1}^{i}\right) \qquad (1)$$

where $|D| = \Sigma i\ |Di|$ is the total dataset size. This data-size-weighted aggregation ensures that clients with larger local datasets exert proportionally greater influence on the global model, which is important for convergence under non-IID distributions (Zhao et al., 2018).

3.6 Differential Privacy via DP-SGD

Differential privacy is enforced at the gradient level using DP-SGD (Abadi et al., 2016) via the Opacus library (Yousefpour et al., 2021). During each local training step, per-sample gradients are clipped to a maximum $\ell 2$ norm C = 1.0 to bound sensitivity, and calibrated Gaussian noise is added:

$$\tilde{g} = (1/|B|)\ [\Sigma_{j \in B}\ clip(\nabla_j, C) + N(0, \sigma^2 C^2 I)] \qquad (2)$$

where B is the mini-batch and C = 1.0 is the gradient clipping threshold.

Cumulative privacy accounting across rounds. Each client's Opacus PrivacyEngine and DP optimiser are instantiated once, before the first communication round, and are not reset between rounds: the same privacy engine object persists for the client across all T = 50 global rounds, so that the Rényi Differential Privacy (RDP) accountant used by Opacus (accountant='rdp') tracks privacy loss over the full sequence of T × E = 250 local optimisation steps per client, not merely over one round's E = 5 local epochs. Consequently, the noise multiplier σ for each target ε is calibrated once, before training begins, to satisfy the target (ε, δ) pair with $\delta = 10^{-5}$ over the entire 250-step training horizon rather than over a single round. All ε values reported in this paper — {0.5, 1.0, 5.0, 10.0, 50.0} — are therefore final, cumulative privacy budgets that already account for the repeated access to each client's data across all 50 rounds, and not per-round budgets. This is essential for correctness: reporting an uncomposed per-round ε while training for 50 rounds would substantially understate the true cumulative privacy loss.

The Poisson subsampling rate for client i is qi = B / min(|Di|, 3000), where B = 32 is the batch size and the cap of 3,000 samples corresponds to the per-round subsampling applied to larger clients described in Section 3.2. Because the Dirichlet partition produces clients of different sizes, qi varies across clients; the noise multiplier σ is calibrated per client using that client's own qi so that every client independently satisfies the same target (ε, δ) over its own 250 local steps. Table 2 reports the resulting noise multiplier for a representative client at the subsampling cap (q = 32/3000 ≈ 0.0107), computed with Opacus's RDP accountant under $\delta = 10^{-5}$ and 250 steps; smaller clients with larger qi require a correspondingly larger σ to reach the same ε and are calibrated individually in the same manner.

Table 2. Noise multiplier required to reach each target cumulative ε ($\delta = 10^{-5}$, 250 local steps, representative client at the subsampling cap, q ≈ 0.0107, via Opacus's RDP accountant)

Target ε	Noise multiplier σ
0.5	1.660
1.0	1.177
5.0	0.640
10.0	0.506
50.0	0.291

As expected, smaller (stricter) target ε values require a larger noise multiplier and therefore inject

more gradient noise, reducing model utility. Five complete experimental runs are conducted, one per target ε ∈ {0.5, 1.0, 5.0, 10.0, 50.0}, representing a spectrum from strong to weak privacy enforcement, each calibrated as described above.

4 Experimental Results

4.1 Experimental Setup

All experiments are conducted using the PyTorch framework, with Opacus providing differentially private gradient computation within a custom FedAvg implementation simulating the ten clients described in Section 3 as independent processes on a single machine. Simulations were performed on a workstation with 16 GB RAM and an 8-core CPU. Three scenarios are evaluated in this study, namely (1) a centralised baseline IDS trained on the full aggregated noisy dataset, (2) plain FL without DP using per-client heterogeneous noise, and (3) FL with DP-SGD at five ε values. Full experimental parameters are summarised in Table 3.

Table 3. Experimental configuration parameters

Parameter	Value
Dataset	CICIoT2023 (DDoS vs. Benign)
FL Clients (N)	10 (simulated city districts)
Local Epochs (E)	5
Global Rounds (T)	50
Batch Size	32
Grad Clip Norm (C)	1.0
Delta (δ)	1×10^{-5}
ε values	0.5, 1.0, 5.0, 10.0, 50.0
MLP Hidden Layers	128–64–32
Normalisation Layer	GroupNorm (DP-safe)
DP Library	Opacus (PyTorch)
Client Noise σ	0.05–0.50 (heterogeneous)
Max Samples/Client	3,000 (subsampled/round)
Non-IID Param α	0.5 (Dirichlet)

4.2 Overall Detection Performance

Table 4 reports Accuracy, Precision, Recall and F1-Score for all experimental conditions evaluated on the noisy held-out test set. The centralised baseline achieves 96.8% accuracy and 96.3% F1 score, serving as the upper-bound reference. Plain FL achieves 94.5% accuracy and 93.9% F1 score, a modest reduction of 2.3 and 2.4 percentage points respectively, attributable to non-IID data partitioning and federated aggregation overhead.

At weak privacy enforcement (ε = 50), FL+DP achieves 93.2% accuracy and 92.7% F1, confirming that low noise injection is effectively absorbed. As ε decreases to 10.0, F1-score drops moderately to 90.8%, remaining above the 90% operational threshold. At ε = 5.0, F1 score reaches 88.1%, still within practical operational bounds. At ε = 1.0, F1 score drops to 82.7%, and at the strongest setting (ε = 0.5), F1 score reaches 76.3%, reflecting the disproportionate noise penalty at very tight privacy budgets.

Table 4. Performance comparison centralized vs. FL vs. FL+DP

Method	Acc. (%)	Prec. (%)	Rec. (%)	F1 (%)
Centralised IDS	96.8	96.2	96.5	96.3
Plain FL (No DP)	94.5	93.8	94.1	93.9
FL+DP (ε=50)	93.2	92.5	92.9	92.7
FL+DP (ε=10)	91.4	90.6	91.0	90.8
FL+DP (ε=5)	88.7	87.9	88.4	88.1
FL+DP (ε=1)	83.1	82.4	83.0	82.7
FL+DP (ε=0.5)	76.8	76.1	76.5	76.3

4.3 Privacy Budget vs. F1 Score Analysis

As expected, there exists a consistent monotonic relation between F1 score and ε. Namely, higher values of ε (synonymous with less privacy) result in higher F1 scores, while smaller ε results in reduced F1 score. However, the degradation is not linear: the fastest reduction happens when going from ε = 1.0 to ε = 0.5. This can be explained by noting that the noise level of DP-SGD scales inversely with ε, resulting in strong distortion for tight budgets. These results are summarised in the form of a bar plot in Fig. 1.

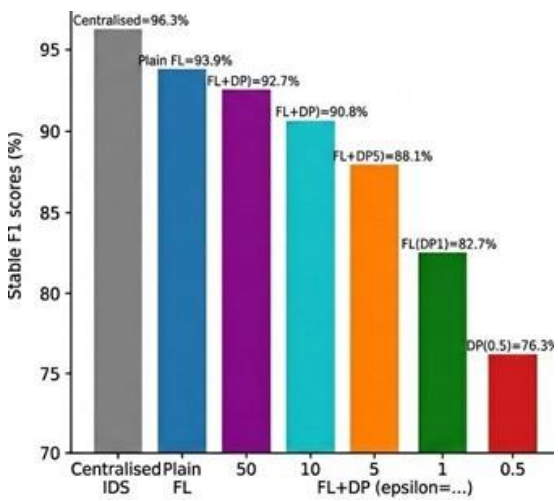


Figure 1. F1 score comparison across all evaluated methods, demonstrating the monotonic privacy-utility trade-off as ε decreases from 50 to 0.5.

These results confirm that ε ∈ [5.0, 10.0] represents a practical operating region where meaningful formal privacy guarantees are maintained alongside detection accuracy above 88%, which is operationally acceptable for real-time smart city DDoS detection. The consistent ordering of results across all four metrics indicates that the privacy-utility trade-off is stable and reproducible, not an artefact of any single metric.

4.4 Convergence Analysis

Convergence of the federated model over 50 rounds of global communication was measured for each scenario. Fig. 2 displays the F1-score round-by-round measurement for both vanilla FL and FL+DP with all five ε values. Vanilla FL reaches convergence within just the first 5 to 10 rounds, maintaining steady performance afterwards. FL+DP with ε = 5.0 and 10.0, due to noise induced by gradient perturbation, takes roughly 30 to 35 rounds to reach the same level of performance, with oscillations in the last ten rounds at a variance of less than ±0.01%.

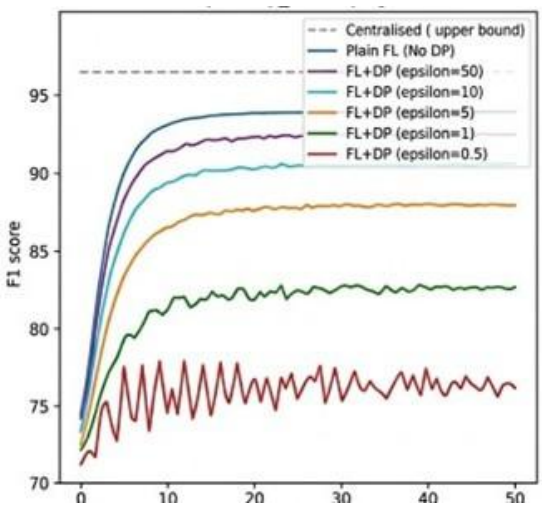


Figure 2. Convergence: F1 score over 50 FL rounds for plain FL and all five FL+DP configurations. Tighter ε values exhibit slower and noisier convergence due to increased gradient noise.

At strict privacy settings (ε = 0.5 and ε = 1.0), convergence is slower and less stable, with persistent F1 oscillations in early rounds attributable to high gradient noise injection. These configurations would benefit from additional communication rounds beyond T = 50 or adaptive noise scheduling mechanisms, consistent with prior theoretical analysis of DP-SGD convergence bounds under high-noise regimes (Abadi et al., 2016; Thakkar et al., 2021).

4.5 Communication Cost Analysis

Each federated round requires transmission of MLP model parameters (≈12,576 parameters at float32 precision) between server and clients, requiring approximately 49 KB per client per round. Over the full T = 50 communication rounds, the total cost per client is approximately 2.4 MB, and approximately 24 MB in aggregate across all 10 clients. This lightweight communication profile, consisting only of model parameter updates rather than raw traffic data, is well within the bandwidth available to typical smart city IoT gateway devices over 4G/LTE or Wi-Fi backhaul links, while also preserving privacy by never transmitting raw traffic. As noted in Section 5, this is a bandwidth estimate based on parameter size alone; it does not yet account for real network conditions or device-level constraints, which would need to be validated on physical hardware.

5 Discussion and Limitations

The experimental results demonstrate that the proposed FL+DP framework achieves a meaningful and practical balance between privacy protection and DDoS detection capability. The framework provides formal, mathematically grounded privacy guarantees absent from conventional centralised IDS deployments, eliminates dangerous single points of failure, and maintains detection performance above 88% F1-score at ε values that provide meaningful real-world privacy protection.

The identification of ε ∈ [5.0, 10.0] as a practical operating range provides concrete, actionable guidance for smart city security practitioners.

The successful training of the lightweight MLP architecture, together with GroupNorm normalisation and ghost clipping, shows that differentially private training can be carried out without GPU acceleration on a standard workstation. This is a necessary, but not sufficient, condition for deployment on Raspberry Pi-type edge devices: it suggests such deployment is plausible, but the present study does not measure actual CPU time, memory footprint or thermal behaviour on such hardware, so this remains a claim to be validated rather than a demonstrated result (see limitations below). The heterogeneous noise injection per client is intended to reflect a more realistic environment for IoT deployments than uniform noise models, capturing the varying measurement quality, calibration drift and environmental interference expected across different sensors and gateway devices in real cities.

Several important limitations constrain the current evaluation. *First*, the experimental setup simulates all 10 edge clients as independent processes on a single machine, meaning real-world network latency, packet loss and hardware heterogeneity are not fully captured; in particular, the framework has not yet been executed on physical resource-constrained hardware such as a Raspberry Pi or Jetson Nano, so claims regarding edge deployment feasibility (Section 4.5) are based on parameter and communication size estimates rather than measured on-device latency, memory use or power draw. *Second*, the threat model assumes all FL

clients are honest, which does not account for adversarial clients attempting model poisoning or gradient inversion attacks. *Third*, while non-IID partitioning and per-client noise injection provide a reasonable approximation of district-level heterogeneity, real smart city traffic distributions may exhibit more extreme temporal drift not captured in the static partition. *Fourth*, the evaluation is confined to binary DDoS-versus-benign classification; multi-class detection across the full CICIoT2023 taxonomy would provide a more comprehensive assessment. *Fifth*, the dataset was collected in a controlled laboratory environment; field validation on live smart city infrastructure remains an important future direction.

6 Conclusion and Future Work

This paper presented a privacy-preserving federated learning framework incorporating differential privacy via DP-SGD for DDoS intrusion detection in smart city IoT networks. The proposed architecture enables distributed edge devices representing city districts to collaboratively train effective intrusion detection models without exposing sensitive raw network traffic data, providing formal privacy guarantees alongside competitive detection performance. Experiments on five real IoT traffic CSV files from the CICIoT2023 dataset with realistic per-client heterogeneous sensor noise and Dirichlet non-IID partitioning demonstrated that plain federated learning achieves a 93.9% F1 score with minimal accuracy loss compared to centralised training. Differential privacy at $\varepsilon \in [5.0, 10.0]$ maintains F1 scores above 88% while providing meaningful formal privacy protection, representing a practical operating range for smart city deployments. Communication cost analysis shows that the framework transmits only ≈49 KB per client per round, suggesting good potential for deployment on resource-constrained smart city edge infrastructure, though this potential remains to be confirmed through testing on physical edge hardware.

Future work will pursue the following directions. Adaptive differential privacy mechanisms that dynamically adjust noise levels based on real-time training convergence signals will be investigated. The system will be extended to include multi-class attack detection over the entire CICIoT2023 taxonomy. Resilience to adversarial poisoning attacks will be examined through the use of Byzantine-robust aggregation methods such as Krum and coordinate-wise median. Deployment and performance will be investigated on actual Raspberry Pi 4 and Jetson Nano edge devices to assess practical feasibility while measuring real power, memory and thermal characteristics. Finally, integrating dynamic spatio-temporal graph models of threat spread across districts is another promising direction for advancing coordinated urban-scale threat intelligence.

Acknowledgments

The authors would like to express their sincere gratitude to The Islamia University of Bahawalpur for the research facilities. Special thanks to Asst. Prof. Iftikhar Rasheed for guidance and to friends (Arooj Fatima, Aqsa Faryad and Nimra Ishfaq) for their support during this work.

References

Abadi, M., et al. (2016). Deep learning with differential privacy. *In Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS)* (pp. 308–318). Vienna, Austria.

Abeshu, A., & Chilamkurti, N. (2018). Deep learning: The frontier for distributed attack detection in fog-to-things computing. *IEEE Communications Magazine*, 56(2), 169–175.

Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4), 211–407.

European Parliament. (2016). General Data Protection Regulation (GDPR), Regulation (EU) 2016/679. *Official Journal of the European Union.*

Geyer, R. C., Klein, T., & Nabi, M. (2017). Differentially private federated learning. *In Workshop on Privacy-Preserving Machine Learning, NeurIPS*. Long Beach, CA, USA.

Kairouz, P., et al. (2021). Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1–2), 1–210.

Li, Q., et al. (2022). Federated learning on non-IID data silos: An experimental study. *In Proceedings of the IEEE International Conference on Data Engineering (ICDE).*

Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3), 50–60.

McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. *In Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)* (pp. 1273–1282). Fort Lauderdale, FL, USA.

McMahan, H. B., Ramage, D., Talwar, K., & Zhang, L. (2018). Learning differentially private recurrent language models. *In Proceedings of the International Conference on Learning Representations (ICLR)*. Vancouver, Canada.

Neto, E. C. P., et al. (2023). CICIoT2023: A real-time dataset and benchmark for large-scale attacks in IoT environment. *Sensors*, 23(13), 5941.

Rey, V., Sánchez, P. M. S., Celdrán, A. H., & Bovet, G. (2022). Federated learning for malware detection in IoT devices. *Computer Networks*, 204, 108693.

Roman, R., Lopez, J., & Mambo, M. (2018). Mobile edge computing, fog et al.: A survey and analysis of security threats and challenges. *Future Generation Computer Systems*, 78, 680–698.

Sarhan, M., Layeghy, S., Moustafa, N., & Portmann, M. (2022). Federated network intrusion detection systems: Aggregation approaches. *In Proceedings of the IEEE Symposium on Computers and Communications (ISCC)*. Rhodes Island, Greece.

Thakkar, O., Andrew, G., & McMahan, H. B. (2021). Differentially private learning with adaptive clipping. *In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS).*

Yousefpour, A., et al. (2021). Opacus: User-friendly differential privacy library in PyTorch. *arXiv preprint arXiv:2109.12298.*

Yu, L., Liu, L., Pu, C., Gursoy, M. E., & Truex, S. (2019). Differentially private model publishing for deep learning. *In Proceedings of the IEEE Symposium on Security and Privacy (S&P)*. San Francisco, CA, USA.

Zanella, A., Bui, N., Castellani, A., Vangelista, L., & Zorzi, M. (2014). Internet of things for smart cities. *IEEE Internet of Things Journal*, 1(1), 22–32.

Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., & Chandra, V. (2018). Federated learning with non-IID data. *arXiv preprint arXiv:1806.00582.*

Zhu, L., Liu, Z., & Han, S. (2019). Deep leakage from gradients. *In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)* (pp. 14774–14784).

Towards Autonomous Resilience: Architecting Agentic Middleware for the Cloud-Edge Continuum

Lovro Posarić, Tara Nikolić
University of Zagreb
Faculty of Organization and Informatics
IoT Laboratory
Pavlinska 2, 42000 Varaždin
{lposaric, tnikolic}@foi.hr

Abstract. *The Cloud-Edge-IoT continuum reduces latency for smart-city services but also introduces operational fragility: cascading, cross-tier failures can disrupt orchestration frameworks designed for steady-state optimisation. This paper introduces a three-tier agentic middleware coupling lightweight deterministic edge heuristics with selective escalation to a Cloud-tier orchestrator. A deterministic Escalation Score, constrained-decoding output schema, pre-execution validation, and decision logging bound the escalation pathway and make it reconstructible after an event. Alongside the architectural specification, we contribute a failure-first evaluation paradigm designed to quantify the cost of failure against the cost of intelligence through discrete-event simulation over open-data-derived urban dependency graphs. The paper is positioned as a design contribution: the architecture and evaluation framework are specified at the design level, while the empirical study is planned rather than completed.*

Keywords. Cloud-Edge continuum, smart-city infrastructure, agentic middleware, fault tolerance, large language models

1 Introduction

Modern smart cities increasingly distribute computation across the Cloud-Edge-IoT (CEI) continuum rather than concentrating it centrally, because the growing volume of devices and cognitive-service demand has outgrown what a centralised Cloud can absorb on its own (Chaturvedi et al., 2026; Wei et al., 2026). Pushing computation to the network edge reduces latency and bandwidth bottlenecks, enabling real-time urban services. However, distributing computing power across multi-operator, heterogeneous environments sharply increases operational vulnerability, and Cloud-oriented orchestration tools were not designed for the extreme device heterogeneity and dynamic conditions of the continuum (Belcastro et al., 2026). Current orchestration literature addresses dynamic placement, monitoring, and adaptation, but evaluations provide limited evidence on containment of correlated compound failures propagating across service dependencies and infrastructure tiers (Angelis & Kousiouris, 2025), leaving an open question about operational resilience under compound, cross-tier failures.

Consider a storm that unfolds progressively. Rising water causes some Far-Edge acoustic sensors to stop reporting while others continue to transmit plausible but corrupted baseline values. Because conventional gateways act as passive forwarders, missing and integrity-violating readings can propagate upward without coordinated handling. As the storm severs a fibre link, the Smart-Edge tier is partitioned from the Cloud, the static orchestrator becomes unreachable, and the unhandled anomalies cascade, exhausting local resources and driving critical emergency-routing services into severe latency spikes. This architectural fault stems from an inability to reason about collapsing cross-tier dependencies. Beyond single-node recovery, such systems would benefit from a layer that recognises compound failures and coordinates reconfiguration before the cascade propagates widely.

The proposed architecture keeps its escalation and governance mechanisms separate from the domain models, so the former transfer across Cloud-Edge-IoT domains while the dependency, criticality, workload, and fault models remain specific to the smart-city setting. Two smart-city properties drive the design directly. First, multi-operator custody means no individual operator holds the complete service-dependency graph. The design therefore assumes a municipality-level platform that composes a versioned structural graph from operator declarations. The structural graph is shared, whereas live telemetry stays local to each Smart-Edge district. Each escalation therefore carries the relevant topology and criticality context. Second, physical-world events such as storms and accidents can produce correlated, potentially out-of-distribution failures across independently managed services, motivating the novelty-aware trigger and failure-first evaluation. Mixed criticality constrains the permitted recovery actions, while heterogeneity and scale shape implementation require-

ments but are not unique to smart cities. Addressing these limitations calls for an orchestration layer that can reason about collapsing cross-tier dependencies and coordinate reconfiguration beyond single-node recovery. This paper makes three contributions:

1. Architectural design: The formal specification of a three-tier agentic middleware with a transparent, threshold-governed escalation model for safely integrating Large Language Model (LLM) reasoning into physical infrastructure.
2. Theoretical foundation: A formal escalation criterion consisting of a deterministic score over a compact, interpretable system-state vector that defines the boundary at which reactive edge heuristics escalate to Cloud-tier semantic reasoning.
3. Methodological framework: A failure-first evaluation paradigm that quantifies the "cost of intelligence" against the "cost of failure."

This is a design and blueprint contribution. The evaluation framework specified below has not yet been executed, so the hypotheses are stated as planned tests, and the paper reports no empirical results.

2 Background and related work

Literature predominantly models the CEI continuum as a hierarchy reflecting differences in hardware and network reliability (Gkonis et al., 2026). At the perimeter, the Far-Edge comprises massive deployments of constrained sensors and microcontrollers; relying purely on centralised Cloud processing introduces communication overhead and single points of failure. A Smart-Edge tier of robust localised nodes (e.g., lamppost-mounted edge nodes) therefore bridges the fragile Far-Edge and the distant Cloud, and the consensus is that it must possess high local autonomy to execute latency-constrained workloads without continuous Cloud validation (Liu et al., 2025).

Recent work pushes increasingly capable models onto this tier: lightweight edge autoencoders reconstruct and analyse raw Far-Edge telemetry locally, avoiding the Cloud round-trip (He et al., 2026), and quantised deep Q-self-attention transformers run on localised hardware for high-accuracy tasks such as real-time PM2.5 prediction (Balasubramanian et al., 2026). Such models, however, are optimised for data-stream prediction rather than operational resilience: if the hosting node loses power or is partitioned, the service simply terminates. Giving edge nodes local intelligence is necessary, but coordinating isolated components during failure remains unresolved.

2.1 Current orchestration methodologies

Two approaches dominate, both using graph-theoretic state representations where nodes are vertices and communication paths are edges. This allows graph neural networks and attention mechanisms to restrict peer-to-peer communication to the most relevant neighbours, minimising overhead (Chen et al., 2025). First, major European initiatives, including aerOS, NEMO, and FLUIDOS, develop continuum-spanning Meta-Operating Systems that expose heterogeneous infrastructure as a single virtualised resource pool (Gkonis et al., 2026). NEMO's reference architecture identifies frequent node arrival and departure as a core challenge demanding adaptive management (Chochliouros et al., 2023), though this churn is addressed primarily as a membership and resource-discovery concern, which differs in scope from the simultaneous cross-tier collapse considered here.

Second, the field has widely pivoted to Deep and Multi-Agent Reinforcement Learning (DRL/MARL) for dynamic scheduling (Liu et al., 2025; Zhang et al., 2026). Recent work removes its practical bottlenecks: bi-level and hierarchical designs contain the action space during task offloading (Babaiyan et al., 2026), neuro-symbolic frameworks pair neural optimisation with symbolic constraints to fit edge resource limits (Abbass et al., 2026), and proactive migration absorbs sudden load spikes (Alqerm et al., 2026). These methods scale from single-agent scheduling to multi-operator settings through game-theoretic negotiation (Wang et al., 2026) and cooperative stochastic games (Kuchuk et al., 2026). Yet these approaches are predominantly evaluated under conditions of routine load and mobility, with limited testing against simultaneous cross-tier infrastructure collapse. While hierarchical DRL excels at routing through a healthy continuum, its ability to recover services when the infrastructure itself collapses has not been demonstrated under compound failure conditions.

2.2 The transition to semantic AI and its limits

Mathematical optimisation presumes a well-specified objective and state space, so it is poorly suited to failures where the conflict is which objective applies rather than how to satisfy a fixed one. This limitation has motivated recent research on LLMs and agentic architectures. In orchestration, AgentEdge employs a simulate-before-acting validation step to contain unsafe actions (Gort et al., 2025), while SORA-ATMAS uses policy-aligned governance and high-risk fallbacks to coordinate complex urban services safely (Antuley et al., 2026). In diagnosis, frameworks such as BANS (Xiong et al., 2025) and IntentContinuum (Akbari et al., 2025) employ LLMs to perform root-cause analysis over structured telemetry.

These diagnostic systems are notable because they demonstrate that system-level decoupling is achievable by pairing lightweight edge models for low-latency detection with robust Cloud-tier models for heavy se-

mantic reasoning. A common limitation of these integrations is their evaluation environment. They are predominantly tested against software-level faults, such as bottlenecks or service-level objective (SLO) violations, all on stable physical infrastructure. How these state-of-the-art semantic frameworks perform under the stress of a degrading physical topology remains insufficiently characterised.

2.3 Open challenges

Beyond these evaluation-environment limits, three open challenges limit the evidence for operational resilience. The first is a predominant focus on *steady-state conditions* within current research. For instance, a review of self-adaptive Cloud studies found that availability and stability were the primary goals in only about 9% of the reviewed works (Angelis & Kousiouris, 2025). Consequently, established approaches rely on single-failure models that cannot dynamically restructure system topology during concurrent disruptions.

Second, the *limited generalisation* of reinforcement learning: RL offers limited semantic reasoning under novel conditions. DRL policies tend to overfit their training environments and generalise poorly (Kirk et al., 2023), and simultaneous degradation across multiple tiers is precisely the kind of out-of-distribution state that exposes this weakness. The core difficulty is novelty rather than priority: priority can be encoded as a numerical reward weight, whereas structural novelty is harder to represent in a fixed scalar objective. A policy may observe a symptom, such as a latency spike on a saturated queue, without representing why the spike arose or which cross-domain dependencies interact. Furthermore, the multi-agent credit-assignment problem makes it difficult to attribute a systemic outcome to the local decision that produced it (Foerster et al., 2018), complicating post-event attribution and audit (T. Xu et al., 2026).

Third, the *empirical gap and the Opaque Autonomy Paradox* identified by Caldeira et al. (2026), in which increasing an autonomous system's capability tends to erode the transparency and auditability that high-stakes deployment demands. While recent frameworks propose LLM-based self-healing edge architectures, the operational latency overhead and containment effect of semantic reasoning under physical stress remain insufficiently characterised across this line of work (R. Xu et al., 2024). Consequently, running full LLMs on constrained edge nodes is often impractical, and municipal operators cannot accept opaque restructuring of critical services, which together motivate separating where reasoning happens from where action is taken. Pre-execution validation approaches that simulate recovery plans before committing them (Zhao & Han, 2025) point towards auditable deployment but have been tested only on Cloud fault-injection, not true cross-tier collapse. Published evidence on how an intelligent orchestrator behaves under compound failure therefore remains limited, motivating a failure-first paradigm and a decoupled architecture.

3 Agentic middleware architecture

To achieve operational resilience under cascading physical failures, this paper proposes the integration of an active, semantic orchestration layer over the conventional urban deployment model. This three-tier agentic middleware is designed to overcome the structural limitations of reactive frameworks by dynamically restructuring component interactions during infrastructure collapse.

3.1 The three-tier Cloud-Edge continuum

The smart-city environment is modelled as a three-tier physical continuum of heterogeneous devices, from Far-Edge microcontrollers to Cloud servers. The baseline AS-IS architecture (Fig. 1) relies on static or reactive orchestrators. It reproduces the failure chain of the storm scenario: passive Far-Edge gateways, no peer-to-peer coordination at the Smart-Edge, and an orchestrator rendered unreachable during a Cloud partition.

The proposed TO-BE middleware (Fig. 2) addresses these limitations by reconfiguring component interactions as the failure unfolds. Gateways detect missing publications through localised heartbeat timeouts and, where applicable, the Last Will and Testament (LWT) feature of Message Queuing Telemetry Transport (MQTT); lightweight, sensor-specific integrity checks flag telemetry that remains available but violates plausibility or drift constraints. Detection is represented in simulation by a class-specific delay between fault occurrence and visibility to the local agent. Intercepted anomalies then reach the Smart-Edge's deterministic heuristic policies, which resolve isolated faults through peer-to-peer (P2P) coordination (e.g., migrating stateless tasks to a healthy neighbour when CPU utilisation exceeds a threshold).

Each Smart-Edge cluster agent governs a defined geographic zone, providing inter-district failure isolation and keeping routine decisions local to minimise LLM invocation. When a cascade exceeds local capacity, a deterministic Escalation Score triggers the Cloud-tier reasoning orchestrator, which is architecturally modular and, as a proof of concept, instantiated with a quantised, open-weight LLM on a municipality-operated on-premises server.

3.2 System state vector and local heuristics

To evaluate system health locally without continuously streaming raw telemetry to the Cloud, each Smart-Edge agent periodically updates a multidimensional state vec-

tor S by aggregating local node telemetry and anomaly flags intercepted from the Far-Edge gateways:

$$S = \{F_c,\, D_{cs},\, N_i,\, C_h,\, T_s\}. \tag{1}$$

Its components are: the concurrent fault count F_c (active, non-isolated faults in the zone); the cross-service dependency depth D_{cs} (maximum topological depth of cascading impact across intersecting systems); a novelty index $N_i \in [0, 1]$ measuring structural dissimilarity of the current failure pattern against a local cache of resolved anomalies; a local offload-capacity score $C_h \in [0, 1]$ derived from remaining resource margins on adjacent healthy nodes; and a tier span $T_s \in \{1, 2, 3\}$, the highest tier the failure has reached (1 = Far-Edge, 2 = Smart-Edge, 3 = Cloud). Representing local health as an explicit, low-dimensional feature vector rather than a latent embedding follows the rationale of explainable feature learning (Vlahek & Mongus, 2023) and makes the escalation trigger easier to inspect. It does not, by itself, explain the LLM's internal action-selection process.

Because the state vector mixes continuous and categorical features, N_i is computed using Gower's distance, which handles mixed types natively. Each district cache contains one entry for every eligible calibration signature successfully contained by the heuristic-only baseline. It has no separately imposed capacity limit, as its realised size and lookup cost are reported with the experimental configuration. N_i is the mean Gower distance between the current signature (F_c, D_{cs}, C_h, T_s) and its k nearest eligible cache entries, where k is a fixed calibration constant. The cache is populated only from calibration cases the heuristic layer successfully contained, and no case contributes to scoring itself, preventing leakage. After model selection, the caches are reconstructed from all calibration seeds and frozen for evaluation, with no online update or eviction. If a district has fewer than k eligible entries, the remaining neighbours are drawn without duplication from a pooled calibration cache built under the same eligibility rules.

3.3 Deterministic escalation and constrained reasoning

To address the Opaque Autonomy Paradox, a deterministic trigger governs the transition to Cloud-tier reasoning, and outputs are strictly constrained.

The Escalation Score is a sigmoid over a linear combination of the state-vector features, with weights $\boldsymbol{\beta}$ and intercept β_0 calibrated from labelled simulation data before evaluation:

$$\mathrm{ES}(S) = \sigma\Big(\beta_0 + \boldsymbol{\beta}^\top S\Big), \qquad \mathrm{ES} \in (0, 1), \tag{2}$$

where $\sigma(x) = 1/(1+e^{-x})$ is the logistic function and S is the state vector of Eq. (1). Calibration labels are generated in simulation under the heuristic-only baseline. A state sampled at time t is labelled 1 if, during a fixed look-ahead horizon, at least 1% of the safety-critical tasks issued in that horizon miss their class-level application deadlines; otherwise, it is labelled 0. The horizon is selected to be long enough to capture service breaches that timely escalation could still prevent, and both the horizon and the 1% threshold are fixed before labelling begins. The operating threshold is then selected through seed-grouped cross-validation, with all state vectors originating from the same simulated scenario assigned to the same fold. The threshold is chosen to minimise unnecessary escalations while detecting a pre-specified minimum proportion of states that lead to service impact and remaining within the measured capacity of the shared Cloud-tier orchestrator. Calibration and evaluation use disjoint scenario seeds, and the fitted parameters and operating threshold are frozen before the final evaluation.

Escalation is governed by a hysteretic trigger on the action $a_t \in \{0, 1\}$ at update step t:

$$a_t = \begin{cases} 1 & \text{if } \mathrm{ES}_t \geq \tau_{\mathrm{high}}, \\ 0 & \text{if } \mathrm{ES}_t \leq \tau_{\mathrm{low}}, \\ a_{t-1} & \text{otherwise.} \end{cases} \tag{3}$$

with $0 < \tau_{\mathrm{low}} < \tau_{\mathrm{high}} < 1$.

The Cloud orchestrator is invoked only once ES rises above τ_{high} and is released only once it falls below τ_{low}; the deadband $[\tau_{\mathrm{low}}, \tau_{\mathrm{high}}]$ suppresses chattering when the state hovers near the boundary.

Once triggered, the agent transmits state vector, the relevant dependency subgraph, the active failure set, and a deterministically generated set of currently feasible actions. Two decoding constraints are applied to reduce sampling randomness and prevent structurally invalid outputs:

- **Greedy decoding:** The municipality-hosted LLM uses greedy decoding with the temperature set to zero.
- **Structural enforcement via constrained decoding:** The orchestrator enforces a strict JSON schema through token-level grammar constraints using the *Outlines* library. Outputs are restricted to valid escalation identifiers, target-agent identifiers, and the four defined orchestration primitives: `MIGRATE`, `SHED`, `RESTART`, and `NO_OP`.

The schema is regenerated from the current feasible-action set, so the grammar admits only action–target combinations whose preconditions hold at the start of the decision cycle. This provides structural and target validity at generation time, but it does not guarantee that the selected action is optimal or beneficial. Greedy decoding with temperature zero reduces sampling variability but is not treated as a guarantee of bitwise reproducibility across different model, hardware, or software configurations.

Because the system state may change during inference, the selected action is revalidated immediately before commitment. An action whose preconditions no

longer hold is rejected and logged with the corresponding reason code. Each decision cycle records the state, Escalation Score, feasible-action set, generated action, validation result, and committed action. These records make the escalation pathway reconstructible after an event without claiming to provide a faithful explanation of the LLM's internal reasoning. Action quality is assessed empirically through the fault-containment score (FCS) and the secondary service metrics. To bound the principal technical risk that Cloud-tier inference may exceed operational limits, the middleware integrates two mitigation strategies. First, Smart-Edge agents use an asynchronous local fallback: while awaiting the LLM response, they continue bounded local coordination and may shed best-effort services to prevent resource exhaustion. Second, every Cloud-tier decision cycle has a control-plane time-to-live (TTL), denoted t_{deadline}. If no valid action is committed before this deadline, the cycle ends in a deadline abort and control remains with the local heuristic fallback. The evaluation therefore tests both containment benefit and the frequency and latency cost of Cloud-tier decision cycles.

4 Proposed evaluation framework

Evaluating an architecture built for infrastructure collapse requires moving beyond steady-state testing towards a failure-first paradigm. Grounded in Design Science Research (DSR), the evaluation tests containment performance together with two governance properties relevant to municipal deployment: adherence to service-level agreements (SLAs) and a reconstructible decision pathway. Because intentionally triggering cascading failures on operational urban infrastructure is unsafe and unethical (Q. Li & Gao, 2026), the architecture is designed to be evaluated through discrete-event simulation mapped to open-data-derived urban dependency graphs.

4.1 Research hypotheses

To quantify the trade-off between the cost of failure and the cost of intelligence, the framework defines the following planned evaluation hypotheses; they are predictions to be tested rather than demonstrated findings:

- H_{1a} **(Cloud decision component):** Under compound failures, the full middleware achieves a higher FCS than a deterministic topology-aware Cloud-tier planner operating through the same escalation pathway.
- H_{1b} **(Complete pathway):** Under compound failures, the full middleware achieves a higher FCS than the heuristic-only strategy.
- H_2 **(Complexity scaling):** The Full–B4 FCS difference is more strongly associated with cross-service dependency depth than with the raw number of faults.
- H_3 **(Cost of intelligence):** In compound-failure scenarios where the Cloud-tier pathway remains reachable, fewer than 1% of initiated decision cycles end in a deadline abort.

All strategies are evaluated on matched, seed-replayed scenarios. H_{1a} isolates the Cloud-tier decision component through the Full-B4 contrast, H_{1b} evaluates the complete escalation pathway against B3, H_2 models the structural conditions under which the Full-B4 difference changes, and H_3 evaluates adherence to the control-plane deadline.

4.2 Empirical simulation grounding

To ensure the simulated environment reflects operational reality, the framework maps the fault taxonomy (silent sensor dropout, drift, link partition, and node overload) directly to empirical failure rates derived from real deployments. These distributions incorporate sensor data-integrity faults (Talayoglu et al., 2025), edge low-power wide-area network (LPWAN) link losses (Yasmin et al., 2020), backbone and Cloud-network failures (Talluri et al., 2026), and Cloud-tier server hardware failures (Xie et al., 2025). The service dependency graph is derived from a published open-data urban infrastructure model (Di Pietro et al., 2025). By applying Monte Carlo sampling over these empirical distributions and the urban topology, the simulation generates realistic, combinatorial fault intersections.

Testing an LLM orchestrator against latency deadlines requires incorporating measured inference delays into simulated control-plane timing. The simulator of choice is Yet (not) Another Intelligent Fog Simulator (YAIFS), developed by Lera and Guerrero (2026). Its Model Context Protocol (MCP) integration lets external LLM agents observe system state and inject orchestration actions. The evaluation records serialisation, network transfer, queueing, inference, validation, and dispatch time for each decision cycle.

4.3 Quantifying the core trade-off

The primary cost-of-failure metric is the FCS, adapted from Mohapatra (2025):

$$\text{FCS} = \frac{N_{\text{isolated}}}{N_{\text{total}}} \times (1-\delta), \qquad \text{FCS} \in [0,1], \quad (4)$$

where N_{total} is the injected ground-truth fault count, N_{isolated} is the number of faults contained within their originating district without degrading a neighbouring district or higher tier, and δ is the propagation-severity coefficient. Mohapatra, 2025 specifies the $(1-\delta)$ form but defines δ only qualitatively; we operationalise it on [0,1] as follows. For multiple injected origins, the denominator of δ is the union of nodes reachable from those origins in the pre-failure dependency graph; injected origin nodes are excluded and overlapping reachable

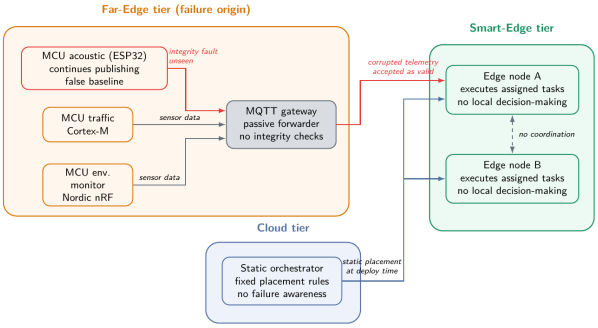


Figure 1. Baseline (AS-IS) architecture: the lack of active edge-tier coordination allows compound physical failures to cascade across tiers.

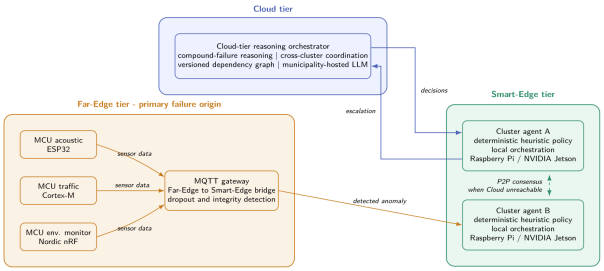


Figure 2. Proposed (TO-BE) three-tier agentic middleware, showing the escalation path from Far-Edge gateways to the Cloud-tier LLM orchestrator.

sets are counted once. The numerator contains the subset that remains degraded at the stabilised end of the scenario. The $(1-\delta)$ multiplier penalises strategies that allow uncontained faults to trigger widespread structural collapse.

To reflect mixed criticality, intentionally shed best-effort services are excluded from the affected-node set used for δ, but shedding is reported separately through shed-instance count and cumulative shed-service duration. Successfully migrated or rerouted services that remain operational are not counted as degraded; only unresolved secondary propagation contributes. A node counts as degraded only if an offline, unreachable, or task-breach condition persists for a full telemetry interval.

Secondary cost-of-failure metrics are critical-service uptime and task completion; cost-of-intelligence metrics are decision-cycle p_{95}/p_{99} latency, Cloud invocation, validation-rejection, and deadline-abort rates. End-to-end critical-service latency is reported separately.

4.4 Adversarial evaluation via threshold evasion

The protocol also includes a slow-onset threshold-evasion stress test. Faults are temporally staggered and constrained to states that keep the Escalation Score below τ_{high} until a service breach is imminent. This "Boiling Frog" cascade tests whether gradual structural degradation can evade the calibrated gate and identifies conditions under which escalation should occur earlier.

4.5 Baseline comparisons

The evaluation compares five strategies arranged as a capability ladder, each adding one mechanism over the previous. The ladder fixes the expected capability ordering for the hierarchy check, while B3 and B4 additionally function as targeted ablations of the full system:

- **B1 (Static orchestration):** Fixed service placement with no dynamic adaptation, included as the non-adaptive lower bound of the ladder.
- **B2 (Reactive rule-based):** Threshold-triggered single-node recovery, modelling traditional self-healing (Boluda-Prieto et al., 2026), included to locate where single-node recovery breaks down under compound collapse.
- **B3 (Heuristic-only ablation):** The proposed local deterministic layer with the Cloud pathway disabled, using resource-aware, TTL-bounded gossip and nearest-neighbour offloading (Y. Li et al., 2026).
- **B4 (Deterministic Cloud-tier planner):** The complete escalation pathway with the LLM replaced by a deterministic topology-aware greedy planner operating over the same payload, feasible-action set, Cloud resource, and decision deadline.

B3 tests the value of the complete escalation pathway, whereas B4 isolates the value added by the LLM decision component from the value of Cloud reachability, city-wide structural context, and escalation itself.

5 Discussion and conclusion

The proposed design separates fast edge heuristics from selective Cloud reasoning during compound smart-city failures. Its Escalation Score, constrained schema, pre-execution validation, and decision logging support a reconstructible escalation pathway, while the failure-first framework compares the *cost of failure* with the *cost of intelligence*.

The architecture's central trade-off is not simply whether Cloud-tier reasoning is more capable than local rules, but whether the structural novelty of a failure justifies delayed action from the Cloud tier. The planned hypotheses make that trade-off falsifiable. A Full–B4 FCS improvement accompanied by frequent deadline aborts would indicate that LLM-based reasoning can help containment but is unsuitable for the most time-critical decisions; acceptable decision-cycle latency without improved containment would suggest that the LLM adds little beyond deterministic Cloud planning. H_2 is especially important for scope: an advantage associated with dependency depth rather than fault count would support the smart-city motivation based on cross-operator dependencies and correlated physical events. Negative results would still identify conditions under which Cloud escalation or LLM-based planning should be disabled.

This design contribution reports no empirical results; its hypotheses remain planned tests. Next, the Escalation Score will be calibrated on held-out data and evaluated in YAIFS against the four baselines, including threshold-evasion cascades and measured Cloud-tier decision delays. Later work may examine lightweight edge RL and generate-and-score verification of orchestration actions (cf. Gort et al., 2025; Zhao & Han, 2025).

Acknowledgements

This work is fully supported under FOI-IIP-2025-13.

References

Abbass, W., Abbas, N., & Majeed, U. (2026). Neuroscale: Scalable and constraint-aware neuro-symbolic optimization for edge-centric IoT systems. *The Journal of Supercomputing*, *82*(3), 127.

Akbari, N., Grundy, J., Cheema, A., & Toosi, A. N. (2025). IntentContinuum: Using LLMs to support intent-based computing across the compute continuum. *2025 IEEE International Conference on Web Services (ICWS)*, 573–583.

Alqerm, I., Cheng, N., & Pan, J. (2026). REPA: Resilient and proactive resource allocation for last-mile heterogeneous edge-IoT systems. *IEEE Internet of Things Journal*, *13*(2), 1988–2005.

Angelis, A., & Kousiouris, G. (2025). An overview on the landscape of self-adaptive cloud design and operation patterns. *Future Internet*, *17*(10), 434.

Antuley, U., Siddiqui, S., Hameed, S., Arif, W., & Shah, S. A. (2026). SORA-ATMAS: Adaptive trust management and multi-LLM aligned governance for future smart cities. *Knowledge-Based Systems*, *337*, 115403.

Babaiyan, V., Bushehrian, O., & Javidan, R. (2026). A bi-level mobility-aware deep reinforcement learning approach for fault-tolerant task offloading in vehicular edge-cloud computing. *Scientific Reports*, *16*(1), 11703.

Balasubramanian, C., Kareemullah, H., Arunkumar, R. S., Kumar, P. S., & Revathi, S. M. (2026). Multi-objective deep Q-self-attention transformer for PM2.5 prediction on edge devices. *GeoInformatica*, *30*(1), 11.

Belcastro, L., Marozzo, F., Orsino, A., Talia, D., & Trunfio, P. (2026). Navigating the edge-cloud continuum: A state-of-practice survey. *IEEE Access*, *14*, 40622–40647.

Boluda-Prieto, M., Mateo-Casali, M. A., Fraile, F., & Alarcon, F. (2026). Resilient edge-to-cloud architecture with self-healing and self-correcting mechanisms for industrial data continuity. *Computers & Industrial Engineering*, *213*, 111795.

Caldeira, I. D. B., Choren, R., & Salles, R. (2026). The opaque autonomy paradox in sovereign and intelligent SD-WAN: A systematic literature review. *Computer Networks*, *282*, 112246.

Chaturvedi, P., Ahmad, S., & Mewada, A. (2026). A comprehensive survey on fog computing: Architectures, techniques, challenges, and future directions. *Archives of Computational Methods in Engineering*.

Chen, S., Li, D., & Mohajer, A. (2025). QoS-aware multi-agent DDPG for adaptive edge service distribution in intelligent wireless communication networks. *Ain Shams Engineering Journal*, *16*(9), 103543.

Chochliouros, I. P., Pages-Montanera, E., Alcázar-Fernández, A., Zahariadis, T., Velivassaki, T.-H., Skianis, C., Rossini, R., Belesioti, M., Drosos, N., & Bakiris, E. (2023). NEMO: Building the next generation meta operating system. *Proceedings of the 3rd Eclipse Security, AI, Architecture and Modelling Conference on Cloud to Edge Continuum*, 1–9.

Di Pietro, A., Cavedon, F., Stergiopoulos, G., & Rosato, V. (2025). Modeling fault dynamics of interdependent assets in complex urban environments using open-data. *SN Computer Science*, *6*(6), 564.

Foerster, J., Farquhar, G., Afouras, T., Nardelli, N., & Whiteson, S. (2018). Counterfactual multi-agent policy gradients. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1).

Gkonis, P. K., Giannopoulos, A., Nomikos, N., Sarakis, L., Nikolakakis, V., Patsourakis, G., & Trakadas, P. (2026). A survey on the computing continuum and meta-operating systems: Perspectives, architectures, outcomes, and open challenges. *Sensors*, *26*(3), 799.

Gort, B., Kibalya, G. M., & Antonopoulos, A. (2025). *AgentEdge: Agentic AI for service orchestration in the edge-cloud continuum* [TechRxiv preprint].

He, J., Zheng, Y., Lei, S., Liao, F., & Wu, L. (2026). Cyber–physical intrusion detection for coordinated renewable energy and storage systems. *Array*, *30*, 100765.

Kirk, R., Zhang, A., Grefenstette, E., & Rocktäschel, T. (2023). A survey of zero-shot generalisation in deep reinforcement learning. *Journal of Artificial Intelligence Research*, *76*, 201–264.

Kuchuk, H., Kosenko, N., Kuchuk, N., Gopejenko, V., & Kosenko, V. (2026). Adaptive resource allocation method for the mobile fog layer of high-density industrial internet of things in industry 5.0 networks. *Innovative Technologies and Scientific Solutions for Industries*, (1(35)), 65–78.

Lera, I., & Guerrero, C. (2026). *YAIFS: Yet (not) another intelligent fog simulator* [arXiv:2604.19181].

Li, Q., & Gao, K. (2026). Beyond the Comfort Zone: A Review and Gap Analysis of Fuzzing in Smart City IoT Ecosystems. *Information*, *17*(3), 218. https://doi.org/10.3390/info17030218

Li, Y., Liu, F., Chang, T.-Y., Hsu, C.-H., & Venkatasubramanian, N. (2026). Microservice provisioning and event-driven adaptation in heterogeneous IoT settings. *Pervasive and Mobile Computing*, *118*, 102195.

Liu, Q., Yang, J., & Yan, Z. (2025). Dynamic resource orchestration in edge computing environments using multi-agent reinforcement learning. *Knowledge and Information Systems*, *67*(10), 9363–9383.

Mohapatra, H. (2025). A comprehensive review on urban resilience via fault-tolerant IoT and sensor networks. *Computers, Materials & Continua*, *85*(1), 221–247.

Talayoglu, B., Vande Velde, J., & Da Silva, B. (2025). Lightweight AI for Sensor Fault Monitoring. *Electronics*, *14*(22), 4532. https://doi.org/10.3390/electronics14224532

Talluri, S., Niewenhuis, D., Chu, X., Kyselica, J., Cetin, M., Balgavy, A., & Iosup, A. (2026). Cloud Uptime Archive: Open-Access Availability Data of Web, Cloud, and Gaming Services. *IEEE Transactions on Parallel and Distributed Systems*, *37*(5), 1136–1152. https://doi.org/10.1109/TPDS.2026.3673519

Vlahek, D., & Mongus, D. (2023). An efficient iterative approach to explainable feature learning. *IEEE Transactions on Neural Networks and Learning Systems*, *34*(5), 2606–2618. https://doi.org/10.1109/TNNLS.2021.3107049

Wang, Y., Guo, S., Pan, Y., Su, Z., Chen, F., Luan, T. H., Li, P., Kang, J., & Niyato, D. (2026). Internet of agents: Fundamentals, applications, and challenges. *IEEE Transactions on Cognitive Communications and Networking*, *12*, 4476–4501.

Wei, Q., Ning, H., Shi, F., Zhu, T., Ding, J., Derhab, A., & Daneshmand, M. (2026). AGI-enabled solutions for service explosion in IoX: A cyber-physical-social-thinking perspective. *ACM Computing Surveys*, *58*(8), 1–32.

Xie, H., Zhou, M., Yu, Q., Yu, J., Sheng, Z., Xie, H., & Lian, D. (2025). M^2-MFP: A Multi-Scale and Multi-Level Memory Failure Prediction Framework for Reliable Cloud Infrastructure. *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, 5083–5094. https://doi.org/10.1145/3711896.3737243

Xiong, Z., Meng, S., Yao, S., & Luo, L. (2025). Fault tolerant and intelligent reliability assurance in cloud-edge systems. *2025 12th International Conference on Dependable Systems and Their Applications (DSA)*, 156–162.

Xu, R., Nagothu, D., & Chen, Y. (2024). AR-Edge: Autonomous and Resilient Edge Computing Architecture for Smart Cities. In Y. Chen & R. Xu (Eds.), *Edge Computing Architecture - Architecture and Applications for Smart Cities*. IntechOpen. https://doi.org/10.5772/intechopen.1005876

Xu, T., Zhang, D., Mitra, K., & Hruschka, E. (2026). Verification-Aware Planning for Multi-Agent Systems. *EACL 2026*. https://doi.org/10.18653/v1/2026.eacl-long.353

Yasmin, R., Mikhaylov, K., & Pouttu, A. (2020). LoRaWAN for Smart Campus: Deployment and Long-Term Operation Analysis. *Sensors*, *20*(23), 6721. https://doi.org/10.3390/s20236721

Zhang, G., Zhang, Q., Feng, L., Zhou, F., Donta, P. K., & Dustdar, S. (2026). Task-aware collaborative inference and fine-grained DNN partitioning in MEC networks. *IEEE Transactions on Mobile Computing*, 1–17.

Zhao, Y., & Han, Z. (2025). Safe and adaptive cloud healing: Verifying LLM-generated recovery plans with a neural-symbolic world model. *Proceedings of the 7th International Academic Exchange Conference on Science and Technology Innovation (IAECST)*, 89–94.

Evaluating event influence on TGN performance in bike sharing station occupancy forecasting

Štefan Horvat, Damjan Strnad, Gregor Horvat, Matej Brumen, Domen Mongus
University of Maribor
Faculty of Electrical Engineering and Computer Science
Koroška cesta 46, 2000 Maribor, Slovenia
{stefan.horvat, damjan.strnad, g.horvat, matej.brumen, domen.mongus}@um.si

Abstract. *Station occupancy forecasting is an important task in bike sharing networks, particularly for anticipating when the stations are empty or full. Trips and redistribution events form a stream of timestamped interactions used to encode station memories. This paper investigates the influence of events is investigated using a hybrid TGN architecture in conjunction with temporal feature encoders used to produce per-station forecasts. Empirical results indicate that event memory improves model performance, but more complex event structure variants underperform the simpler hybrid TGN model. The study highlights the value of event streams while emphasizing the need for caution when adding model complexity.*

Keywords. machine learning, temporal graph network, bike sharing network, event stream

1 Introduction

Bike sharing system changes state over time and station connections are created by user trips and redistribution operations. Most forecasting models try to capture patterns in station time series events in conjunction with static graph structures, where edges usually represent proximity, historical correlations or feature aggregates. This approach captures spatial dependencies but can miss or lack informational structure that can be pres ent in the interaction events between the nodes (Rossi et al., 2020). Better forecasts can support more sustainable network operations by accurate predictions that could reduce the unnecessary service trips.

In this work we study the role of windowed event memory streams for a bike sharing network in Barcelona. The goal is to produce short term forecasts of available station slots. These could be used to optimise the operational structure for easier scheduling of station maintenance and allow better responses when station operational limits are reached. Instead of treating the problem as a snapshot of fixed graphs over time, we employed timestamped events between station which represent trips and redistribution operations. These events are fed through a Temporal Graph Network (TGN) encoder, allowing each station to maintain a compact representation of historic interactions. Instead of using graph convolution over a static graph, the TGN model maintains and updates station memory from a continuous event stream.

The study combines multiple individual components to examine the role of continuous event flow in bike sharing station occupancy forecasting. Besides the temporal baseline, a hybrid TGN based forecasting model is tested that combines station history and event memory representation. Additionally two model variants are tested in the study, a GraphSAGE based neighbourhood aggregation model over event driven connections, and a dual memory hybrid TGN model that maintains a separate memory based on the event type. The results indicate that event information improves forecasting, and that a compact event memory formulation can outperform more complex hybrid TGN variants.

2 Related Work

Forecasting in bike sharing networks is commonly studied as a station level spatio-temporal prediction problem. Most of the literature emphasizes that station states depend on temporal patterns, spatial relationships between stations, and external factors, such as weather, events, and station imbalance. Station level demand predictions can aid the process of redistribution and produce accurate forecasts for unmet demand while considering external factors. Similarly, in excess demand prediction rebalancing is framed as a response to stations becoming empty or full, which motivates forecasting station occupancy rather than only aggregate demand (Hulot et al., 2018).

Bike sharing networks are usually presented as graphs composed of nodes that represent stations, and edges, which can represent a wide variety of relationships, such as spatial proximity or historical flows. There were several attempts to try and extract graph patterns and train the models to predict node states or flows. One of them uses a graph convolutional model to learn hidden pairwise correlations between the stations (Lin et al., 2018). STG2Vec model learns hetero-

geneous spatial temporal representations from demand, geographic context, historic flow, and external events (Li et al., 2019). TAGCN combines graph convolution and temporal attention to predict station inflow and outflow (Zi et al., 2021). These works show, that graph based structures can improve bike demand modeling, but graphs are usually presented as aggregation of graph features in discrete time intervals referred to as discrete-time dynamics graphs (DTDG), rather than a stream of continuous timestamped events which avoids the aggregation of graph features and possibly produce more accurate predictions (Kazemi et al., 2020).

Spatio-temporal graph neural networks were also widely adopted in other transportation problems. DCRNN and STGCN models combine temporal modeling with graph based spatial propagation, which enables capturing patterns across traffic sensors connected via the transport network (Li et al., 2017; Yu et al., 2018). Other techniques, such as GraphSAGE, allow capturing localised relationships patterns with a model that learns node representations by aggregating features from local neighbourhoods (Hamilton et al., 2017).

Temporal graph learning is an alternative to spatio-temporal graph network modeling which avoids explicit construction of graphs and instead uses event processing. A step towards event processing was first made by JODIE, which models temporal interaction networks by learning dynamic embedding trajectories for interacting entities (Kumar et al., 2019). This was followed by DyREP, which models dynamic graphs through temporal point processes capturing both structural evolution and interaction activity (Trivedi et al., 2019). More general modular framework, based on node memory, message functions, aggregation, and temporal embedding, is the TGN model (Rossi et al., 2020). TGN maintains node memories, which are updated by temporal events, thus allowing each node to summarize its history in a compact way. This maps to the problem in this paper, where trips and redistribution events can be interpreted as timestamped interactions between stations.

3 Method

This section consists of two parts. In the first part the problem and data is described, followed by the description of a hybrid TGN architecture.

3.1 Problem Formulation

The main goal of the experiment is to evaluate whether modeling station to station event structure aids the model knowledge for predicting the number of available docks in the Bicing bike sharing network in Barcelona (Bicing Barcelona, 2026). The events exhibit daily and weekly patterns so the input window was set to $T_{in} = 168$ hours. We were only interested in short term predictions (next hour) so $T_{out} = 1$ hour. Every station is described by 8 features, of which six are temporal and two are static (see Table 1). Hourly features are normalised to interval $[-1, 1]$ using a sine transform, while the remaining features are normalised to $[0, 1]$. To integrate static features, each static attribute was repeated for every time step in the input window. This resulted in a unified time series representation across the participating stations. The objective is to forecast the number of available docks per bike station and is treated as a regression task.

Table 1. Node features with ranges before normalisation. Features marked in italic represent static features.

Feature	Range
Available docks	0 - 50
Holiday	0 or 1
Weekend	0 or 1
Hour of day	1 - 24
Hour of week	1 - 168
Hour of year	1 - 8760
Capacity	2 - 50
Altitude	0 - 175

The stations are interconnected via a continuous stream of events, which represent bike movements. An event can be represented as $e_{ij}(t) = (src, dst, tstmp, feat)$. The src and dst contain information about the participating stations. The $tstmp$ represent the event timestamp and $feat$ represents event features. In our case the event type is classified either as a trip, which represents flow produced by the users or a redistribution, which represents flow caused by human interventions in the system.

3.2 Hybrid TGN Model

The main component of the hybrid model is the TGN encoder that is applied on continuous-time dynamic graphs (CTDG), which are general representations of timed lists of events (Kazemi et al., 2020). They are modeled as a sequence of time-stamped events producing, for each time t, the embeddings of the graph nodes $Z(t) = (z_1(t), ..., z_n(t))$ (Rossi et al., 2020). The concept is based on event streams, which describe the interaction of two nodes at a specified time t. The TGN stores its internal node representations using a vector referred to as memory $s_i(t)$. Node memory is a learned hidden state involving the node in question and keeps track of the nodes history in a compact format. When $e_{ij}(t)$ arrives, two messages are generated, one for source node $m_i(t)$ and the other for the target node $m_j(t)$. A message for the source node is constructed using Equation 1:

$$m_i(t) = msg_i(s_i(t^-), s_j(t^-), \Delta t, e_{ij}(t)) \quad (1)$$

where $s_i(t^-)$ is the memory of node i before time t, Δt represents elapsed time of the event and msg_i is a learnable message function. Processing in batches usually involves multiple events using the same node i. In that case batch message aggregation function is used $\bar{m}_i(t) = agg(m_i(t_1), ..., m_i(t_b))$. Choices for agg can range from simple non-learnable function such as most recent message and mean message or complex learnable functions such as attention or recurrent neural networks. This is followed by an update of the node memory using Equation 2:

$$s_i(t) = mem(\bar{m}_i, s_i(t^-)) \tag{2}$$

To avoid the memory staleness problem, the embedding module is used to generate temporal embedding $z_i(t)$ of node i by sampling nodes reachable within k graph hops before time t using Equation 3:

$$z_i(t) = \sum_{j \in_i^k (0,t)} h(s_i(t), s_j(t), e_{ij}) \tag{3}$$

where h is a learnable function and is implementation defined. In our experiments $h = emb(i, t)$, so we used memory of node i before time t to produce the embeddings (Rossi et al., 2020). An illustration of the TGN encoder can be seen in Fig. 1.

To utilise temporal features of stations a hybrid TGN is proposed. The TGN model is designed to utilise the events, but we also wanted to utilise temporal node features. The hybrid TGN model is extended with a separate TimeConvEncoder for encoding station temporal data, the TGN memory is used for encoding event history representation and a small FutureEncoder is utilised to include known-ahead temporal covariates that describe the forecast timestamp, such as calendar and cyclic time features, and are available before the target station occupancy is observed. The outputs of all three encoders are concatenated to produce station state forecats. An illustration of the pipeline can be seen in Fig. 2.

The TGN encoder can be upgraded to include local node awareness using k-hop expansion, where k specifies the number of local neighbourhoods to be searched. We used GraphSAGE as a neighborhood aggregation layer over the node representations produced from TGNEncoder, TimeConvEncoder, and FutureEncoder. In some cases, the event stream might have a wider local effect, so it makes sense to explore the graph neighbourhood and aggregate the results. GraphSAGE enables learning of node embeddings by aggregating feature information from local graph neighbourhoods producing nonlinear aggregations (Hamilton et al., 2017). A similar effect can be reached by replacing h in Equation 3 with Temporal Graph Attention (ATTN) embedding block (Rossi et al., 2020).

4 Results

There are 548 statios evaluated in the data (Fig. 3), but the RMSE is evaluated only on 501 stations. The remaining 47 stations have special statuses, were closed down, have incomplete data or were reorganised, but can still participate in the flow. The dataset covers the period between January 2023 and March 2025, so there is 2 years and 3 months of data available. The events exhibit daily, weekly and yearly patterns which were confirmed with spikes on corresponding lags when the data was examined using the autocorrelation function (ACF).

After slicing the dataset into windows of size $T_{in} + T_{out}$ we obtained 27147 windows. There were 102410 unqiue origin-destination pairs across the full dataset, including approximately 2536 redistribution origin-destination pairs. The first 80 % of windows were used as a training dataset, the next 10 % was used to represent validation dataset, while the remaining 10 % was used to represent the test datasets. After some preliminary testing on the split dataset, we settled with $T_{in} = 168$ because longer windows did not present any improvements while shorter windows negatively affected the forecasts. We were only interested in next hour predictions, so $T_{out} = 1$. Preliminary feature ablations showed that the retained feature set produced the lowest validation RMSE among the tested configurations so all trained models use the same feature representation - Table 1.

The models were implemented in Python 3.11 using Pytorch Geometric and trained on cuda capable device Nvidia GeForce 3060 12 GB GPU. The hyperparameters were chosen based on experimental results feedback, where we prioritised RMSE and training time.

Each model was trained three times with different random intialisations, and the reported results are the means performance over the three runs. Training was run for a maximum of 10 epochs using the Adam optimiser with a learning rate $\eta = 0.001$. During training, validation performance was monitored and the checkpoint with the lowest RMSE was retained for final evaluation on the test set. Best checkpoint was always reached before the full 10 epochs, so the report results are based on early selected model states.

The hybrid TGN model consists of 3 encoders (TimeConvEncoder, TGNEncoder, FutureEncoder) and 1 decoder (OutputDecoder). It consists of two convolutional layers, each followed by a ReLU activation function. After the final activation, the resulting feature representations are passed through mean pooling and a linear layer, which outputs a latent representation with the dimension of 64. The FutureEncoder is used for known future context extraction and consists of a linear layer with a ReLU activation function, followed by another linear layer that produces an embedding of size 16. The function of the TGN encoder was discussed in Section 3.2 and uses events. Finally, the embeddings

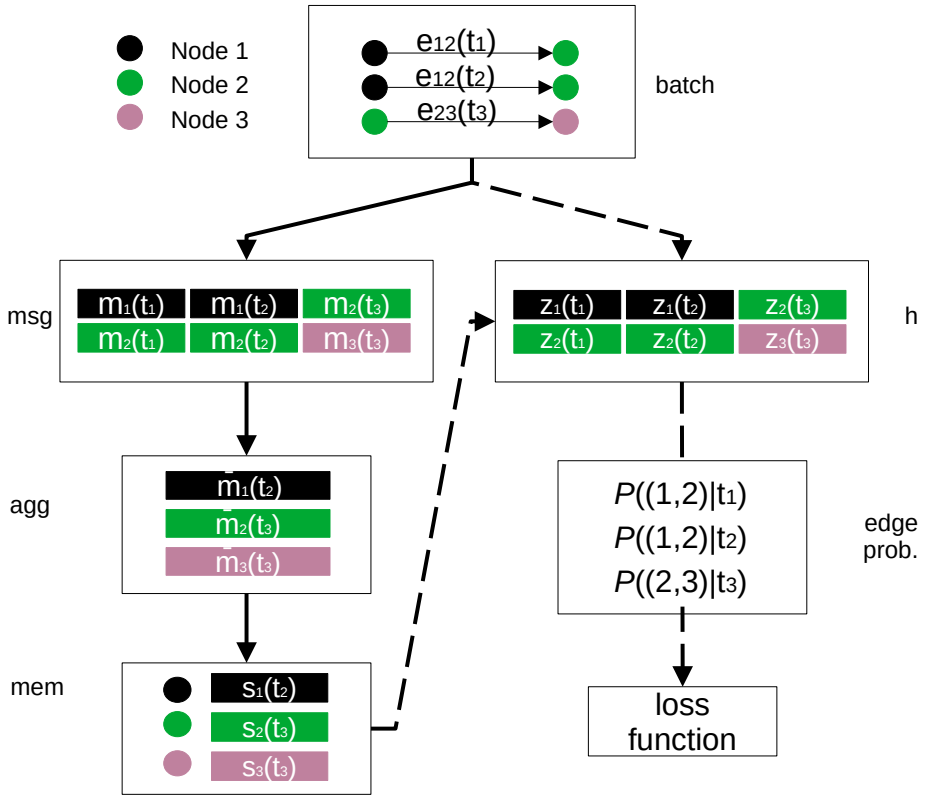


Figure 1. Simplified illustration of a TGN encoder module. Multiple events usually involve the same nodes in the batch. Firstly the left branch (solid black line) gets executed, where event representations are transformed to compact messages. Because a single node can be presented with only one message within the batch, the messages need to be aggregated. In this example the agg always takes the last message. Finally, the node memories get updated. Secondly the right branch gets executed (dashed black line) which creates embeddings using current node events and nodes historic context from memory. The embeddings are used to make edge probability predictions in given t. This is fed to the loss function, which produces gradients that are used to update trainable parameters.

from TimeConvEncoder, FutureEncoder and TGNEncoder are concatenated and passed through the OutputDecoder, which consists of a linear layer followed by a ReLU activation. The last layer is again linear and produces the final prediction of station occupancies in the next hour.

To determine whether using the event processing pipeline improves model performance, a baseline model was designed first that operates solely using per station features. The baseline model consists of TimeConvEncoder, FutureEncoder, and OutputDecoder, and achieved an RMSE of 3.07. In the second experiment, we utilised the hybrid TGN to include the events, but the neighbourhood search was not included is resulted in an improved RMSE of 2.99.

We speculated that including context from local neighbourhoods could have an impact on the model performance, because a flow from one station can influence the flow at the other nearby stations. For instance, if a station was full, the overflow would be redirected to neighbouring stations. We experimented with inclusion of direct, 2-hop and 3-hop neighbourhoods. The best results were achieved by direct neighbourhood integration, that achieved an RMSE of 3.02, but fell short of beating the model from the previous experiment.

In the final experiment we addressed the problem of exogenous shocks in the form of redistribution events. In the previous experiments a flag was included to indicate whether an event represented a redistribution or a trip. We assumed, that the model might get confused if we feed the events to it in this fashion so the idea of a dual hybrid TGN model emerged. The idea was to group events by their type (trips or redistribution). In the last experiment we therefore trained a dual hybrid TGN model, which maintained two independent memories, one for redistribution and the other for trips. For the final prediction, the two memories were concatenated together. However, the performance degraded further, and the model achieved an RMSE of 3.04. The summary of experiments can be found in Table 2.

The results indicate that stations, own temporal and static features probably do most of the work. Adding the events improves the performance slightly, so TGN provides some context signal learned from the events.

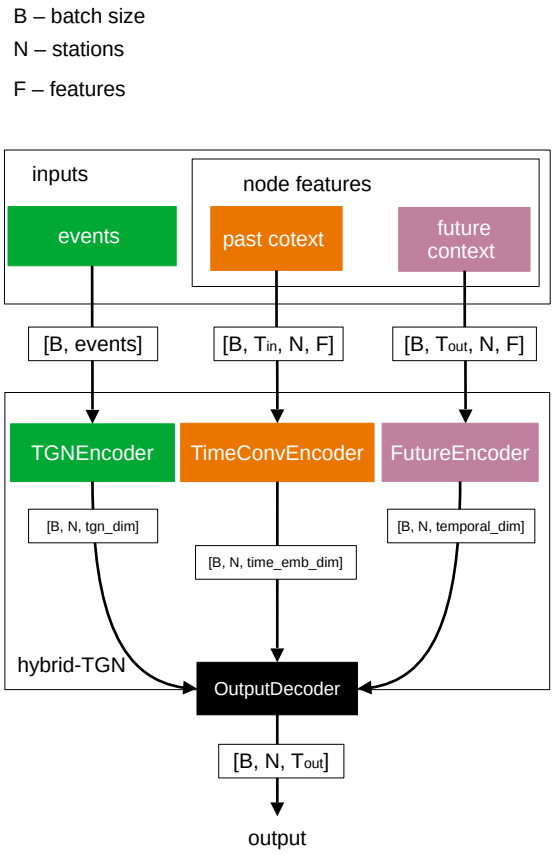


Figure 2. Illustration of a Hybrid TGN model, composed of three encoders. The TimeConvEncoder encodes station history, TGN encodes events, and FutureEncoder encodes known future context. Outputs from encoders are concatenated along the temporal axis and are fed to the OutputDecoder, which produces a forecast for the given horizon T_{out}.

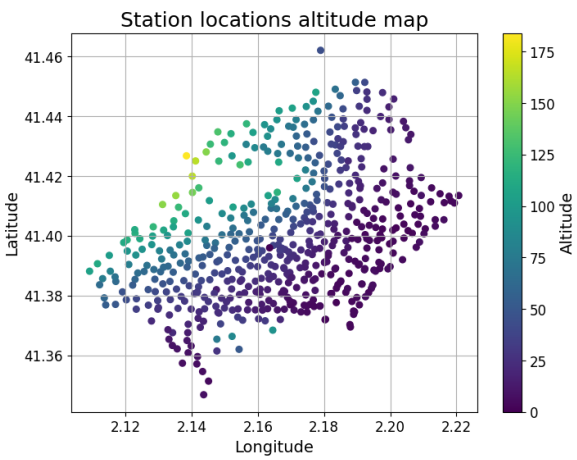


Figure 3. Bike sharing network for the city of Barcelona. The colormap represents station altitudes.

Including the local neighborhood aggregation and dual memory, on the other hand, resulted in worse performance than the simple TGN, but both remained slightly better than the temporal baseline. This indicates, that event information is useful but using more complex models may increase the noise or make optimization harder so extra architectural complexity is currently not helpfull.

Table 2. Model experiments indicate, that all TGN model variants benefited from the additional event information.

Model	No. of params	RMSE
Baseline	30000	3.07
TGNMemory	78769	**2.99**
TGNKHop	142897	3.02
TGNDualMemory	190161	3.04

The hybrid TGN introduces some additional computational overhead compared to the temporal baseline model because the event stream must be loaded, mapped to station indices, and used to update node memories for each training window. This cost is manageable for the direct event memory formulation, but becomes more pronounced in the TGNKHop variant, where events are expanded into local neighbourhoods before training. This presents a big computational bottleneck and increased the preprocessing time, while not improving RMSE over the simpler memory based TGN.

5 Conclusion

This paper reports the results of a case study on how useful temporal event patterns can be to forecasting of bike station occupancy. The experiments indicate that additional model complexity may not automatically improve performance. The simplest hybrid TGN model without neighbourhood aggregation achieved a lower RMSE compared to the baseline temporal convolutional network, which indicates that trip and redistribution events contain some additional patterns that can positively contribute to model performance, although most of the structure was explained using the temporal and static features of bike stations. The simplest hybrid TGN approach, which combined station temporal encoding, event encoding, future encoding and the output decoder, achieved the lowest RMSE. More complex model variants that added the GraphSAGE neighbourhood aggregation and dual memory redistribution and trips, fared worse.

Overall, the results suggest that event memory is valuable, but the most effective use of events is a TGN memory representation rather than explicit graph neighborhood aggregation or memory separation. Future work should therefore focus less on increasing architectural complexity and more on understanding how temporal memory is updated, how event types con-

tribute to predictions, and under which station or time conditions event information provides the greatest benefit.

Better forecasts could lead to reduction of unnecessary redistribution trips that could affect the emissions, noise, and operational wear. Future work should evaluate whether event-aware forecasts can reduce such trips through improved redistribution optimization, and also strengthen the evaluation protocol in three directions: reporting variability across random seeds, complementing RMSE with MAE and threshold-based empty/full station metrics, and using seasonal or rolling temporal holdouts to test robustness beyond a single final chronological split. These additions would better align the evaluation with the operational goal of identifying critical station states.

Acknowledgments

This research has received funding from the Slovenian Research and Innovation Agency under Research Project J7-70247. We would also like to thank Serveo for providing the Bicing bike sharing network dataset.

References

Bicing Barcelona. (2026). *Bicing, el servei de bicicleta compartida de barcelona*. Retrieved June 10, 2026, from https://www.bicing.barcelona/

Hamilton, W. L., Ying, R., & Leskovec, J. (2017). Inductive representation learning on large graphs. *Advances in Neural Information Processing Systems*, *30*.

Hulot, P., Aloise, D., & Jena, S. D. (2018). Towards station-level demand prediction for effective rebalancing in bike-sharing systems. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 378–386. https://doi.org/10.1145/3219819.3219873

Kazemi, S. M., Goel, R., Jain, K., Kobyzev, I., Sethi, A., Forsyth, P., & Poupart, P. (2020). Representation learning for dynamic graphs: A survey. *J. Mach. Learn. Res.*, *21*(1).

Kumar, S., Zhang, X., & Leskovec, J. (2019). Predicting dynamic embedding trajectory in temporal interaction networks. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1269–1278. https://doi.org/10.1145/3292500.3330895

Li, Y., Yu, R., Shahabi, C., & Liu, Y. (2017). Graph convolutional recurrent neural network: Data-driven traffic forecasting. *CoRR*, *abs/1707.01926*. http://arxiv.org/abs/1707.01926

Li, Y., Zhu, Z., Kong, D., Xu, M., & Zhao, Y. (2019). Learning heterogeneous spatial-temporal representation for bike-sharing demand prediction. *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*. https://doi.org/10.1609/aaai.v33i01.33011004

Lin, L., He, Z., & Peeta, S. (2018). Predicting station-level hourly demand in a large-scale bike-sharing network: A graph convolutional neural network approach. *Transportation Research Part C: Emerging Technologies*, *97*, 258–276. https://doi.org/https://doi.org/10.1016/j.trc.2018.10.011

Rossi, E., Chamberlain, B., Frasca, F., Eynard, D., Monti, F., & Bronstein, M. M. (2020). Temporal graph networks for deep learning on dynamic graphs. *CoRR*, *abs/2006.10637*. https://arxiv.org/abs/2006.10637

Trivedi, R., Farajtabar, M., Biswal, P., & Zha, H. (2019). Dyrep: Learning representations over dynamic graphs. *International Conference on Learning Representations*. https://openreview.net/forum?id=HyePrhR5KX

Yu, B., Yin, H., & Zhu, Z. (2018). Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, 3634–3640. https://doi.org/10.24963/ijcai.2018/505

Zi, W., Xiong, W., Chen, H., & Chen, L. (2021). Tagcn: Station-level demand prediction for bike-sharing system via a temporal attention graph convolution network. *Information Sciences*, *561*, 274–285. https://doi.org/https://doi.org/10.1016/j.ins.2021.01.065

From industrial data to business decisions: Industrial case studies as teaching instruments in interdisciplinary analytics education

Dino Vlahek, Sandra Lovrenčić, Dijana Oreški, Nina Begičević Ređep, Tena Jagačić
University of Zagreb Faculty of Organization and Informatics
Pavlinska 2, 42000 Varaždin, Croatia
{dvlahek, sandra.lovrencic, dijana.oreski, nina.begicevic, tjagacic}@foi.unizg.hr

Renato Punčec
Emil Frey Digital d.o.o.
Ulica Ivane Brlić Mažuranić 28, Varaždin
renato.puncec@emilfreydigital.hr

Hrvoje Aničić, Marko Mrkonjić
SICK Mobilisis Hrvatska
Varaždinska ulica - Odvojak II 7 Varaždin – Jalkovec
{hrvoje.anicic, marko.mrkonjic}@sick.com

Abstract. *This paper analyses industrial problems as teaching instruments in interdisciplinary analytics education, based on 14 student project reports from two industrial case studies: SICK Mobilisis (AI-based fuel-event detection from vehicle telemetry data) and Emil Frey Digital (digital vehicle sales optimisation). Reports were analysed using a qualitative coding framework covering business-process understanding, analytical-unit construction, target design, feature availability, validation, metric interpretation, decision support, implementation, and risk. Students usually completed expected sections but had more difficulty connecting them into a coherent process from data to decision, particularly regarding analytical-unit definition, leakage-aware feature selection, validation design, and linking model outputs to business decisions.*

Keywords. data science education, business analytics, work-based learning, project-based learning, industrial case studies, decision support, data leakage, AHP

1 Introduction

Data science and business analytics education often introduces students to predictive models, evaluation metrics, and standard data-preparation procedures. These elements are necessary, but they are not sufficient for applied industrial analytics (Brüngel, Rückert, & Friedrich, 2020; De Veaux et al., 2017; Ramasamy & Ramamoorthy, 2023). In real projects, the analytical task is rarely given in a clean and predefined form. Students have to understand the process that generates the data, define the analytical unit, construct a meaningful target variable, identify which features are available at the moment of decision, choose an appropriate validation design, and interpret model results in relation to business risk or economic value. This corresponds to the broader logic of end-to-end analytics processes, where business understanding, data understanding, data preparation, modelling, evaluation, and deployment-oriented interpretation are treated as connected phases rather than isolated technical steps (Chapman et al., 2000).

This problem is especially visible in project-based and case-based data science education. Industrial projects can help students develop practical skills, teamwork, and problem-solving abilities that are difficult to obtain from isolated algorithmic exercises (Brüngel et al., 2020; Ramasamy & Ramamoorthy, 2023). However, an authentic industrial problem does not by itself ensure methodological understanding. A student team may build a model and report standard evaluation metrics even when the target does not correspond to the business decision, the feature set contains information unavailable at prediction time, or the validation design does not reflect future use. The central limitation is then the transition from the business process to the analytical task rather than model selection alone. Work-based learning (WBL) provides a useful lens for this setting because it connects academic learning with professional practice through real-world and industry-oriented activities (Perusso & Wagenaar, 2021; Rienties, Divjak, Iniesto, Pažur Aničić, & Zizak, 2023).

This paper examines the transition from an educational setting to a work-oriented analytics setting through two industrial case studies used in an interdisciplinary project-based course. The student teams included students from information systems and business/economics study programmes, which made the projects suitable for analysing how technical modelling, business-process understanding, and decision-support reasoning are connected. The first case, provided by SICK Mobilisis, involved sensor-based ve-

hicle telemetry data for AI-based fuel-event detection. The second case, provided by Emil Frey Digital, involved business and sales-process data for the optimisation of digital vehicle sales through analytical recommendations. The two cases differ in domain and data structure. The SICK Mobilisis case required students to reason about temporal sensor data, rare events, false alarms, missed detections, and the difference between row-level classification and event-level detection. The Emil Frey Digital case required students to reason about a sales funnel, linked business tables, marketplace attribution, target construction, return on investment, conversion, and the connection between machine learning outputs and multi-criteria decision-making. A central difficulty in both cases is that the analytical dataset is not given. It has to be constructed from the available data and aligned with the intended decision. If features are derived from later stages of the process, or if the same entity appears in both training and test data, model performance can become overly optimistic because of leakage (Apicella, Isgrò, & Prevete, 2025; Kapoor & Narayanan, 2023; Sasse et al., 2025). Therefore, leakage and validation problems are not only technical mistakes. In student projects, they also indicate the extent to which students understand the temporal, entity-based, and process-based structure of the industrial problem.

The decision-support part introduces a second difficulty. Predictive models estimate classes, probabilities, or scores, but these outputs still have to be translated into alternatives, criteria, risks, and business actions. Multi-criteria decision-making methods such as the Analytic Hierarchy Process (AHP) can support this translation because they require explicit criteria and priorities (Saaty, 1980). However, AHP can also remain detached from the predictive model if the criteria are not linked to model outputs, feature importance, business risk, or economic value. This is particularly important in interdisciplinary analytics education, where students are expected to connect information-systems modelling with economic decision-making (Pang & Quek, 2026).

The main argument of this paper is that real industrial case studies are useful not only because they provide authentic data, but because they expose the methodological transitions that students must learn to manage. In simplified classroom datasets, the row, target variable, feature set, validation design, and interpretation of metrics are often already defined. In industrial data, these elements have to be justified. For this reason, student reports can be analysed as evidence of how students understand applied analytics as a methodological and decision-oriented process.

This paper contributes to data science and business analytics education by analysing how student teams translate real industrial problems into analytical tasks and business decisions. The study uses 14 student project reports from the two industrial case studies and codes them according to methodological transitions that are often hidden in simplified classroom datasets, including business-process understanding, analytical-unit construction, target design, feature availability, validation design, metric interpretation, and decision support. The reports are analysed as empirical evidence of how students reason through real industrial analytics problems. The analysis identifies recurring strengths, weaknesses, and teaching implications, with a particular focus on the transition from business process to analytical task and from model output to business decision. In addition, this study addresses the following research questions:

- RQ1: How do students describe and understand the business process in industrial analytics projects?
- RQ2: How do students construct the analytical dataset, target variable, and features?
- RQ3: Which modelling and validation choices do students most frequently make?
- RQ4: How do students interpret model results in relation to the business problem and economic decision?
- RQ5: Which teaching interventions can be derived from the observed patterns?

The remainder of the paper is organised as follows. Section 2 reviews related literature. Section 3 describes the educational setting, case studies, corpus, and coding framework. Section 4 presents the findings. Finally, Section 5 discusses implications for analytics education and concludes the paper.

2 Related Work

2.1 Project-based and work-based learning with real industrial data

Project-based and case-based learning are frequently used in data science and business analytics education because they expose students to tasks that are closer to professional analytical practice than isolated algorithmic exercises (Brüngel et al., 2020; Raschka, 2022). In such settings, students have to formulate the problem, understand the available data, prepare the dataset, select methods, evaluate results, and communicate recommendations. This is important because professional data science work is usually carried out with incomplete, noisy, context-dependent, and process-generated data rather than with clean benchmark datasets (Chapman et al., 2000; Kutzias, Dukino, Kötter, & Kett, 2023).

Work-based learning extends this logic by placing the learning activity closer to professional practice. It usually includes projects, workplace simulations, industry collaboration, or entrepreneurship-oriented activities in which students develop practical knowledge

through real-world challenges (Perusso & Wagenaar, 2021). This is especially relevant for interdisciplinary analytics education because industrial problems require more than technical modelling. Students have to understand organisational processes, communicate with different stakeholders, work in teams, and justify analytical choices in relation to business value. Recent work on online WBL also shows that work-based approaches can support the connection between academic learning, digital competence, and professional contexts (Rienties et al., 2023). In the present study, the industrial cases function as a WBL environment because students worked with real company problems, raw process-generated data, and decision-oriented outputs rather than with predefined classroom exercises.

Previous work has emphasised the need to align data-centric curricula with industry requirements. Authors in (Ramasamy & Ramamoorthy, 2023) argue that data-centric project curricula can support the development of an industry-ready workforce by connecting technical learning with current analytics requirements. Similarly, in (Brüngel et al., 2020) authors presented project-based learning in a machine learning course with differentiated industrial projects and show that industrial project settings can support both practical and social skills needed for real-world problem solving. Work on labour-market alignment also indicates that higher education curricula may lag behind changing industry needs, especially in applied technical skills and interdisciplinary problem solving (Shiwei & Abas, 2026). More general evidence on project-based learning also supports its positive role in student learning outcomes (Zhang & Ma, 2023). In machine learning education, hands-on project work has been used to connect model development with experimental design, reporting, and communication of results (Raschka, 2022).

These studies support the use of industrial problems, project-based learning, and WBL in data science and business analytics education. However, they mostly focus on curriculum design, industry readiness, skill alignment, or the general value of authentic learning. Less attention is given to what student project reports reveal about the specific methodological transitions that students have to manage when they work with raw industrial data. This is the focus of the present study: the reports are used as evidence of how students move from professional practice to analytical formulation, and from model outputs to decision-oriented recommendations.

2.2 Leakage, validation, and feature availability as learning challenges

A central difficulty in applied machine learning is that technically correct pipelines may still produce invalid conclusions if the data are not split, prepared, or interpreted according to the real prediction setting (Kapoor & Narayanan, 2023; Sasse et al., 2025). Data leakage occurs when information that would not be available in a real use case enters the training or evaluation process. This can lead to overly optimistic performance estimates and models that fail to generalise (Apicella et al., 2025; Kapoor & Narayanan, 2023; Sasse et al., 2025). Recent work shows that leakage is not a marginal issue. In (Apicella et al., 2025), the authors discuss leakage risks in machine learning and transfer learning and warn against "push the button" use of machine learning tools without sufficient methodological control. Moreover, research in (Sasse et al., 2025) provides an overview of leakage scenarios in supervised machine learning and shows that leakage may arise from overlap, preprocessing, feature construction, or evaluation design. The importance of proper data representation is emphasized in (Kapoor & Narayanan, 2023), where authors connect leakage with the reproducibility crisis in machine-learning-based science and show that leakage has affected a wide range of research fields, especially education. Tool-oriented work, such as LeakageDetector 2.0, further shows that leakage can occur inside practical Jupyter-driven machine learning workflows (Truong et al., 2025).

Applied analytics projects require a sequence of connected methodological decisions, from business understanding and data preparation to modelling, evaluation, and deployment-oriented interpretation. Cross-Industry Standard Process for Data Mining provided a classical process view of data mining projects (Chapman et al., 2000), while more recent comparisons of data science process models show that such frameworks remain relevant for structuring data science work across project phases (Kutzias et al., 2023). For educational projects, this process perspective is important because it makes visible that model training is only one part of the analytical task. Students also have to define the unit of analysis, construct the target variable, decide which features are valid, and select an evaluation design that corresponds to the intended use of the model (Chapman et al., 2000; Kutzias et al., 2023). In educational projects, leakage is not only a technical problem, but also a learning signal. If students use features that are generated after the decision point, or if they split data in a way that allows the same entity or future information to appear in both training and test data, this indicates a weak understanding of the temporal or process structure of the industrial problem (Apicella et al., 2025; Kapoor & Narayanan, 2023; Sasse et al., 2025).

2.3 Business analytics, case-based learning, and decision support

Business analytics education requires students to connect analytical outputs with decisions, risks, costs, and organisational value. A predictive model is useful only if its output can be translated into an action, recom-

mendation, or decision criterion. For this reason, case-based and problem-driven approaches are especially relevant in business analytics education. Research on case-based learning suggests that real-life scenarios can support analytical reasoning, deeper cognitive processing, and decision-making skills (Silitubun, Mustaji, & Dewi, 2025). More generally, a meta-analysis of problem-based, project-based, and case-based learning found a positive effect of student-centred, problem-driven methods on student motivation (Wijnia, Noordzij, Arends, Rikers, & Loyens, 2024). These findings support the use of authentic problems as teaching instruments, especially when the goal is not only technical knowledge but also the ability to reason about decisions.

Industry-oriented business analytics programmes also emphasise the need to bridge academic learning and organisational practice. The authors in (Pang & Quek, 2026) describe a structured industry-collaboration framework in business analytics education and argue that industry-ready analytics professionals require both analytical competence and exposure to practical organisational problems. This is consistent with the central premise of the present study: students should not only build models, but also explain how model outputs support decisions. Decision-support literature similarly emphasises that analytical results should be connected to alternatives, criteria, organisational goals, and decision processes rather than treated as isolated model outputs (Turban, Sharda, & Delen, 2010). Multi-criteria decision-making methods provide a structured way to compare alternatives when several criteria have to be considered at the same time (Velasquez & Hester, 2013). In this context, AHP is relevant because it requires alternatives, criteria, and priorities to be made explicit (Saaty, 1980; Velasquez & Hester, 2013). However, decision-support methods do not automatically integrate predictive modelling with business decision-making. If decision criteria are not connected to model outputs, feature importance, business risk, or economic value, the decision-support component remains separate from the analytical model. This is a central issue in interdisciplinary student projects: students may build a model and also perform multi-criteria decision-making, but still fail to explain how prediction, criteria, alternatives, and economic value are connected.

3 Methodology

3.1 Educational context

The empirical material consisted of student project reports produced as part of interdisciplinary projects at the University of Zagreb Faculty of Organization and Informatics (UNIZG FOI). The student teams included students from Artificial Intelligence for Business Applications (AIBA), within the Information and Business Systems (IBS) study programme, and Business Decision Making (BDM), within the Economics of Entrepreneurship (EE) study programme. This made the setting suitable for connecting technical modelling with business-oriented interpretation. Since the projects were based on real industrial problems and embedded in student assessment, the setting can be interpreted as a work-based learning environment. The industrial partners provided real business problems and process-generated data, while the student teams were expected to transform these problems into analytical and decision-support tasks. This required students to analyse a real business case, describe the company and its market environment, examine the available data, prepare the dataset, construct analytical features, select and evaluate predictive models, compare model-based results with multi-criteria decision-making, and propose an implementation-oriented business recommendation. Thus, the reports combined data science, information systems, business analytics, and economic decision-making.

All reports followed a common documentation template. The required sections included company description, product or service description, market and competitor analysis, problem description, dataset description, exploratory data analysis, data-quality assessment, data preparation, modelling, AHP-based criteria weighting, comparison between sensitivity analysis and AHP weights, evaluation, alternatives, implementation risks, ethical reflection, implementation planning, and final recommendations. This common structure made the reports comparable and allowed the analysis to focus not only on the presence of project components, but also on how students connected them into a coherent analytical and decision-support process.

3.2 Case studies

The first case study was provided by SICK Mobilisis, a company that develops intelligent fleet management, traffic monitoring, and parking management solutions. The case focused on AI-based fuel-event detection using vehicle telemetry data. The data type was sensor-based, telemetry-oriented, and temporal. The main educational challenge was that students had to understand the difference between classifying individual records and detecting operational events. A single telemetry record is not necessarily equivalent to a fuel event. A suspicious event may appear as a sequence of measurements, a fuel-level change over time, a threshold crossing, or a pattern that has to be interpreted in relation to vehicle movement and operational context. Therefore, students had to reason about time, rare events, false alarms, missed detections, event construction, and operational use. The SICK Mobilisis reports generally included standard sections on the company, products, market, problem, data, data quality, preparation, modelling, and AHP. Some reports also included more ad-

vanced modelling components, such as Random Forest, XGBoost, multilayer perceptrons (MLP), recurrent neural networks (RNN), gated recurrent units (GRU), bidirectional long short-term memory (LSTM), conditional random field (CRF)-based extensions, threshold tuning, and production-oriented testing. These elements made the case useful for analysing how students respond to temporal and sensor-driven industrial problems.

The second case study was provided by Emil Frey Digital, a company that develops digital solutions for car sales and after-sales processes. The case focused on the optimisation of digital used-vehicle sales through analytical recommendations. The data reflected a business process involving vehicle publishing, marketplace platforms, customer leads, offers, and contracts. The main educational challenge was that students had to understand the sales funnel, construct a valid unit of analysis, define a meaningful target variable, reason about attribution of sales to platforms, and connect model results with return on investment, conversion, and decision support. The Emil Frey Digital reports followed a structure similar to the SICK Mobilisis reports. They included modelling, sensitivity analysis, AHP, alternatives, risk analysis, ethical reflection, and implementation planning. However, the methodological challenge was different. Students had to determine what one row represented: a vehicle, a vehicle-marketplace combination, a published advertisement, a lead, an offer, or a contract-related outcome. This choice determined the meaning of the target variable, the availability of predictors, the validation design, and the interpretation of the model as a recommendation tool.

3.3 Corpus

Table 1. Corpus summary

Element	Value
Student reports	14
Industrial cases	2
SICK Mobilisis reports	7
Emil Frey Digital reports	7
Listed students	122
Students in SICK Mobilisis teams	60
Students in Emil Frey Digital teams	62
Typical team size	8–9
Report-length range	14–146 pages
Approx. corpus size	660 pages
Report type	Business, analytics, AHP

The corpus consisted of 14 student project reports produced by interdisciplinary student teams. Seven reports addressed the SICK Mobilisis fuel-event detection case and seven reports addressed the Emil Frey Digital sales-optimisation case. Based on the cover pages, the reports involved 122 listed students, 52 from AIBA and 70 from BDM, with teams typically consisting of eight or nine members. Table 1 summarises the corpus. The reports varied in the methodological detail provided across the coding dimensions, which made it possible to identify recurring patterns. Grades were neither collected nor used as coding variables, and the analysis did not evaluate or rank individual teams.

The reports were treated as the primary unit of analysis. Individual report sections were used as secondary analytical units when coding specific methodological and pedagogical dimensions. These sections included the problem description, dataset description, data-quality assessment, data preparation, modelling, validation, sensitivity analysis, AHP-based decision support, evaluation, risk analysis, ethical reflection, and implementation planning. The aim was not to grade individual student teams, but to identify recurring learning patterns and methodological difficulties that appear when students work with real industrial data. Because all reports followed the same assignment template, the coding focused on differences in methodological reasoning instead of differences in report structure.

3.4 Coding framework and scale

A qualitative coding framework was developed to analyse how students translated industrial problems into analytical projects. The framework followed the expected stages of an applied analytics project: understanding the business process, constructing the analytical dataset, defining the target, preparing the data, selecting and validating models, interpreting metrics, and connecting analytical results to a decision. The coding framework included eleven dimensions, shown in Table 2. Each dimension captured a specific transition that students had to manage. Business-process understanding captured the transition from a general company description to the operational process that generated the data. Target design captured the transition from a business problem to a measurable outcome. Validation design captured the transition from a trained model to a credible estimate of future performance. AHP and decision support captured the transition from predictive modelling to economic or managerial decision-making. Each report was assessed independently on every coding dimension using the four-point ordinal scale shown in Table 3. Evidence was taken from the report sections relevant to that dimension rather than from the report's overall presentation or grade. The scale was designed for descriptive comparison rather than statistical hypothesis testing. The authors reviewed the reports and resolved ambiguous cases through discussion and consensus. Higher scores indicated a clearer connection between the coded element, case-specific evidence, methodological reasoning, and the business decision.

For example, in the target-design dimension, a report received a score of 0 if the target variable was not clear. It received a score of 1 if the target was named but its business meaning was not explained. It received a score of 2 if the target was explained but still involved potential leakage or weak alignment with the decision.

Table 2. Coding dimensions

Dimension	Focus
Business process	Understanding of the process and decision context.
Analytical unit	Definition of row, event, vehicle, advertisement, lead, offer, or contract.
Target design	Clarity and business validity of the target variable.
Feature availability	Availability of predictors at the moment of decision and leakage risk.
Data preparation	Missing values, duplicates, outliers, feature construction, and integration.
Modelling choice	Suitability of selected models for the data structure and task.
Validation design	Appropriateness of random, temporal, group-based, or event-level validation.
Metric interpretation	Connection between model metrics and business or operational risk.
AHP / decision support	Connection between criteria, model outputs, alternatives, and economic decisions.
Implementation	Explanation of how the proposed solution could be used in practice.
Ethics / risk	Specificity of ethical and implementation risks.

Table 3. Coding scale

Score	Meaning
0	The element is not present.
1	The element is present only superficially or generically.
2	The element is partially developed and supported by case-specific examples.
3	The element is methodologically clear and connected to the business decision.

It received a score of 3 if the target was clearly connected to the business process, the moment of decision, and the evaluation design.

The same scoring logic was applied within the other dimensions. For validation design, the scale progressed from no stated design (0), through an unjustified generic split (1) and a partly justified but incomplete design (2), to a temporal, group-based, per-entity, or event-level design aligned with intended use (3). For metric interpretation, it progressed from no interpretation (0), through generic reporting (1) and partial case-specific interpretation (2), to an explicit connection between metrics and operational or economic consequences (3). For AHP and decision support, it progressed from absence (0), through a separate or generic AHP exercise (1), partial linkage to analytical evidence (2), and an explicit connection among model outputs, criteria, alternatives, and business decisions (3).

The scale did not define binary or global "strong report" and "weak report" classes. Scores were not aggregated into an overall report-quality label, and no threshold was used to classify an entire report. Instead, every dimension retained its own score. A report could therefore provide methodologically clear evidence in one dimension, such as business-process understanding, while providing only generic evidence in another, such as validation. Such mixed profiles were expected and were central to the analysis because they revealed which transitions were understood and which still required instructional support.

3.5 Qualitative comparison

The analysis followed a comparative qualitative design. First, each report was reviewed according to the coding dimensions. Second, recurring patterns were identified within each case group. Third, the two case groups were compared to determine how the nature of the industrial problem shaped student reasoning.

The comparison was structured around four analytical questions. First, how did students understand the business process behind the data? Second, how did they construct the analytical dataset and define the unit of analysis? Third, how did they choose models and validation procedures? Fourth, how did they interpret results and connect them to decision support?

This design allowed the analysis to move beyond individual report quality and focus on the pedagogical value of real industrial cases. The two cases created different learning pressures. The SICK Mobilisis case exposed problems related to temporal data, rare events, event construction, and operational alarms. The Emil Frey Digital case exposed problems related to business-process data, target construction, attribution, recommendation, and economic evaluation.

4 Results

Table 4 summarises the main patterns observed in the reports. These frequencies are binary indicators of explicit evidence and are reported separately from the four-point dimensional scores. Formal sections were counted when the required report component was present. A methodological-integration pattern was counted only when the report explicitly stated or justified the relevant connection to the business process, analytical unit, target definition, validation design, model interpretation, or decision-support logic. The totals therefore describe evidence for individual patterns; they do not classify complete reports. A report could be counted in one row and not in another, producing mixed dimensional profiles. Fig. 1 visualises the contrast between formal completion and methodological integration.

4.1 Business-process understanding

Students generally described the companies, products, services, markets, competitors, and SWOT analyses in sufficient detail. However, the operational process behind the data was less consistently explained. In the SICK Mobilisis case, most reports identified fuel-event detection as the core problem. Reports that clearly represented the process treated it as a temporal telemetry

Table 4. Observed methodological patterns in the analysed reports

Observed pattern	SICK Mobilisis	Emil Frey Digital	Total
Formal data-quality section included	7/7	7/7	14/14
Data-preparation section included	7/7	7/7	14/14
Modelling section included	7/7	7/7	14/14
AHP or decision-support section included	7/7	7/7	14/14
Risk, ethics, and implementation sections included	7/7	7/7	14/14
Analytical unit explicitly discussed	3/7	3/7	6/14
Target variable explicitly linked to the business decision	4/7	3/7	7/14
Feature availability or leakage explicitly discussed	3/7	3/7	6/14
Validation justified beyond a generic random split	4/7	3/7	7/14
Metrics linked to operational or economic risk	3/7	3/7	6/14
AHP criteria explicitly connected to model outputs or feature importance	3/7	3/7	6/14
Concrete deployment or operational-use scenario included	4/7	3/7	7/14

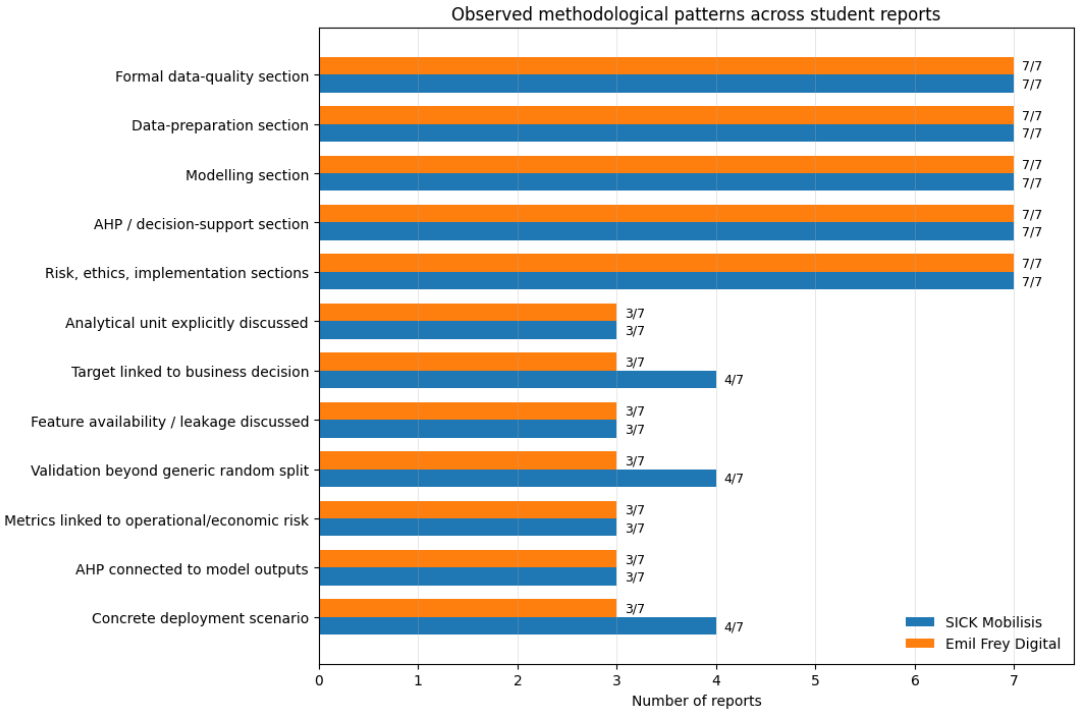


Figure 1. Observed methodological patterns across the 14 student reports. Formal project sections were present in all reports, while methodological integration was less consistent, especially for analytical-unit definition, feature availability, metric interpretation, and the connection between predictive modelling and AHP-based decision support.

problem in which an event may involve a sequence of measurements or a fuel-level change over time. Other reports treated the task mainly as ordinary row-level classification and did not explain how individual predictions would become operational alarms. In the Emil Frey Digital case, most reports described digital vehicle sales and online marketplaces, but their treatment of the sales funnel varied. Some reports treated the data as a standard tabular classification dataset. Others explicitly recognised that publishing, leads, offers, contracts, and sales outcomes occur at different stages and cannot all be used at the same decision point. Across both cases, company-context descriptions were more complete than explanations of the process logic connecting data, prediction, and action.

4.2 Construction of the analytical dataset

The analytical unit was one of the main difficulties. Several reports used the available rows as given, whereas others explicitly discussed what the rows should represent for the intended decision.

In the SICK Mobilisis case, the raw unit was usually a telemetry record, but the business unit was a fuel event. This created a mismatch between row-level modelling and event-level interpretation. Some reports addressed this mismatch by considering time windows, vehicle trajectories, sequences, thresholding, or alarm construction. Others reported classification metrics without explaining how record-level predictions would be aggregated into an event.

In the Emil Frey Digital case, the unit of analysis was less obvious because the data came from several linked business tables. A row could represent a vehicle,

a vehicle-marketplace combination, an advertisement, a lead, an offer, or a contract. Reports that did not define this unit clearly were more exposed to repeated-vehicle leakage, target ambiguity, and the use of post-publication information in pre-publication recommendations.

The cross-case pattern is simple: in industrial data, the row is not self-evident. It has to be constructed according to the decision that the model is expected to support.

4.3 Modelling and validation choices

The reports used a wide range of models. In the SICK Mobilisis case, teams used classical tabular models such as Random Forest and XGBoost, as well as MLP, RNN, GRU, LSTM, and CRF-based variants. Some reports connected these choices to sequence construction, temporal order, threshold selection, and event interpretation. In others, more complex models were used without a clear explanation of their suitability for the telemetry problem.

In the Emil Frey Digital case, tabular models such as logistic regression, Random Forest, XGBoost, Naive Bayes, and MLP were generally appropriate for the structured business data. The main problem was not the choice of algorithm, but the target. Predicting a historical marketplace label does not necessarily recommend the best marketplace. Predicting sale probability also does not directly determine return on advertising investment or marketing-budget allocation.

Validation was a recurring difficulty. Some reports used generic random splits, whereas others recognised the need for temporal, group-based, or process-aware validation. In the SICK Mobilisis case, validation should respect temporal order and event structure. In the Emil Frey Digital case, it should avoid placing records of the same vehicle in both training and test data. Validation design therefore provided a useful indicator of process understanding.

4.4 Feature availability and leakage

Feature availability was one of the clearest indicators of methodological understanding. Most reports included data preparation and model evaluation, but fewer reports explicitly discussed which variables would be available at the moment when the model would be used. This distinction was important in both cases because some variables were generated after the relevant business decision or were linked to the same entity across multiple records.

In the SICK Mobilisis case, the main leakage risk was connected to temporal order and event construction. A model for fuel-event detection should not use information that becomes available only after the event has unfolded. Some reports recognised this problem by discussing chronological splits, per-vehicle validation, threshold selection, sequence construction, or the removal of variables that could reveal the target too directly. Other reports treated the data mainly as independent tabular records, which made the validation design less aligned with the intended operational use.

In the Emil Frey Digital case, the leakage risk was connected to the sales funnel. Some variables related to leads, offers, contracts, or later customer actions may be useful for analysing historical sales performance, but they are not necessarily available when the system has to recommend where or how to publish a vehicle advertisement. Some reports carefully separated publishing-time information from post-publication outcomes. Others used later-stage variables as predictors, which may improve apparent model performance but weaken the validity of the recommendation task.

Across both cases, leakage was therefore not only a technical data-splitting problem. It indicated the extent to which students understood the temporal and process-based structure of the industrial task. Reports that explicitly discussed feature availability, entity overlap, chronological order, or the moment of decision also tended to provide more credible validation and more meaningful business interpretation.

4.5 Interpretation and decision support

Most reports included model metrics and an AHP-based decision-support component. The main difference was how well these parts were connected.

In the SICK Mobilisis case, some reports interpreted metrics through operational risk. Precision, recall, F1-score, and threshold choice were connected to false alarms, missed fuel withdrawals, real-time monitoring, or inspection decisions. Other reports presented the metrics without explaining their operational consequences. In the Emil Frey Digital case, some reports linked model outputs to conversion, time-to-sale, advertising cost, lead quality, or return on investment. Others treated feature importance or prediction accuracy as sufficient for a marketplace recommendation without explaining its economic meaning.

AHP was present in the project structure, but it was not always integrated with the predictive model. In some reports, AHP appeared as a separate economic exercise. In others, model results, feature importance, operational criteria, and business alternatives were discussed together. This integration is important because AHP becomes more educationally meaningful when its criteria are justified by model evidence, domain constraints, and economic consequences rather than by generic preferences alone. Overall, the least developed transition was not from data preparation to modelling, but from model performance to a justified business decision.

4.6 Summary of findings by research question

For **RQ1**, students generally described the companies, markets, and business problems in sufficient detail. However, they less consistently explained the operational process that generated the data and determined the analytical task.

For **RQ2**, the main difficulties concerned the analytical unit, target variable, and feature availability. The reports did not always distinguish telemetry records from fuel events or clearly determine what a row represented: a vehicle, advertisement, marketplace entry, lead, offer, or contract. Variables generated after the decision point also created potential leakage.

For **RQ3**, most teams used appropriate supervised-learning models, including Random Forest, XGBoost, logistic regression, Naive Bayes, MLP, and sequence-oriented models. Model selection was generally less problematic than validation. Several reports used generic random splits, while others recognised the need for temporal, group-based, per-vehicle, or process-aware validation.

For **RQ4**, students usually reported standard model-evaluation metrics, but did not always connect them to operational or economic consequences. The clearest interpretations related false alarms and missed events to operational risk or connected model outputs to conversion, advertising cost, return on investment, lead quality, and time-to-sale.

For **RQ5**, the findings indicate that students should explicitly define the analytical unit, identify the moment of decision, review feature availability, justify the validation design, interpret metrics in relation to business consequences, and connect predictive modelling with AHP-based decision support.

4.7 Pedagogical implications

The observed patterns suggest several pedagogical implications for interdisciplinary analytics education:

- **Define the analytical unit before modelling.** Students should explicitly state if the unit of analysis is a record, event, vehicle, advertisement, lead, offer, contract, or another process-specific object. This prevents them from treating the available row in the dataset as automatically suitable for prediction or decision-making.
- **Use WBL checkpoints to connect academic and professional reasoning.** Industrial analytics projects should include moments where students explicitly relate their analytical choices to the professional context of use. This includes explaining how raw company data become an analytical dataset, how model outputs support a business action, and how recommendations would be communicated to non-technical stakeholders.
- **Introduce the moment of decision as a mandatory checkpoint.** Before selecting features, students should identify when the model is expected to be used and which variables are available at that point. This can reduce leakage and help students distinguish between historical explanation and usable prediction.
- **Teach validation design as part of problem formulation.** Students should justify if a random, temporal, group-based, per-entity, or event-level split is appropriate for the intended use of the model. This is especially important when repeated entities, temporal order, or process stages make random splits misleading.
- **Connect model metrics to operational or economic consequences.** Students should explain what false positives, false negatives, precision, recall, F1-score, or related metrics mean in the specific business case. Evaluation should therefore become part of decision-making rather than only a technical reporting requirement.
- **Integrate AHP with predictive modelling and feature importance.** AHP criteria should not be selected independently from the analytical model. Students should explain how criteria are informed by model outputs, feature importance, business risks, costs, and implementation constraints. For example, if a model identifies lead quality, marketplace source, price, time-to-sale, or anomaly patterns as important predictors, students should explain how these variables influence the criteria used in AHP. This would help them avoid treating machine learning and multi-criteria decision-making as separate project sections.

These interventions suggest that industrial analytics projects should include explicit checkpoints for analytical-unit definition, decision-point identification, feature-availability review, validation-design justification, metric-to-risk interpretation, and model-to-decision integration. Such checkpoints can make methodological reasoning more visible and help students move from formal project completion toward coherent analytical decision support.

5 Conclusion

This paper analysed how students translate real industrial problems into analytical and decision-support tasks. Across the two industrial cases, students generally completed the required project components, including data preparation, modelling, evaluation, and AHP-based decision support. However, they were less consistent in defining the analytical unit, constructing a business-valid target variable, identifying feature availability at the decision point, justifying the validation design, and linking model results to operational or economic consequences. The main difficulty was therefore

not model training itself, but the transition from the business process to the analytical task and from model output to business decision. These findings indicate that industrial analytics projects should include explicit methodological checkpoints for analytical-unit definition, decision-point identification, feature availability, validation design, metric interpretation, and model-to-decision integration.

The study is limited to one interdisciplinary project, two industrial cases, and 14 student reports. The coding was descriptive and was not intended for statistical generalisation. Nevertheless, the analysis shows that student reports can provide useful evidence about how learners understand applied analytics as a methodological, technical, and decision-oriented process. Future work could extend the analysis to additional industrial cases, compare cohorts across academic years, and examine how targeted teaching interventions affect leakage awareness, validation quality, and decision-oriented interpretation. Further research could also examine team dynamics, collaboration challenges, and the implications of students' use of generative AI in interdisciplinary projects.

Acknowledgments

The authors acknowledge institutional support from the Faculty of Organization and Informatics, University of Zagreb. The broader research environment was supported by the Institutional Research Project through the National Recovery and Resilience Plan 2021–2026 under the project PiPi: Data infrastructure and augmented intelligence for sustainable development and the project StartUp.Score: Development of a model for deciding on the readiness of an entrepreneurial venture for incubation.

References

Apicella, A., Isgrò, F., & Prevete, R. (2025). Don't push the button! exploring data leakage risks in machine learning and transfer learning. *Artificial Intelligence Review*. doi: 10.1007/s10462-025-11326-3

Brüngel, R., Rückert, J., & Friedrich, C. (2020). Project-based learning in a machine learning course with differentiated industrial projects for various computer science master programs. In *Conference on software engineering education and training.* doi: 10.1109/CSEET49119.2020.9206229

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). Crisp-dm 1.0: Step-by-step data mining guide [Computer software manual].

De Veaux, R. D., Agarwal, M., Averett, M., Baumer, B. S., Bray, A., Bressoud, T. C., … Ye, P. (2017). Curriculum guidelines for undergraduate programs in data science. *Annual Review of Statistics and Its Application*, *4*, 15–30. doi: 10.1146/annurev-statistics-060116-053930

Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*. doi: 10.1016/j.patter.2023.100804

Kutzias, D., Dukino, C., Kötter, F., & Kett, H. (2023). Comparative analysis of process models for data science projects. In *Icaart (3)* (pp. 1052–1062).

Pang, Y., & Quek, K. P. (2026). Bridging academia and industry in business analytics education: Insights from the national university of singapore. *INFORMS Journal on Applied Analytics*. doi: 10.1287/inte.2026.0315

Perusso, A., & Wagenaar, R. (2021). The state of work-based learning development in eu higher education: learnings from the wexhe project. *Studies in Higher Education*, *47*(7), 1423–1439. doi: 10.1080/03075079.2021.1904233

Ramasamy, V., & Ramamoorthy, S. (2023). Sustainable data-centric project curriculum for industry-ready workforce. *Journal of Engineering Education Transformations*. doi: 10.16920/jeet/2023/v36is2/23014

Raschka, S. (2022). Deeper learning by doing: integrating hands-on research projects into a machine learning course. In *The second teaching machine learning and artificial intelligence workshop* (pp. 46–50).

Rienties, B., Divjak, B., Iniesto, F., Pažur Aničić, K., & Zizak, M. (2023). Online work-based learning: A systematic literature review. *International Review of Education*, *69*. doi: 10.1007/s11159-023-10008-y

Saaty, T. L. (1980). *The analytic hierarchy process*. New York: McGraw-Hill.

Sasse, L., et al. (2025). Overview of leakage scenarios in supervised machine learning. *Journal of Big Data*. doi: 10.1186/s40537-025-01193-8

Shiwei, Y., & Abas, A. (2026). Data science talent demand in china: A large-scale job posting analysis and implications for curriculum alignment. *International Journal of Latest Technology in Engineering Management & Applied Science*. doi: 10.51583/ijltemas.2026.150300052

Silitubun, E., Mustaji, & Dewi, U. (2025). The effectiveness of case-based learning in enhancing students' critical thinking skills. *Journal of Neonatal Surgery*. doi: 10.52783/jns.v14.1844

Truong, O., et al. (2025). Leakagedetector 2.0: Analyzing data leakage in jupyter-driven machine learning pipelines. In *Ieee international conference on software maintenance and evolution.* doi: 10.1109/ICSME64153.2025.00102

Turban, E., Sharda, R., & Delen, D. (2010). *Decision support and business intelligence systems* (9th ed.). Pearson.

Velasquez, M., & Hester, P. T. (2013). An analysis of multi-criteria decision making methods. *International Journal of Operations Research*, *10*(2), 56–66.

Wijnia, L., Noordzij, G., Arends, L. R., Rikers, R. M. J. P., & Loyens, S. M. M. (2024). The effects of problem-based, project-based, and case-based learning on students' motivation: A meta-analysis. *Educational Psychology Review*. doi: 10.1007/s10648-024-09864-3

Zhang, L., & Ma, Y. (2023). A study of the impact of project-based learning on student learning effects: A meta-analysis study. *Frontiers in Psychology*, *14*. doi: 10.3389/fpsyg.2023.1202728

An AI Companion for Self-Regulated Learning: Integrating Conversational AI, Predictive Analytics, and Explainability

Hana Ivančić, Tomislav Jagušt
University of Zagreb
Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Croatia
{hana.ivancic, tomislav.jagust}@fer.unizg.hr

Abstract. *Artificial intelligence in higher education often focuses on conversational assistants rather than broader learning processes. This paper presents ongoing research towards FER AI Companion, a self-regulated learning (SRL) platform integrating retrieval-augmented generation, predictive analytics, and explainable AI, with reflection and gamification planned as extensions. The system follows Zimmerman's SRL cycle and is evaluated using historical data from five first-year courses (3,958 students). Tree-based models achieve MAE of approximately 3.0 ISVU points at mid-semester and 1.8–2.0 at the late checkpoint, outperforming a training-mean baseline (11.5). Inter-model disagreement also provides a relative error signal. No user study has been conducted.*

Keywords. predictive learning analytics, self-regulated learning, retrieval-augmented generation, explainable artificial intelligence, gamification, higher education

1 Introduction

Artificial intelligence is rapidly transforming higher education by enabling personalized learning experiences and providing students with continuous access to academic support (Zawacki-Richter, 2019). Recent advances in large language models (LLMs) have led to the widespread adoption of AI-powered assistants capable of answering questions, explaining course concepts, and assisting with administrative tasks (McGrath et al., 2025; Kasneci et al., 2023). While such systems improve information accessibility, they typically function as conversational interfaces and provide limited support for broader learning processes such as planning, monitoring, and reflecting on learning (Lan & Zhou, 2025).

Self-regulated learning (SRL) emphasizes that successful students actively set learning goals, monitor their progress, and evaluate their performance through a continuous cycle of forethought, performance, and self-reflection (Zimmerman, 2002). Supporting these phases requires more than question answering alone. Predictive learning analytics can provide students with early feedback regarding expected academic outcomes, explainable artificial intelligence (XAI) can clarify why predictions were made, and gamification can encourage sustained engagement with learning activities (Khosravi et al., 2022). Integrating these complementary technologies within a unified system has the potential to provide substantially richer educational support than standalone AI chatbots.

Despite the rapid emergence of AI-powered educational tools, current systems are typically optimized for isolated functions (most commonly question answering or administrative support) rather than for sustained learning process orchestration (McGrath et al., 2025). This creates a structural gap between what learning science identifies as essential for student success and what AI systems currently deliver in practice. In particular, self-regulated learning requires continuous interaction between goal setting, performance monitoring, adaptive feedback, and reflection (Panadero, 2017), yet such integration remains uncommon in the literature (Khosravi et al., 2022). This work addresses this gap by proposing a unified design approach in which predictive analytics, explainable AI, conversational assistance, and behavioral scaffolding are not treated as independent modules, but as coordinated components of an SRL-aligned feedback loop. In doing so, the system moves beyond the paradigm of AI as an information provider toward AI as an active learning process companion.

This paper presents the first phase of the development of FER AI Companion, an AI-supported web application designed to support first-year university students within an SRL framework. The proposed architecture combines retrieval-augmented generation (RAG) (Lewis et al., 2021) for grounded conversational assistance, machine learning models for grade prediction, AI-generated personalized study plans, automated reflection prompts, and gamification mechanisms within a single platform structured around Zimmerman's SRL model (Zimmerman, 2002). Prediction transparency is achieved through SHAP (SHapley Additive exPlanations), an explainable AI

method that identifies the features contributing most to individual predictions. The contribution of this work is twofold. First, it presents an SRL-aligned architecture that connects predictive analytics and prediction explanations with grounded conversational assistance, while explicitly identifying reflection and gamification as future extensions. Second, it provides an initial empirical evaluation of the predictive component on a chronological held-out cohort, including baseline comparisons and analysis of inter-model disagreement. The paper therefore contributes an initial technical and empirical foundation for an SRL-oriented learning companion rather than evidence of educational effectiveness.

This paper addresses the following research questions: (RQ1) How accurately can final course outcomes be predicted at different semester checkpoints using machine learning models trained on historical student data? (RQ2) Does disagreement among predictive models provide a useful indication of prediction error? (RQ3) How are the implemented and planned system functions mapped to theoretically defined SRL processes? (RQ4) What components of the proposed platform are implemented and technically evaluated at the current stage of development? The present study empirically addresses RQ1 and RQ2 through predictive-model evaluation; RQ3 is addressed at the design level in Section 3; and RQ4 through the implementation and prototype descriptions in Sections 4 and 5.

2 Related Work

The application of artificial intelligence in higher education has gained significant momentum in recent years, with a growing number of systems designed to support students through conversational interfaces. Recent reviews indicate that most applications focus on conversational assistance, information retrieval, writing support, or administrative guidance, while a relatively small number of systems are explicitly grounded in learning theory or designed to support the complete learning process (McGrath et al., 2025; Zawacki-Richter, 2019; Lo, 2023). Some early institutional deployments include Georgia State University’s chatbot that provided students with timely administrative information (Mainstay, 2026), Deakin University's Genie assistant (Deakin University, 2026), the University of Michigan's Maizey advising assistant (LSA Newnan Advising Center, 2026), and the Croatian AIVA system, which integrates a generative AI advisor into Moodle and explicitly adopts Zimmerman’s self-regulated learning framework (Faculty of Organization and Informatics, 2026). While these systems demonstrate the potential of AI in student support, they share a common limitation: their primary function is answering questions and delivering information, without integrating predictive analytics, explainability mechanisms, or systematic support for the full self-regulated learning cycle.

Predictive learning analytics has become one of the most extensively studied applications of artificial intelligence in higher education. Recent systematic reviews report that machine learning methods are widely used to predict academic performance, dropout risk, and student success, with ensemble techniques such as Random Forest and XGBoost consistently achieving strong predictive performance on educational datasets (Sghir et al., 2023; Namoun & Alshanqiti, 2021). The shift from traditional statistical approaches toward ensemble machine learning methods and gradient boosting algorithms was driven by their ability to capture complex relationships within heterogeneous educational data (Sghir et al., 2023; Namoun & Alshanqiti, 2021). Beyond improving prediction accuracy, recent research increasingly emphasizes the practical usefulness of predictive learning analytics in supporting decision-making by both students and instructors. To achieve this objective, predictive models must not only produce accurate estimates but also provide outputs that are sufficiently transparent and actionable to inform educational interventions (Sghir et al., 2023).

In educational settings, predictive accuracy alone is insufficient. Khosravi et al. propose the XAI-ED framework, arguing that explanations should serve a pedagogical purpose and be tailored to different types of users rather than functioning merely as technical model outputs (Khosravi et al., 2022). More generally, explainable artificial intelligence has emerged as an essential prerequisite for responsible AI, enabling users to understand, trust, and appropriately act upon model predictions (Barredo Arrieta et al., 2020). Among existing methods, SHAP (SHapley Additive exPlanations) provides theoretically grounded local explanations applicable to both tree-based and linear models (Lundberg & Lee, 2017) and has been successfully applied in educational prediction tasks to identify the factors most strongly influencing individual student outcomes. Despite these advances, predictive analytics and explainability are still rarely integrated with conversational AI or broader learning support systems that help students interpret predictions and translate them into concrete learning actions.

Self-regulated learning (SRL) is widely recognized as one of the most influential theoretical frameworks for understanding successful learning. Zimmerman (2002) introduced the three-phase cycle of forethought, performance, and self-reflection, while subsequent reviews have demonstrated that different SRL models largely converge on these core regulatory processes (Panadero, 2017). Recent research suggests that AI technologies can effectively support planning, monitoring, and reflection when designed to augment rather than replace learners' self-regulation (Lan & Zhou, 2025). However, existing AI applications often support only individual SRL activities instead of providing coordinated assistance throughout the

complete learning cycle. Prior studies have shown that digital systems supporting all three SRL phases can improve academic outcomes compared with systems focused solely on information delivery (Rahal & Zainuba, 2016).

Gamification has become an established approach in higher education, for increasing learner engagement and motivation in digital learning environments (Khaldi et al., 2023; Navarro-Espinosa et al., 2022). Meta-analyses demonstrate statistically significant positive effects on cognitive, motivational, and behavioral learning outcomes when game elements are meaningfully aligned with educational objectives (Sailer & Homner, 2020). In higher education, systematic literature reviews identify points, badges, and leaderboards as the most commonly implemented mechanisms, while emphasizing that their effectiveness depends on pedagogically meaningful integration rather than the addition of isolated reward systems (Khaldi et al., 2023; Koivisto & Hamari, 2019). Nevertheless, gamification has usually been studied as a standalone intervention rather than as an integrated component of AI-supported learning systems, limiting its potential to reinforce adaptive learning behaviors within a broader self-regulated learning process (Khaldi et al., 2023).

Table 1 summarizes the key functional dimensions of representative existing systems in comparison to the system developed in this work. Representative systems were selected based on their documented use of AI-supported student assistance in higher education and the availability of publicly accessible descriptions of their functionality. The selection was intended as an illustrative comparison rather than a systematic review or exhaustive survey. The comparison indicates that none of the selected systems simultaneously integrates AI-assisted conversation, grade prediction, prediction explainability, systematic SRL support, and gamification within a unified platform, highlighting the research gap addressed in this work.

Table 1. Comparison of selected AI-based student support systems and the proposed architecture

System	AI	Pred	XAI	SRL	Game
Georgia State	Yes	No	No	No	No
Deakin Genie	Yes	No	No	Partial	No
AIVA	Yes	No	No	Partial	No
Maizey	Yes	No	No	No	No
WGU Labs	Yes	Partial	No	Partial	No
This work	Yes	Yes	Yes	Partial	Planned

Legend:
AI conversation = conversational interaction using generative or AI-based dialogue
Predictive analytics = prediction of academic outcomes based on student data
XAI = explicit explanation of individual model predictions
SRL support = explicit support for one or more SRL regulatory processes
Gamification = game elements such as points, badges, streaks or leaderboards

3 System Architecture

The FER AI Companion is implemented as a three-tier web application consisting of a React 18 frontend, a FastAPI backend, and a dual-database persistence layer consisting of a PostgreSQL relational database hosted on Supabase and a ChromaDB vector store for semantic search over course materials. The backend coordinates the implemented AI components: a large language model (OpenAI GPT-4o) for conversational and generative tasks, locally stored machine learning models for predictive analytics, and a SHAP-based explainability layer. Communication between frontend and backend is exclusively via REST API with JWT authentication.

The architecture is organized around the three phases of Zimmerman's self-regulated learning cycle (Zimmerman, 2002), with each system component explicitly mapped to one or more phases, as illustrated in Fig. 1. At the current development stage, the predictive analytics, RAG, and study-planning components are implemented, while reflection and gamification mechanisms represent planned extensions incorporated into the architectural design.

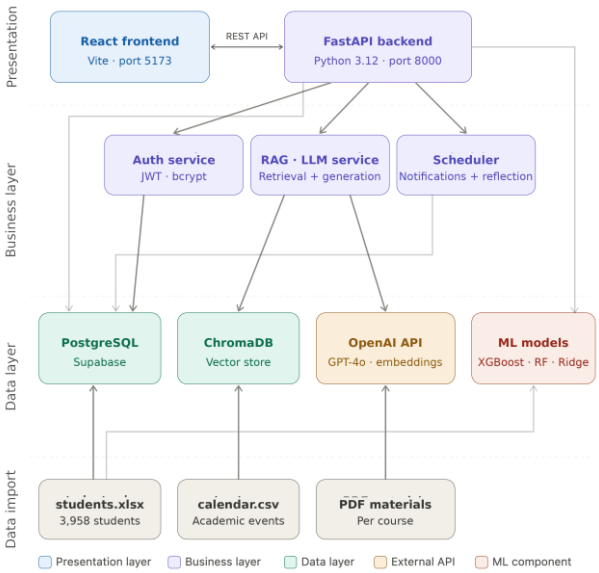


Figure 1. The system's architecture

3.1 Forethought Phase: Goal Setting and Study Planning

In the forethought phase, students set target grades for each of the five first-year courses (Digital Logic, Linear Algebra, Mathematical Analysis 1, Introduction to Programming, and Communication Skills) and request personalized study plans. The system maps the desired grade to a minimum score threshold using per-course grading scales and passes this information, together with the student's current scores and upcoming calendar events, to the language model as a structured prompt context. The model generates a markdown-formatted study plan that is stored in the database for subsequent retrieval.

The AI assistant component is built on a Retrieval-Augmented Generation (RAG) architecture (Lewis et al., 2021; Gao et al., 2024). Course materials in PDF format are segmented into approximately 500-word chunks with 50-word overlap and converted into 1536-dimensional vector representations using the text-embedding-3-small model, then stored in ChromaDB per-course collections. At query time, the five semantically closest chunks are retrieved via cosine similarity and injected into the language model prompt alongside the student's academic context and the last ten conversational turns. This architecture grounds the language model in verified course material and is expected to reduce the likelihood of hallucinations in comparison to a general-purpose LLM without retrieval.

3.2 Performance Phase: Prediction, Explainability, and AI Assistance

The predictive analytics component trains 45 models – five courses × three checkpoints × three algorithms (XGBoost (n_estimators=300, learning_rate=0.05, max_depth=4), Random Forest (n_estimators=200, max_depth=8, min_samples_leaf=3), Ridge regression (α=1.0)) – on historical student data from academic years 2020-2024, evaluated on the 2025 cohort as a held-out test set. Three prediction checkpoints reflect the information available at different points in the semester: before the semester (enrollment features only), after the midterm exam, and after the final exam. For courses with detailed component scores available (MATAN1, LINEALG, UUP), models directly predict ISVU points on a 0-100 scale (score mode); for courses with only final grades available (DIGLOG, VJEKOM), models predict the grade on a 1-5 scale with subsequent interpolation to points (grade mode). At the mid-semester checkpoint, the midterm score is included as a predictor despite contributing directly to the final target, as the goal is to provide the best available estimate based on all information known at that point; at the late checkpoint, some continuous assessment elements may still be pending, so results should be interpreted as a consolidation estimate rather than a fully prospective prediction.

The consensus mechanism combines predictions from all three algorithms: the spread between the highest and lowest prediction serves as a model disagreement indicator, classified as high agreement (spread $\leq$ 5 points), medium agreement (5-12 points), or low agreement (>12 points). These thresholds were defined with reference to the average 12-point span associated with a single grade category on the 0-100 point scale used in the system. Thus, a spread greater than 12 points indicates that the models differ by more than the approximate width of one grade category and is therefore interpreted as low model agreement. The thresholds are intended as practical, interface-level categories of inter-model disagreement rather than as calibrated measures of predictive uncertainty. This multi-model architecture is motivated not only by ensemble accuracy benefits but by the need to communicate prediction confidence, a dimension that no single model can provide on its own.

For prediction explainability, tree-based models (XGBoost, Random Forest) use SHAP TreeExplainer to compute feature-level Shapley-value-based contributions, while Ridge regression uses weighted feature contributions computed as the product of each coefficient and the mean-centered feature value, which approximates the Shapley value for linear models under the assumption of feature independence. The six features with the largest absolute contributions are returned alongside their values and directions (positive/negative), converted into human-readable explanations (e.g., "Mathematical Analysis 1 – midterm exam: 47 pts, contribution: +8.3") and visualized as a horizontal bar chart for every model separately. This implementation follows the XAI-ED pedagogical function principle (Khosravi et al., 2022): the explanation is designed to help students understand which features contributed to their prediction and potentially use this information when planning subsequent study activities.

During the performance phase, the AI assistant responds to student questions in Croatian or English, identifies the relevant course from the query, retrieves matching course material segments, and generates responses grounded in those materials. The system prompt explicitly prohibits the model from providing direct solutions to exam problems and instructs it to acknowledge uncertainty rather than fabricate information.

3.3 Self-Reflection Phase: Planned Reflection and Gamification Extensions

The components described in this section are planned extensions and have not yet been implemented. Their design illustrates the intended SRL mapping. In the self-reflection phase, future development will introduce automated reflection support and gamification mechanisms aligned with students' learning activities. A scheduler service will be designed to periodically analyze the academic calendar and generate reflection prompts for students within 24 hours after completed exams or laboratory exercises,

provided that they are enrolled in the corresponding course.

Gamification elements are planned to be integrated directly into existing student activities rather than implemented as a separate interaction flow. The proposed design includes awarding points for actions such as daily login, interaction with the AI assistant, exercise completion and viewing the progress dashboard. Badges are planned to be introduced based on predefined milestones, such as first login, first question, first completed exercise, etc. This planned approach follows the widely adopted points-and-badges model identified in the gamification literature (Khaldi et al., 2023) while aiming to reinforce self-monitoring and reflection behaviors within the SRL framework.

4 Implementation and Initial Evaluation

Since the predictive analytics component constitutes the core decision-support mechanism of the proposed architecture, the present study focuses on its initial quantitative evaluation. The remaining components are integrated into the proposed architecture, while their educational effectiveness will be investigated in future user studies.

The system is implemented in Python 3.12 (FastAPI backend) and React 18 (frontend), with SQLAlchemy as the ORM layer, Pydantic for schema validation, and JWT-based stateless authentication with bcrypt password hashing. Machine learning models are trained using scikit-learn and XGBoost and serialized as .joblib files alongside JSON metadata containing evaluation metrics and feature lists. The training dataset (students_enriched.xlsx) comprises records for 3,958 student-year observations from the 2020-2025 academic years across five first-year courses, with a chronological train/test split: models are trained on the 2020-2024 cohorts and evaluated on the 2025 cohort (663 student-year observations) to simulate real deployment conditions and avoid data leakage. Because the available dataset does not contain a persistent student identifier, the study cannot determine whether observations from different academic years correspond to the same individual. Missing values in assessment components (e.g., midterm scores for students who did not sit the exam) are imputed with zero using SimpleImputer prior to model training, reflecting the assumption that absent scores indicate non-participation rather than missing data. Five-fold cross-validation with shuffling (KFold, n_splits=5, shuffle=True, random_state=42) was applied to the training cohorts (2020–2024) for model comparison purposes. Folds were not stratified by academic year; instead, student-course records from all training years were randomly distributed across folds. Feature preprocessing is applied within each fold to prevent data leakage from the validation set into the imputation statistics. Hyperparameters were not tuned via cross-validation; instead, values were selected based on standard recommendations from the literature for tabular datasets of comparable size.

Table 2 reports both average cross-validation performance on the training cohorts (2020-2024) and held-out test performance on the 2025 cohort for the three predictive models on score-mode courses (MATAN1, LINEALG, and UUP), addressing RQ1. Two naive baselines are included for context: a training-set mean baseline, which predicts the training average score for every student (MAE = 11.49, R^2 = -0.05 on the test set at all checkpoints), and a current-score baseline, which uses only the student's accumulated score at the given checkpoint as a single predictor (not applicable before the semester, as no current-semester scores are yet available). As expected, prediction accuracy improves substantially as additional assessment information becomes available. For the three score-mode courses reported in Table 2, average test-set R^2 across tree-based models increases from 0.27 before the semester to 0.84 at mid-semester and 0.91 at the late checkpoint, suggesting that students' actual academic performance data provides increasingly informative predictors of final outcomes.

Table 2. Average CV and held-out test metrics by checkpoint and algorithm for the three score-mode courses (MATAN1, LINEALG, and UUP); XGB = XGBoost, RF = Random Forest, RR = Ridge Regression, TM = training-set mean baseline; CS = current-score baseline (student's accumulated ISVU points at checkpoint)

Alg.	before_sem		mid		late	
Cross-validation						
	MAE	R^2	MAE	R^2	MAE	R^2
XGB	9.76	0.21	3.38	0.82	2.37	0.88
RF	9.69	0.23	3.26	0.82	2.33	0.88
RR	10.29	0.23	6.56	0.61	5.68	0.68
Test						
	MAE	R^2	MAE	R^2	MAE	R^2
XGB	9.41	0.25	3.05	0.84	2.00	0.90
RF	9.23	0.29	2.98	0.84	1.83	0.91
RR	9.92	0.19	7.99	0.49	5.83	0.66
TM	11.49	-0.05	11.49	-0.05	11.49	-0.05
CS	n/a	n/a	7.45	0.55	6.55	0.64

Among the evaluated algorithms, Random Forest consistently achieves the best overall test-set performance, although its advantage over XGBoost is small (test MAE difference ≤ 0.25 points at all checkpoints). Both tree-based methods considerably outperform Ridge regression, particularly at the mid-semester checkpoint where the test MAE gap reaches approximately 5 points, suggesting that nonlinear

relationships among enrollment characteristics, continuous assessment results, and examination scores play an important role in predicting student performance. These findings are consistent with previous studies reporting the effectiveness of ensemble methods for educational prediction tasks. Importantly, both XGBoost and Random Forest outperform the training-set mean baseline at all checkpoints, and outperform the current-score baseline at the mid-semester and late checkpoints; at the mid-semester checkpoint, tree-based models achieve a test MAE of approximately 3 points compared to 7.45 for the current-score baseline, suggesting that the additional enrolment and cross-course features contribute meaningful predictive information beyond what a student's current accumulated score alone can provide.

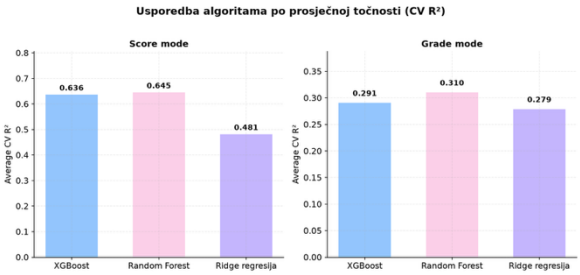


Figure 2. Comparison of predictive performance (CV R^2) across XGBoost, Random Forest, and Ridge Regression for score-mode and grade-mode courses.

To better understand the contribution of individual feature groups, a feature ablation study was performed on MATAN1 at the mid-semester checkpoint. Using only enrollment-related features yields a CV R^2 of 0.257. Adding the midterm examination score increases performance substantially to 0.908, while incorporating all remaining activity components (continuous assessment, homework, and auditory exercises) further improves performance only marginally to 0.926. This result suggests that the midterm examination is the most informative predictor among the evaluated feature groups.

Since students are presented with predictions from three different models, the prototype reports their inter-model disagreement as the spread between the highest and lowest prediction. The three agreement categories (high: $\leq$5 points; medium: 5-12 points; low: $>$12 points) are used only as interface-level descriptors rather than calibrated measures of uncertainty. At the mid-semester checkpoint, high-, medium-, and low-agreement predictions yield mean absolute consensus errors of 3.1, 4.1, and 5.0 points, respectively, with the same ordering observed at the other checkpoints. Thus, greater model disagreement is associated with higher prediction error, but the agreement measure should be interpreted as a relative disagreement signal rather than as calibrated prediction confidence. Spearman correlation between prediction spread and absolute consensus error is negligible before the semester (ρ = 0.049, p = 0.081) but increases to a strong positive association at mid- and late-semester checkpoints for score-mode courses (mid: ρ = 0.561; late: ρ = 0.717, both p < 0.001), while remaining non-significant for grade-mode courses at all checkpoints (mid: ρ = 0.050, p = 0.102), suggesting that spread functions as a reliable uncertainty proxy only when direct score prediction is possible.

An important limitation revealed by the evaluation concerns early identification of at-risk students. Although the regression models achieve satisfactory overall prediction accuracy, they perform poorly when detecting students who ultimately fail the course. Recall for failing students is zero at both the pre-semester and mid-semester checkpoints, increasing to only 0.194 at the final checkpoint (precision = 0.75; 6 of 31 failing students correctly identified). This finding reflects the optimization objective of regression models, which minimize overall prediction error rather than emphasizing minority cases. This challenge is consistent with previous findings in student-risk prediction research, where imbalanced outcomes make metrics such as recall and precision more informative than accuracy alone (Prenkaj et al., 2020). Future work should therefore complement the current regression approach with dedicated classification models and appropriate class-balancing techniques for early at-risk student detection.

5 Prototype and Discussion

The developed prototype provides students with a unified interface accessible through any web browser. Fig. 3 shows the prediction view, where the student sees predicted scores from all three models alongside the consensus and the agreement level.

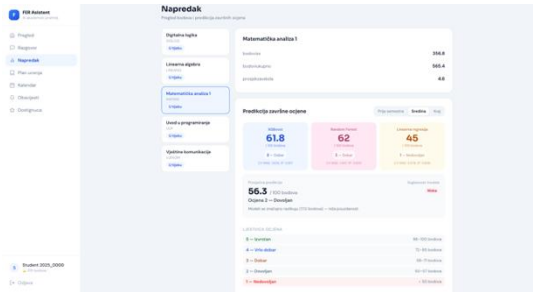


Figure 3. Prediction dashboard showing outputs of the three predictive models, consensus estimate and agreement level

The primary aim of the current stage of this research is to integrate multiple complementary AI-supported functionalities within a single student support platform. The intended practical implication of this integration is qualitatively different from what a chatbot alone can offer. For example, the prototype can present a predicted score together with model disagreement and feature-level explanations, which are

intended to provide context for subsequent study planning, while future reflection mechanisms are designed to use similar information after relevant learning activities. The planned gamification mechanisms would further connect these interactions with motivational support. The intended value of the system emerges from their integration around the SRL framework rather than from any individual component alone. The RAG architecture is intended to reduce unsupported responses by grounding generation in indexed course materials; this effect remains unevaluated. The agreement mechanism serves a similar transparent function for the prediction component: rather than presenting a single authoritative number, the system communicates its own inter-model disagreement to the student.

Limitations of the current prototype include the weaker predictive performance for courses with limited input features (DIGLOG, VJEKOM), the absence of real-user evaluation (prototype testing has not yet been conducted with students, while predictive models were evaluated using historical student data), and the inability of the current regression-based risk detection to reliably identify at-risk students before the final exam. In addition, shuffled cross-validation was used only for model comparison within the 2020–2024 training data; temporal generalization is assessed solely through the held-out 2025 cohort. The disagreement measure is a relative inter-model signal rather than calibrated uncertainty. Future work should therefore add course- and time-specific uncertainty analysis, dedicated classification for at-risk detection, broader cohorts, and user evaluation of the explanations, study plans, and planned SRL mechanisms.

6 Conclusion

This paper presented the design and initial implementation of FER AI Companion, an SRL-oriented student learning platform integrating retrieval-augmented generation, predictive learning analytics, explainable artificial intelligence, and AI-assisted study planning. Reflection support and gamification are currently represented as planned architectural extensions rather than implemented components.

Initial results suggest the technical feasibility of integrating these components into a unified web application. Experimental evaluation of the predictive models suggests that prediction accuracy improves as additional assessment data become available across three semester checkpoints, with tree-based models (XGBoost, Random Forest) outperforming Ridge regression in all evaluated conditions. The model agreement mechanism provides a relative indicator of inter-model disagreement, with high-agreement predictions exhibiting lower absolute error than low-agreement predictions.

The work presented in this paper represents the first stage of a broader research effort. While the current study focuses on the design, implementation, and technical evaluation of the platform, future work will investigate its educational effectiveness through user studies with students, evaluate the impact of explainable predictions and AI-generated study plans on self-regulated learning, and assess long-term engagement supported by the gamification mechanisms. Potential directions for future research include extending the platform to additional courses and academic years, investigating dedicated classification models for earlier identification of at-risk students, and exploring integration with existing university learning management systems.

References

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., et al. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Deakin University. (2026). Deakin's Genie: a virtual digital assistant out of the bottle. Retrieved from https://www.deakin.edu.au/about-deakin/news-and-media-releases/articles/deakins-genie-a-virtual-digital-assistant-out-of-the-bottle

Faculty of Organization and Informatics. (2026). AIVA – prvi pedagoški utemeljeni AI savjetnik za studente integriran u Moodle razvijen na FOI-ju. Retrieved from https://www.foi.unizg.hr/hr/novosti/aiva-prvi-pedagoski-utemeljeni-ai-savjetnik-za-studente-integriran-u-moodle-razvijen-na-foi

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2024). Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv. https://doi.org/10.48550/arXiv.2312.10997

Hodgson, J. (2025). How WGU Labs built an AI student support assistant. Retrieved from https://www.wgulabs.org/posts/exploring-multi-agent-and-single-agent-systems-how-labs-is-building-an-ai-powered-student-support-assistant

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., et al. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, 103, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Khaldi, A., Bouzidi, R., & Nader, F. (2023). Gamification of e-learning in higher education: A systematic literature review. Smart Learning

Environments, 10(1). https://doi.org/10.1186/s40561-023-00227-z

Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y.-S., Kay, J., Knight, S., Martinez-Maldonado, R., Sadiq, S., & Gašević, D. (2022). Explainable artificial intelligence in education. Computers and Education: Artificial Intelligence, 3, 100074. https://doi.org/10.1016/j.caeai.2022.100074

Koivisto, J., & Hamari, J. (2019). The rise of motivational information systems: A review of gamification research. International Journal of Information Management, 45, 191–210. https://doi.org/10.1016/j.ijinfomgt.2018.10.013

Lan, M., & Zhou, X. (2025). A qualitative systematic review on AI empowered self-regulated learning in higher education. npj Science of Learning. https://doi.org/10.1038/s41539-025-00319-0

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2021). Retrieval-augmented generation for knowledge-intensive NLP tasks. arXiv. https://doi.org/10.48550/arXiv.2005.11401

Lo, C. K. (2023). What Is the Impact of ChatGPT on Education? A Rapid Review of the Literature. Education Sciences, 13(4), 410. https://doi.org/10.3390/educsci13040410

Lundberg, S., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. arXiv. https://doi.org/10.48550/arXiv.1705.07874

LSA Newnan Advising Center. (2026). LSA Maizey AI advising. University of Michigan. Retrieved from https://lsa.umich.edu/advising/connect-with-an-advisor/lsa-maizey-ai-advising.html

Mainstay. (2026). Georgia State University chatbot supports every student. Retrieved from https://mainstay.com/case-study/how-georgia-state-university-supports-every-student-with-personalized-text-messaging/

McGrath, C., Farazouli, A., & Cerratto-Pargman, T. (2025). Generative AI chatbots in higher education: A review of an emerging research area. Higher Education. https://doi.org/10.1007/s10734-024-01288-w

Namoun, A. & Alshanqiti, A. (2021). Predicting Student Performance Using Data Mining and Learning Analytics Techniques: A Systematic Literature Review. Retrieved from https://www.mdpi.com/2076-3417/11/1/237

Navarro-Espinosa, A., et al. (2022). Gamification as a promoting tool of motivation for creating sustainable higher education institutions. International Journal of Environmental Research and Public Health, 19(5), 2599. https://doi.org/10.3390/ijerph19052599

Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. Frontiers in Psychology, 8, 422. https://doi.org/10.3389/fpsyg.2017.00422

Prenkaj, B., Veltri, P., & others. (2020). A Survey of Machine Learning Approaches for Student Dropout Prediction in Online Courses. Proceedings of the IEEE International Conference on Data Science and Advanced Analytics (DSAA). https://dl.acm.org/doi/10.1145/3388792

Rahal, A., & Zainuba, M. (2016). Improving students' performance in quantitative courses: The case of academic motivation and predictive analytics. The International Journal of Management Education, 14(1), 8–17. https://doi.org/10.1016/j.ijme.2015.11.003

Sailer, M., & Homner, L. (2020). The gamification of learning: A meta-analysis. Educational Psychology Review, 32(1), 77–112. https://doi.org/10.1007/s10648-019-09498-w

Sghir, M., et al. (2023). Recent advances in Predictive Learning Analytics: A decade systematic review (2012–2022). Education and Information Technologies. https://doi.org/10.1007/s10639-022-11536-0

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education—where are the educators? International Journal of Educational Technology in Higher Education, 16, 39. https://doi.org/10.1186/s41239-019-0171-0

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. Theory Into Practice, 41(2), 64–70. https://doi.org/10.1207/s15430421tip4102_2

Artificial Intelligence Disclosure and Magical Perception: An Experimental and Correlational Study of Consumer Trust in Chatbot Interactions

Claudia-Maria Miu, Delia Elena Vranceanu, Andreea Cristina Stoica, Alin Ioan Cretu, Carmen Ecaterina Catrina
Bucharest University of Economic Studies, Faculty of Marketing
Roman Square, no. 6, Bucharest, Romania
{claudia.miu, alin.cretu}@mk.ase.ro, {vranceanuelena21, stoicaandreea21}@stud.ase.ro, carmen.viasu@expand.ro

***Abstract**. The integration of AI chatbots has created a dilemma: mandated transparency undermines consumer trust. We propose disclosure acts as demystification cues, reducing the perceived AI "magic". Two online experiments (N = 200) manipulate disclosure timing (Study 1) and framing (Study 2); hypotheses were unsupported, as 86% identified the chatbot regardless of condition, so the experimental component cannot support causal claims. However, full-sample analysis revealed AI literacy was associated with lower trust via diminished magical perception (indirect = -0.030, 95% CI [-0.054, -0.014]), alongside an offsetting direct path, an inconsistent mediation pattern. Because AI literacy was measured, not manipulated, this finding is correlational and hypothesis-generating.*

Keywords. AI disclosure, chatbot, magical perception, consumer trust, AI literacy, demystification

1 Introduction

Artificial intelligence-powered chatbots have become integral to commercial service delivery, yet a growing body of evidence reveals a paradox at the heart of their deployment. Regulatory frameworks such as the European Union's AI Act increasingly mandate that consumers be informed when they are interacting with an AI system. This imperative for transparency, however, comes at a measurable commercial cost: disclosing chatbot identity consistently undermines consumer trust and engagement (Luo et al., 2019; Schilke & Reimann, 2025; Gelbrich et al., 2025). This tension constitutes what Mozafari, Weiger and Hammerschmidt (2022) termed the 'chatbot disclosure dilemma.'

Existing research has explained the negative effects of disclosure primarily through cognitive-evaluative mechanisms: reduced perceived competence (Luo et al., 2019), diminished perceptions of legitimacy (Schilke & Reimann, 2025), and the activation of algorithm aversion (Dietvorst, Simmons, and Massey, 2015). What remains largely unexplored is whether disclosure also operates through an affective pathway, specifically, whether it diminishes a sense of wonder and mystery that consumers experience when interacting with capable AI systems. Tully, Longoni and Appel (2025) recently demonstrated that consumers with lower AI literacy perceive AI as more 'magical' and that this magical perception predicts greater receptivity to AI recommendations.

This paper bridges the chatbot disclosure literature with the AI magical perception literature through a dual-process theoretical framework (Kahneman, 2011). We propose that AI identity disclosure may shift consumers from intuitive, wonder-preserving processing (System 1) to analytical, demystifying processing (System 2), thereby diminishing magical perception and, through it, trust. Two online experiments (total N = 200) test this mechanism, with AI literacy examined as an exploration moderator.

Our central empirical finding is a significant correlational mediation: AI literacy is associated with lower trust exclusively through diminished magical perception, with no direct negative effect (bootstrapped indirect effect = -0.030, 95% CI [-0.054, -0.014]). This 'inconsistent mediation' pattern, where literacy is simultaneously associated with lower trust indirectly through diminished magical perception and with higher trust directly through a separate pathway, is consistent with magical perception operating as a candidate psychological mechanism linking AI understanding to trust outcomes, though this correlational pattern cannot establish causal direction. The experimental manipulations did not produce significant main effects, a null result we interpret as an informative methodological finding about the difficulty of manipulating disclosure perceptions in AI-experienced populations.

2. Problem Statement

2.1 Chatbot disclosure and its consequences

Luo et al.'s (2019) large-scale field experiment found that disclosing chatbot identity reduced purchase rates

by approximately 80%, driven by perceptual shifts rather than performance differences. Mozafari, Weiger and Hammerschmidt (2022) demonstrated that selective self-presentation strategies could partially mitigate this penalty. Schilke and Reimann (2025) confirmed generalizability across thirteen experiments, proposing reduced legitimacy as the mechanism but leaving affective pathways untested. The Gelbrich et al. (2025) meta-analysis confirmed the robustness of the negative disclosure effect across contexts.

2.2 Trust in algorithm-mediated interactions

Trust is operationalized following Mozafari et al. (2022) as a multidimensional construct (competence, benevolence, integrity) drawing on McKnight, Choudhury and Kacmar (2002), with recent work by Lankton, McKnight and Tripp (2015) distinguishing system-like from human-like trust constructs in technology contexts. Glikson and Woolley (2020) highlighted that trust in virtual AI is particularly sensitive to transparency cues. Dietvorst, Simmons and Massey (2015) showed that consumers who observe algorithms err become dramatically less willing to use them; Jussupow, Benbasat and Heinzl (2024) recently provided an integrative perspective conceptualizing algorithm aversion and appreciation as the dual outcomes shaping human-algorithm decision-making. Jakesch, Hancock and Naaman (2023) demonstrated that people's heuristics for detecting AI-generated content are systematically flawed. Nordheim, Følstad and Bjørkli (2019) developed an initial model of chatbot trust in customer service, identifying perceived expertise and responsiveness as key chatbot-related factors alongside risk and brand perceptions. Shi et al. (2024) found that chatbot disclosure effects depend on customer expectations and service context. Longoni, Bonezzi, and Morewedge (2019) showed that consumers resist medical AI because they believe it cannot account for their unique characteristics, illustrating how perceived AI limitations interact with trust formation.

2.3 AI literacy, anthropomorphism, and magical perception

While anthropomorphism asks, 'Does this seem human?', magical perception asks 'Does this seem extraordinary?' (Epley, Waytz and Cacioppo, 2007; for a recent meta-analysis of anthropomorphism in service contexts, see Blut et al., 2021). Tully, Longoni and Appel (2025) demonstrated across eight studies that lower AI literacy predicts greater magical perception and, through it, greater AI receptivity. Risen (2016) provided the cognitive science foundation for how magical intuitions persist even when consciously rejected. AI literacy is measured using the objective test developed by Tully, Longoni and Appel (2025), based on the competencies identified by Long and Magerko (2020); the instrument and its scoring are described in Section 3.4.

2.4 Theoretical integration: Disclosure as demystification

This paper proposes that disclosure serves as an externally imposed demystification cue that diminishes magical perception and, in turn, trust. Drawing on Kahneman's (2011) dual-process framework, with recent refinements by De Neys (2023), we propose that disclosure may activate analytical processing, diminishing the sense of wonder that intuitive processing supports. The mechanism through which analytical processing diminishes magical attributions can be understood through Risen's (2016) acquiescence model: magical intuitions arise automatically but require the absence of corrective processing to persist. We test this proposition experimentally in two studies, though, as reported in Section 4, the experimental manipulations did not yield the variation needed for causal inference.

This framework draws on Nass and Moon's (2000) CASA paradigm, formally extended by Lombard and Xu (2021) as the Media Are Social Actors (MASA) paradigm and by Gambino, Fox and Ratan (2020), who argued that people have developed human-media scripts (learned interaction patterns specific to chatbots), distinct from human-human scripts. Puntoni et al. (2021) identified a tension between connection and alienation in AI social experiences, proposing that the timing of disclosure may influence alienation (RQD5) and that growing familiarity may reshape sensitivity to AI interventions (RQA4). Together, these suggest that demystification may operate differently depending on prior AI exposure, consistent with Hancock, Naaman and Levy's (2020) framework for AI-mediated communication.

3. Research questions and hypotheses

RQ1: Does AI identity disclosure reduce consumers' magical perception of AI chatbots?
RQ2: Does magical perception mediate the relationship between AI literacy and consumer trust?
RQ3 (exploratory): Does AI literacy moderate the effect of disclosure on magical perception?
H1: Participants receiving disclosure after the interaction will report higher magical perception than those receiving it before. This prediction draws on Zajonc's (1980) affective primacy hypothesis and Ross, Lepper, and Hubbard's (1975) work on belief perseverance.
H2: Magical perception mediates the relationship between disclosure, timing and trust.
H3: Participants exposed to ambient framing will report higher magical perception than those exposed to

prominent framing. This prediction follows from the demystification framework outlined in Section 2.4: prominent labelling foregrounds the system's artefactual nature and prompts analytical processing, whereas ambient labelling permits intuitive engagement to persist.
H4: Magical perception mediates the relationship between disclosure framing and trust.

Table 1. Participant demographics by condition

Variable	Total	Sunrise	Horizon	Atlas	Compass
Female	131 (65.5 %)	33 (66%)	30 (60%)	33 (66 %)	35 (70%)
Age 18-29	137 (68.5 %)	33 (66%)	33 (66%)	31 (62 %)	40 (80%)
Liceal	45 (22.5 %)	13 (26%)	7 (14%)	9 (18 %)	16 (32%)
AI Lit M(SD)	13.01 (3.18)	12.44 (3.07)	12.82 (3.32)	13.24 (3.13)	13.52 (3.16)

Note. Gender χ^2=1.13, p=.82; Age χ^2=7.39, p=.60; Education χ^2=10.38, p=.34; AI Literacy F(3,196)=1.11, p=.35.

4. Research Methods

4.1 Design

Two between-subjects online experiments (total N = 200). Study 1 (N = 100): disclosure timing (Before n = 50 vs After n = 50). Study 2 (N = 100): disclosure framing (Prominent n = 50 vs Ambient n = 50).

4.2 Participants

Two hundred participants (131 female, 69 male) were recruited via convenience sampling from the Bucharest University of Economic Studies and social media. The sample was predominantly young adults (68.5% aged 18-29). Demographic equivalence across conditions was confirmed (Table 1; all ps > .10).

4.3 Platform and procedure

TravelSmart (travelsmart-interview.vercel.app) simulates interactions with a commercial travel-planning chatbot. Procedure: (1) consent, (2) AI literacy test (18 items: 17 substantive items plus 1 attention check), (3) experimental manipulation, (4) chatbot conversation, (5) post-interaction scales, (6) manipulation checks and demographics.

4.4 Measures

AI literacy. An objective multiple-choice test developed by Tully, Longoni and Appel (2025), based on the AI literacy competencies identified by Long and Magerko (2020). The instrument contained 18 items administered to participants (17 substantive AI literacy items plus 1 attention-check item used to flag inattentive responding); the AI literacy score reported throughout is the sum of the 17 substantive items, with a possible range of 0-17. M = 13.01, SD = 3.18, observed range 5-17.

Magical perception. Six items adapted from Tully et al. (2025). 7-point Likert. α = .818.

Trust. Six items from Mozafari et al. (2022). 7-point Likert. α = .774.

Manipulation checks. MC1a: AI knowledge (Yes/No). MC2: Awareness (7-point). MC3 (Study 2): Label noticed.

4.5 Analysis strategy

The analysis proceeds in two stages. First, H1-H4 are tested using independent-samples t-tests and bootstrapped mediation following Hayes (2022), using 5,000 bootstrap resamples to construct bias-corrected confidence intervals for indirect effects. Second, a correlational mediation test of AI literacy → magical perception → trust in the full sample. Correlational mediation serves as the primary evidential contribution; experimental tests are informative secondary analyses. For the four primary hypothesis tests (H1-H4) and the associated mediation paths, Bonferroni-adjusted thresholds ($p < .0125$) are reported alongside conventional p-values in Section 4.4; exploration and descriptive analyses are flagged as such throughout.

These two analyses carry different evidentiary support. H1-H4 involves random assignment to disclosure condition; had the manipulation produced the intended contrast, between-condition comparisons would support causal inference about disclosure. The AI literacy analysis, by contrast, treats literacy as a measured individual-difference variable rather than an assigned one, so it is necessarily correlational: it cannot establish that literacy causes reduce magical perception and lower trust. Plausible alternative explanations include reverse causation (lower trust may motivate some consumers to seek out AI knowledge, raising their literacy scores) and third-variable accounts (a general disposition toward skepticism could independently lower both magical perception and trust). We will return to this distinction when interpreting the findings in Section 6.

5. Findings

5.1 Preliminary analyses

Both scales demonstrated adequate internal consistency: magical perception (α = .818) and trust (α = .774). No individual item deletion improved alpha for the magical perception scale. Shapiro-Wilk tests indicated departures from normality for magical perception in all conditions (all ps < .05). Non-parametric Mann-Whitney U tests yielded substantively identical results.

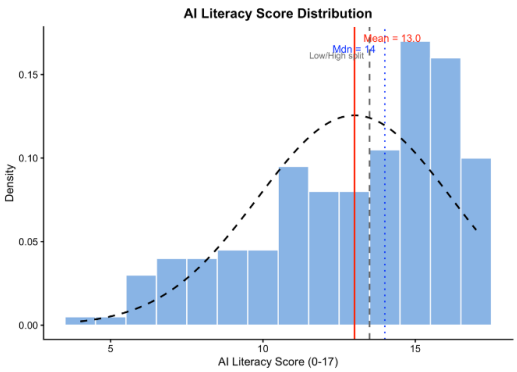


Figure 1. AI literacy score distribution (N = 200). Dashed line = Low/High split (≤13/≥14). M = 13.01, Mdn = 14.

The AI literacy distribution (Fig. 1) showed adequate variance (M = 13.01, SD = 3.18, range 5-17), with 39.5% scoring ≤13 (Low) and 60.5% scoring ≥14 (High).

5.2 Manipulation checks

Table 2. Manipulation check results

Study	MC	Cond A	Cond B	Statistic	p
Study 1	Awareness	6.14 (1.20)	6.12 (1.27)	t(97.6) =0.08	.94
Study 1	Knew AI	76%	82%	-	ns
Study 2	Awareness	6.18 (1.16)	6.42 (1.16)	t(98)=-1.04	.30
Study 2	Knew AI	90%	96%	-	ns
Study 2	Label	88%	68%	χ^2=14.62	<.001

Note. Cond A = Before/Prominent; Cond B = After/Ambient.

The Study 2 label-salience manipulation succeeded at the surface level: 88% of Prominent participants noticed the label compared with 68% in the Ambient condition (χ^2 = 14.62, p < .001). However, this surface-level salience difference did not translate into perceptually distinct interaction experiences: AI-awareness ratings and 'knew it was AI' rates did not differ across conditions in either study (Table 2). Among 'After' participants in Study 1, the most common explanation for knowing it was AI was 'I assumed based on my previous experience with chatbots' (30%), consistent with Gambino et al.'s (2020) prediction that experienced users apply learned chatbot-identification scripts. This pattern itself constitutes a substantive finding: in AI-experienced populations, pre-formed assumptions about chatbot identity override surface-level disclosure cues, indicating that the human-media scripts identified by Gambino, Fox and Ratan (2020) operate independently of, and prior to, the explicit cues researchers manipulate. The implication is that disclosure manipulations cannot bypass these scripts once consumers have acquired baseline AI familiarity, a methodological constraint that shapes interpretation of the null experimental results reported in Section 5.3.

5.3 Experimental effects (H1-H4)

Table 3. Hypothesis tests

Test	Cond A M(SD)	Cond B M(SD)	t	p	d	95% CI
H1: Magic	2.60 (1.06)	2.89 (1.13)	1.32	.190	0.26	[-.14,.72]
H3: Magic	2.61 (0.99)	2.88 (1.10)	1.28	.206	0.26	[-.14,.68]
Trust S1	4.58 (0.88)	4.78 (0.89)	1.15	.253	0.23	-
Trust S2	4.62 (0.77)	4.67 (0.87)	0.35	.730	0.07	-

Note. Cond A = Before/Prominent; Cond B = After/Ambient.

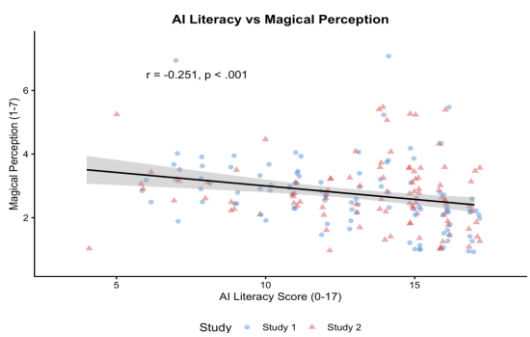


Figure 2. Magical perception and trust by condition (both studies). Error bars = ±1 SE.

H1 predicted that late disclosure would preserve higher magical perception. The After condition (M = 2.89) did not differ significantly from Before (M = 2.60), t(97.6) = 1.32, p = .190, d = 0.26. H1 is not supported.

H3 predicted that ambient framing would preserve higher magical perception. The Ambient condition (M = 2.88) did not differ significantly from Prominent (M = 2.61), t(96.8) = 1.28, p = .206, d = 0.26. H3 is not supported (Table 3, Fig. 2). Sensitivity analysis: N = 50 per cell provides 80% power for d ≥ 0.57; the observed d = 0.26 would require approximately 234 per cell.

H2 predicted that magical perception mediates the effect of disclosure timing on trust. The condition-based indirect effect in Study 1 was non-significant (indirect = 0.106, 95% CI [-0.035, 0.303]). H2 is not supported.

H4 predicted that magical perception mediates the effect of disclosure framing on trust. The condition-based indirect effect in Study 2 was also non-significant (indirect = 0.070, 95% CI [-0.018, 0.221]). H4 is not supported. Both null mediations are consistent with the failed a-paths: because the experimental conditions did not significantly affect magical perception, condition-based mediation could not emerge.

5.4 Correlational mediation: AI literacy → magical perception → trust

Table 4. Pearson correlations (N = 200)

	AI Literacy	Magic	Trust
AI Literacy	-		
Magic	-.251***	-	
Trust	.049	.403***	-

****p < .001. Full matrix including demographics: see Figure 6.*

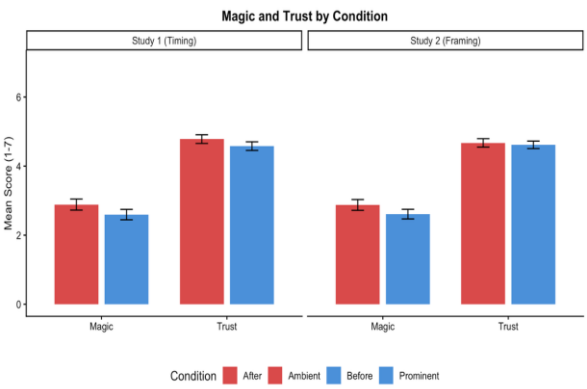


Figure 3. AI literacy vs magical perception by study. r = -.251, p < .001.

As shown in Table 4, AI literacy correlated negatively with magical perception (r=-.251, p<.001; Fig. 3) and magical perception correlated positively with trust.

Table 5. Bootstrapped mediation: AI Literacy → Magical Perception → Trust

Path	b	SE	p	95% CI
a (Literacy→Magic)	-0.085	0.023	<.001	
b (Magic→Trust)	0.351	0.053	<.001	
c (total)	0.013	0.019	.488	
c' (direct)	0.043	0.018	.017	[0.008, 0.078]
Indirect (a×b)	-0.030	0.010	-	[-0.054, -0.014]

Note. 5,000 bootstrap resamples—bias-corrected CIs.

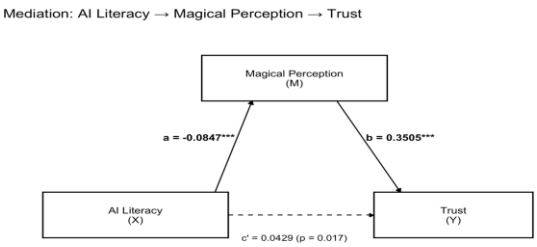


Figure 4. Mediation path diagram: AI Literacy → Magical Perception → Trust. ***p < .001.

This constitutes 'inconsistent mediation' (MacKinnon, Fairchild and Fritz, 2007): the indirect and direct effects operate in opposite directions, producing a near-zero total effect (Table 5, Fig. 4). In suppression patterns like this, AI literacy is associated with lower trust through diminished magical perception (a negative indirect path). In comparison, it is independently associated with higher trust via a separate positive direct path. The positive direct effect (c' = 0.043, p = .017) may reflect a confidence effect: individuals who understand AI may trust its outputs more because they recognize its computational strengths. This direct effect also held when controlling demographics (c' = 0.044, p = .014). Under a Bonferroni correction for the four primary hypothesis tests (H1-H4), the adjusted threshold is p < .0125; the indirect effect (p < .001) survives this threshold, while the c' direct effect (p = .017) does not. Interpretation of the direct path, therefore, proceeds with appropriate caution.

The full-sample mediation survived demographic controls (indirect = -0.031, 95% CI [-0.055, -0.014]) and exclusion of straight-line respondents (n = 20 removed; indirect = -0.021, 95% CI [-0.043, -0.007]). The attenuation of the indirect effect upon excluding

straight-liners suggests these respondents contributed disproportionately to the effect magnitude.

We consider the full-sample literacy-based mediation (N = 200) as the primary estimate, while acknowledging that the literacy-magic correlation differed between Study 1 (r = -.38) and Study 2 (r = -.15), introducing heterogeneity. The stronger relationship in Study 1 should be weighed more heavily in interpretation.

5.5 Exploratory moderation (Study 1)

A 2×2 ANOVA (condition × literacy group) yielded a non-significant interaction, $F(1, 96) = 0.63$, $p = .43$, $\eta^2p = .007$. Among low-literacy participants (≤13), the After condition (M = 3.35) scored higher than Before (M = 2.83), $t(44) = 2.17$, $p = .035$, $d = 0.63$ (Fig. 5). Among high-literacy participants, no difference emerged ($p = .58$, $d = 0.15$). Given the non-significant interaction, this is hypothesis-generating rather than confirmatory.

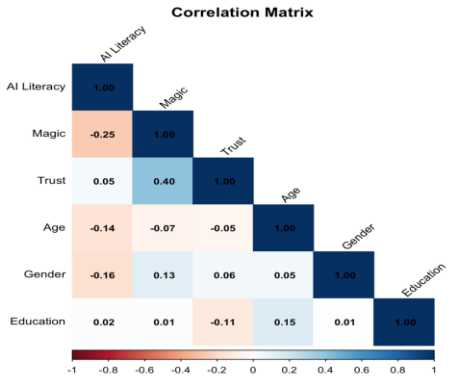


Figure 5. Magical perception by disclosure timing and AI literacy (Study 1). Error bars = ±1 SE.

5.6 Trust subdimensions (exploratory)

Magical perception was most strongly associated with integrity ($r = .434$, $p < .001$), then competence ($r = .332$, $p < .001$), then benevolence ($r = .233$, $p = .001$) (Fig. 6).

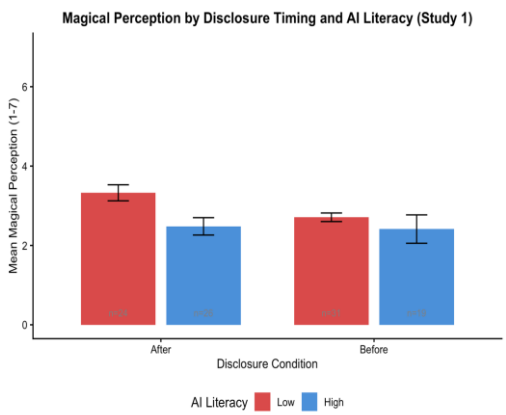


Figure 6. Correlation matrix heatmap (N = 200).

6. Conclusions

6.1 Summary

The four experimental hypotheses (H1-H4) were not supported: neither disclosure timing nor framing significantly affected magical perception, and the null a-paths rendered mediation hypotheses H2 and H4 untestable via the experimental design. The correlational evidence, however, supports the theoretical model: AI literacy is associated with lower trust exclusively through diminished magical perception (indirect = -0.030, 95% CI [-0.054, -0.014]). This mediation survives demographic controls and exclusion of suspect respondents.

6.2 Why the manipulations failed

Manipulation strength. Interface cues signaled AI identity regardless of disclosure. Future designs should create genuine ambiguity using human-mimicking interfaces and free-text conversation.

Prior experience and human-media scripts. Gambino et al.'s (2020) extended CASA framework predicts that experienced users develop chatbot-specific scripts, including default AI categorization. In our data, 30% of 'After' participants assumed AI based on prior experience. Puntoni et al. (2021) anticipated this. Zhu et al. (2025) empirically demonstrated that post-ChatGPT users interact with chatbots differently.

AI-literate sample. Mean literacy = 13.01/17 (76.5%). The significant effect among low-literacy participants ($d = 0.63$, $p = .035$) supports this interpretation. Other unmeasured mechanisms may also contribute to the observed pattern among low-literacy participants and represent directions for future research.

6.3 Study 2 and within-study mediation

Within-study condition-based mediations were non-significant in both studies, consistent with null a-paths. The literacy-magic correlation was substantially weaker in Study 2 (r = -.15) than in Study 1 (r = -.38). Pooling across studies with such different correlations introduces heterogeneity. Study 2's manipulation also simultaneously varied visual salience, register, and content. The Study 1 result should be weighted more heavily, given its stronger internal relationships. The weaker literacy-magic correlation in Study 2 may reflect a confounded manipulation that simultaneously varied visual salience, linguistic register, and content, potentially introducing noise that attenuated the relationship between literacy and magical perception.

6.4 Theoretical contributions

First, the correlational findings are consistent with magical perception operating as a candidate mediating mechanism, complementing cognitive accounts (Luo

et al., 2019; Schilke & Reinmann, 2025; Mozafari et al.,20222) with a possible affective pathway; because this pattern was not tested experimentally, it should be treated to pending future causal tests. Second, the inconsistent mediation pattern is compatible with an interpretation in which understanding AI is not inherently associated with lower trust and associated with lower trust only through diminished magical perception. The positive direct path is one plausible account of why higher AI literacy might otherwise be associated with trust, though we did not test this directly and other explanations can't be ruled out.

Beyond these primary contributions, the null experimental results, interpreted through Gambino et al.'s (2020) framework and Puntoni et al.'s (2021) experiential perspective, suggest that disclosure manipulations effective in earlier research may become inert as consumer AI exposure increases.

6.5 Practical implications

The correlational findings suggest a possible basis for segmenting disclosure strategy by customer AI-familiarity: because magical perception was concentrated among lower-literacy participants (Section 5.5), firms operating in markets with heterogeneous AI exposure may see more consistent trust effects from disclosure design among less AI-experienced customer segments than among highly AI-literate ones, though this segmentation logic was not tested experimentally and remains a proposal for future work. Second, the failure of both disclosure manipulations to produce a perceptible contrast is itself a practical signal for regulators and platform designers: in populations with substantial prior chatbot exposure, standard disclosure cues such as timing and label prominence may already be functionally redundant, since consumers default to assuming AI involvement regardless of the specific cue presented. This suggests that regulatory transparency requirements may need to be evaluated against a moving baseline of consumer AI familiarity rather than assumed to have constant effects over time. These implications are offered cautiously, given the exploratory and largely null status of the experimental evidence reported here.

6.6 Limitations and future research

Several limitations warrant discussion. First, the convenience sample (68.5% aged 18-29) limits generalizability. Second, experimental manipulations did not produce perceptually distinct interaction experiences despite a successful surface-level salience manipulation in Study 2. Third, Study 2's manipulation simultaneously varied salience, register, and content. Fourth, no separate measure of awe was included. Fifth, the dual-process mechanism (System 1/System 2) is theoretically motivated but not directly measured. Sixth, the cross-sectional design does not permit causal inference. Seventh, although the four primary tests are corrected via Bonferroni, exploratory and descriptive analyses are not corrected and should be interpreted with appropriate caution. Eighth, the literacy-magic correlation differed substantially between studies.

Future research should test stronger manipulations in populations with varied AI exposure, include general-population samples, examine cultural contexts, and test the full serial mediation chain (disclosure → magical perception → awe → trust).

References

Blut, M., Wang, C., Wünderlich, N.V. and Brock, C. (2021) 'Understanding anthropomorphism in service provision: A meta-analysis of physical robots, chatbots, and other AI', Journal of the Academy of Marketing Science, 49(4), pp. 632-658. doi: 10.1007/s11747-020-00762-y.

De Neys, W. (2023) 'Advancing theorizing about fast-and-slow thinking', Behavioral and Brain Sciences, 46, e111. doi: 10.1017/S0140525X2200142X.

Dietvorst, B.J., Simmons, J.P. and Massey, C. (2015) 'Algorithm aversion: People erroneously avoid algorithms after seeing them err', Journal of Experimental Psychology: General, 144(1), pp. 114-126. doi: 10.1037/xge0000033.

Epley, N., Waytz, A. and Cacioppo, J.T. (2007) 'On seeing human: A three-factor theory of anthropomorphism', Psychological Review, 114(4), pp. 864-886. doi: 10.1037/0033-295X.114.4.864.

Gambino, A., Fox, J. and Ratan, R.A. (2020) 'Building a stronger CASA: Extending the computers are social actors paradigm', Human-Machine Communication, 1, pp. 71-86. doi: 10.30658/hmc.1.5.

Gelbrich, K., Roschk, H., Miederer, S. and Kerath, A. (2025) 'Automated versus human agents: A meta-analysis of customer responses to robots, chatbots, and automated agents', Journal of the Academy of Marketing Science, 53, pp. 288-317. doi: 10.1007/s11747-024-01040-9.

Glikson, E. and Woolley, A.W. (2020) 'Human trust in artificial intelligence: Review of empirical research', Academy of Management Annals, 14(2), pp. 627-660. doi: 10.5465/annals.2018.0057.

Hancock, J.T., Naaman, M. and Levy, K. (2020) 'AI-mediated communication: Definition, research agenda, and ethical considerations', Journal of Computer-Mediated Communication, 25(1), pp. 89-100. doi: 10.1093/jcmc/zmz022.

Hayes, A.F. (2022) Introduction to mediation, moderation, and conditional process analysis: A regression-based approach. 3rd edn. New York: Guilford Press.

Jakesch, M., Hancock, J.T. and Naaman, M. (2023) ‘Human heuristics for AI-generated language are flawed’, Proceedings of the National Academy of Sciences, 120(11), e2208839120. doi: 10.1073/pnas.2208839120.

Jussupow, E., Benbasat, I. and Heinzl, A. (2024) 'An integrative perspective on algorithm aversion and appreciation in decision-making', MIS Quarterly, 48(4), pp. 1575-1590.

Kahneman, D. (2011) Thinking, fast and slow. New York: Farrar, Straus and Giroux.

Lankton, N.K., McKnight, D.H. and Tripp, J. (2015) 'Technology, humanness, and trust: Rethinking trust in technology', Journal of the Association for Information Systems, 16(10), pp. 880-918. doi: 10.17705/1jais.00411.

Lombard, M. and Xu, K. (2021) 'Social responses to media technologies in the 21st century: The Media are Social Actors paradigm', Human-Machine Communication, 2, pp. 29-55. doi: 10.30658/hmc.2.2.

Long, D. and Magerko, B. (2020) ‘What is AI literacy? Competencies and design considerations’, in Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. New York: ACM, pp. 1-16. doi: 10.1145/3313831.3376727.

Longoni, C., Bonezzi, A. and Morewedge, C.K. (2019) ‘Resistance to medical artificial intelligence’, Journal of Consumer Research, 46(4), pp. 629-650. doi: 10.1093/jcr/ucz013.

Luo, X., Tong, S., Fang, Z. and Qu, Z. (2019) ‘Frontiers: Machines vs. humans: The impact of artificial intelligence chatbot disclosure on customer purchases’, Marketing Science, 38(6), pp. 937-947. doi: 10.1287/mksc.2019.1192.

MacKinnon, D.P., Fairchild, A.J. and Fritz, M.S. (2007) ‘Mediation analysis’, Annual Review of Psychology, 58, pp. 593-614. doi: 10.1146/annurev.psych.58.110405.085542.

McKnight, D.H., Choudhury, V. and Kacmar, C. (2002) ‘Developing and validating trust measures for e-commerce: An integrative typology’, Information Systems Research, 13(3), pp. 334-359. doi: 10.1287/isre.13.3.334.81.

Mozafari, N., Weiger, W.H. and Hammerschmidt, M. (2022) ‘Trust me, I’m a bot - repercussions of chatbot disclosure in different service frontline settings’, Journal of Service Management, 33(2), pp. 221-245. doi: 10.1108/JOSM-10-2020-0380.

Nass, C. and Moon, Y. (2000) ‘Machines and mindlessness: Social responses to computers’, Journal of Social Issues, 56(1), pp. 81-103. doi: 10.1111/0022-4537.00153.

Nordheim, C.B., Følstad, A. and Bjørkli, C.A. (2019) 'An initial model of trust in chatbots for customer service: Findings from a questionnaire study', Interacting with Computers, 31(3), pp. 317-335. doi: 10.1093/iwc/iwz022.

Puntoni, S., Reczek, R.W., Giesler, M. and Botti, S. (2021) ‘Consumers and artificial intelligence: An experiential perspective’, Journal of Marketing, 85(1), pp. 131-151. doi: 10.1177/0022242920953847.

Risen, J.L. (2016) ‘Believing what we do not believe: Acquiescence to superstitious beliefs and other powerful intuitions’, Psychological Review, 123(2), pp. 182-207. doi: 10.1037/rev0000017.

Ross, L., Lepper, M.R. and Hubbard, M. (1975) ‘Perseverance in self-perception and social perception: Biased attributional processes in the debriefing paradigm’, Journal of Personality and Social Psychology, 32(5), pp. 880-892. doi: 10.1037/0022-3514.32.5.880.

Schilke, O. and Reimann, M. (2025) ‘The transparency dilemma: How AI disclosure erodes trust’, Organizational Behavior and Human Decision Processes, 188, 104405. doi: 10.1016/j.obhdp.2025.104405.

Shi, S., Gong, Y. and Gursoy, D. (2024) ‘Chatbot service encounters: An examination of AI chatbot disclosure’s impact’, International Journal of Hospitality Management, 116, 103617. doi: 10.1016/j.ijhm.2023.103617.

Tully, S.M., Longoni, C. and Appel, G. (2025) ‘Lower artificial intelligence literacy predicts greater AI receptivity’, Journal of Marketing, 89(5), pp. 1-20. doi: 10.1177/00222429251314491.

Zajonc, R.B. (1980) ‘Feeling and thinking: Preferences need no inferences’, American Psychologist, 35(2), pp. 151-175. doi: 10.1037/0003-066X.35.2.151.

Zhu, J., Dong, A., Wang, C., Veldhuizen, S., Abdelwahab, M., Brown, A., Selby, P. and Rose, J. (2025) ‘The impact of ChatGPT exposure on user interactions with a motivational interviewing chatbot: Quasi-experimental study’, JMIR Formative Research, 9, e56973. doi: 10.2196/56973.

Trends in Software Engineering

Melita Draganić, Robertina Zdjelar, Velibor Božić
The Role of Artificial Intelligence in Objectifying the CMM Model for Assessing the Maturity of Information Systems

Maja Benkus, Neven Vrček
The Adoption of Vibe Coding and the Transformation of Programming Practices Induced by Generative Models: A SonarQube-Based Analysis

Sandra Melovic, Dusanka Dakic, Darko Stefanovic, Teodora Vuckovic
SwiftUI Adoption in Enterprise Mobile Development: Evaluating the Impact of Incremental UIKit to SwiftUI Migration on Presentation-Layer Complexity

Matúš Nečas, Juraj Ďuďák, Roman Budjač
A State-Indexed Execution Model for Embedded Systems

Tara Nikolić, Dario Ljubas, Goran Nikšić
Detection without disposition: Security technical debt, scanner-context artefacts, and the triage gap in SME GitHub repositories

The Role of Artificial Intelligence in Objectifying the CMM Model for Assessing the Maturity of Information Systems

Melita Draganić, Robertina Zdjelar
University of Zagreb
Faculty of Organization and Informatics
Pavlinska 2, Varaždin, Croatia
{melita.draganic, rzdjelar}@foi.unizg.hr

Velibor Božić
General Hospital Koprivnica
Željka Selingera 1, Koprivnica, Croatia
vebozic@obkoprivnica.hr

Abstract. This study addresses subjectivity in Capability Maturity Model (CMM) assessments by proposing a hybrid artificial intelligence approach that combines process mining and natural language processing. Using event logs, system records, and process documentation, the model converts qualitative characteristics into measurable indicators and maps them to maturity levels through weighted scoring. Developed through design science and tested on healthcare information system processes, it produced more consistent, transparent, and traceable results than expert evaluations. The approach supports continuous and predictive assessment while preserving CMM principles. It advances information systems maturity, objectification, machine learning, and process analytics in organizational evaluation practices.

Keywords: artificial intelligence, CMM, information systems maturity, objectification, machine learning, process analytics.

1 Introduction

The effectiveness of organizational units depends on the business value of information systems (IS), their support for the organization's business goals and maintaining competitiveness. Organizational changes affect business models, decision-making and effective use of IS. The purpose of the paper is to apply Capability Maturity Models (CMM) in the context of the software development process, and to explore the role of artificial intelligence (AI) in objectifying the results of the CMM framework. The CMM framework is based on the Plan-Do-Check-Act (PDCA) concept [3] by following the steps: an assessment the current maturity level of information system, a comparison between the current maturity levels and the target values, and defining improvements based on the gap analysis. In the application of the CMM framework to assess the maturity of information systems within a company there are limitations regarding the objectiveness of the results. The output results of the assessment using the CMM framework are mostly based on subjective perceptions and responses when it is necessary to decide on the level of maturity for a particular business aspect and management. In the conditions of application of new digital technologies, a large amount of data is generated, such as information, system records and digital records of business processes. Accordingly, the main idea of this study is to investigate the role of AI in the analysis of large amounts of data as well as the generation of new patterns for a more objective assessment of the capability of business processes in relation to the CMM framework. The logic of transforming qualitative assessments into quantitative maturity levels is based on assigning weighting factors to parameters such as process mining and natural language processing (NLP) that process input data and check their compliance with the CMM standard. The main research goals are: (1) defining the criteria and logic of transforming subjective assessments into quantitative data that represent the maturity level according to the CMM framework; (2) building an AI model that links process data and defined transformation logic to CMM maturity levels; (3) comparing AI-supported assessments with CMM framework assessment results; and (4) examining whether maturity assessments can be conducted in real-time monitoring. The paper is based on three research hypotheses. H1: AI-supported assessments are statistically more consistent than traditional CMM assessments. H2: Combining process mining and natural language processing improves the accuracy of CMM level classification. H3: AI models more accurately identify changes between maturity levels in relation to the static CMM framework. This paper is structured as follows. Section 2 provides the theoretical background, covering maturity models and

the CMM framework, the design science research perspective, the role of process analytics as an empirical basis, and the application of natural language processing and predictive AI in maturity assessment. Section 3 describes the research methodology, including the hybrid model design approach, the supervised machine learning techniques employed, and the evaluation strategy based on quantitative and comparative analysis. Section 4 presents the empirical results obtained through a case study conducted in a healthcare organization, including the operationalization of CMM criteria into quantitative variables and the measurement of outcomes across six selected business processes. Section 5 discusses the findings in relation to the three research hypotheses and addresses both theoretical and practical implications. Section 6 acknowledges the limitations and outlines directions for future research, while Section 7 concludes the paper.

2Theoretical Background and Related Research

The Capability Maturity Model (CMM) is a staged maturity framework used to assess the degree to which organizational processes are defined, managed, measured, and continuously improved. In its classical form, it distinguishes maturity levels from ad hoc and individually dependent processes to optimized processes characterized by quantitative management and continuous improvement. In this paper, CMM is used as the reference framework for evaluating the maturity of information-system-supported business processes.

2.1 Maturity Models and Capability Maturity Model

Maturity models are commonly used to describe how organizations progressively improve their capabilities, particularly in information systems, software engineering, and process management [13]. The core proposition is that organizations do not improve uniformly, but typically advance through a series of discrete stages [1], each representing more structured, consistently managed, and measurable practices. The Capability Maturity Model (CMM) was originally developed to introduce greater discipline and predictability into software and information system development. Over time, it evolved into the Capability Maturity Model Integration (CMMI), which generalized the core concepts and sought to harmonize development practices across organizational functions. At its core, the model represents process capability through five maturity levels, from chaotic and unpredictable practices to quantitatively managed and continuously improving ones. Despite its broad adoption, the model has attracted critical commentary. A substantial body of literature identifies both methodological and epistemological limitations. One of the most frequently raised concerns is the subjectivity inherent in the assessment process. Research by Mettler [4] and Becker et al. [5] points out that results can be difficult to replicate across different organizational settings or time periods.

2.2 Maturity Models in the Context of Design Science

Maturity models have increasingly been examined through the lens of Design Science Research. Within this perspective, the objective shifts from describing organizational phenomena to designing artefacts that address concrete operational problems. Röglinger and colleagues [2], [6] argue that contemporary maturity models must meet three core requirements: they must be grounded in empirical evidence; they must be adaptable to diverse organizational contexts; and they should support continuous assessment rather than relying solely on periodic evaluations. The implemented model was an interpretable hybrid rule-based scoring model rather than a data-intensive machine-learning classifier. Process-mining and NLP-derived indicators were normalized to a 0–100 scale and aggregated using weights of 0.6 and 0.4, respectively. The initial weights and maturity thresholds were defined based on CMM criteria and expert consensus and were subsequently calibrated against historical expert assessments. Earlier assessment periods were used for calibration, while the most recent observations were retained for validation. Therefore, the model did not undergo conventional parameter training; instead, its weights and thresholds were calibrated and validated using historical labelled data.

2.3 Process Analytics as an Empirical Basis

To truly understand how an organization operates, it is insufficient to rely solely on process descriptions or individuals' perceptions of how things work. Process analytics represents a significant shift towards more evidence-based evaluations. Rather than depending primarily on interviews or retrospective interpretations, it draws on the digital traces left by information systems. Through event logs [7] and system records, the actual execution of processes can be reconstructed. Analyzing this data enables the identification of patterns that might otherwise remain undetected, including objective metrics such as process duration, deviation frequency from expected execution paths, and the level of automation.

2.4 Natural Language Processing and Predictive AI in Maturity Assessment

One persistent challenge in CMM assessments concerns determining how effectively organizational processes are documented and formalized. Natural Language Processing (NLP) techniques enable the systematic examination of large document sets to extract insights about their structure, content, and semantic coherence [8]. When applied to maturity assessments, these methods enable a more thorough examination of documentation at a scale that would be difficult to achieve through manual review alone. Recent advances in AI are increasingly demonstrating the potential for predictive analysis [9, 10]. By examining operational data over time, AI-supported models can identify patterns that typically precede changes in maturity levels, suggesting that maturity models could gradually transition from static assessment instruments to predictive mechanisms.

2.5 Research Gap

Furthermore, the question of whether AI-based maturity assessments can serve a predictive function — identifying forthcoming changes in maturity levels before they are formally recognized in periodic evaluations — has received limited empirical treatment. This gap is particularly significant given the operational demands of modern organizations, which increasingly require real-time rather than retrospective insight into their process performance. The present study addresses this gap through the design and empirical evaluation of a hybrid model that links process mining indicators and NLP-derived documentation scores to CMM maturity levels, while also examining the model's capacity for predictive assessment. The literature review revealed a gap between traditional maturity models such as CMM and the data-driven methods that are becoming increasingly influential in information systems management. What is notably absent is a robust model that links empirical indicators — derived from operational data — with the maturity levels defined in the CMM framework. This study aims to address that gap by developing and empirically evaluating a hybrid AI-supported maturity assessment model.

3 Research Methodology

3.1 Model Architecture, Features and Training Procedure

Table 1. Description of model

Component	Input data	Extracted features	Output
Process mining	Event logs	Process duration, deviation rate, conformance rate, automation ratio, control-point frequency	PM compliance score
NLP	Process documentation	Process-step coverage, role clarity, input/output definition, and control-point documentation.	NLP compliance score
Scoring module	PM and NLP scores	Weighted aggregation	Maturity score 0-100
Classification / mapping	Weighted score	CMM range thresholds	ML 0-5

The hybrid model consists of three analytical layers. The first layer extracts process mining indicators from event logs, including process duration variability, path conformance, deviation frequency, automation ratio, rework frequency, and the density of formal control points. The second layer applies NLP-based analysis to formal process documentation. The third layer aggregates these indicators into a normalized maturity score and maps the score to the corresponding CMM maturity level. The research adopts a design science research approach [11], which is particularly suited for studies that aim to produce and evaluate functional artefacts rather than purely theoretical constructs. The primary objective is to design, implement, and assess a maturity assessment model enhanced by AI, accompanied by a supporting user interface. The model was operationalised and evaluated through a real-world case study conducted on data originating from a healthcare business information system. The evaluation strategy combined quantitative and comparative analysis. Quantitative assessment relied on statistical methods to measure the consistency and alignment between the AI-generated maturity estimates and those produced through traditional CMM assessments conducted by certified evaluators. Historical CMM evaluation results served as the reference benchmark.

3.2 AI-Supported Feature Extraction and Rule-Based Scoring

In line with the Design Science Research perspective described in Section 2.2, the proposed model is implemented as an interpretable assessment artefact that combines AI-supported feature extraction with explicit rule-based scoring. Process mining is used to extract quantitative indicators from event logs, while NLP is used to identify and quantify predefined elements in process documentation. The final maturity level is not generated by a trained supervised machine-learning classifier. Instead, the extracted indicators are normalized

and aggregated through a transparent weighted-scoring procedure. Therefore, the approach is described as AI-supported rather than as an AI classifier.

All extracted features were transformed to a common 0–100 scale before aggregation. Ratio-based indicators, including automation rate, conformance rate, documentation coverage, role clarity, input/output definition, and control-point documentation, were expressed directly as percentages. For indicators where lower values indicate higher maturity, such as process-duration variability and the coefficient of variation, reverse normalization was applied against predefined historical benchmarks. Control-point density was normalized relative to the target value established from historical process data and expert validation. Indicators within each analytical channel were equally weighted to calculate the process-mining score and the NLP score. The final maturity score was then calculated as Score = (0.6 × PM) + (0.4 × NLP) and mapped to the CMM maturity ranges presented in Table 2.

3.3 Evaluation Strategy

Two CMM assessors independently evaluated the same 23 process-period observations. Both assessors applied the predefined CMM criteria and were blinded to the maturity levels generated by the hybrid model. Because exactly two assessors participated, Cohen's kappa was used to measure agreement beyond chance. The coefficient was calculated from their initial independent assessments, before resolving disagreements, and yielded $\kappa = 0.62$. Cases in which the assessors assigned different maturity levels were subsequently reviewed in a consensus meeting. The agreed maturity level was used as the expert reference label for evaluating the hybrid model.

Table 2. Agreement between the two human assessors before consensus

Assessor 1 / Assessor	ML2	ML3	ML4	ML5	Total
ML2	2	0	0	0	2
ML3	2	6	1	0	9
ML4	0	3	7	1	10
ML5	0	0	0	2	2
Total	4	8	8	3	23

The assessors assigned the same maturity level in 17 of the 23 observations, corresponding to an observed agreement of 73.9%. Cohen's kappa was $\kappa = 0.62$, indicating substantial agreement beyond chance. The six disagreements were resolved through joint review and consensus.

Table 3. Confusion matrix of the hybrid model against the consensus expert labels

Expert reference / Model prediction	ML2	ML3	ML4	ML5	Support
ML2	2	1	0	0	3
ML3	0	7	0	0	7
ML4	0	0	8	0	8
ML5	0	0	2	3	5
Predicted total	2	8	10	3	23

The hybrid model was evaluated against the consensus reference labels established by the two human assessors. As shown in Table Y, the model correctly classified 20 of the 23 process-period observations, resulting in an accuracy of 86.96%, reported as 87%. The class-specific F1-scores were 0.80 for ML 2, 0.93 for ML 3, 0.89 for ML 4, and 0.75 for ML 5. The unweighted average of these values produced a macro-F1 score of 0.84. Cohen's kappa between the model predictions and the consensus expert labels was $\kappa = 0.81$. The complete confusion matrix and class-specific results are provided to make the calculation of the reported metrics transparent and reproducible.

4 Research and Results

4.1 Hybrid Model Concept

The hybrid model proposed in this study addresses one of the most persistent limitations of traditional CMM-based assessments: the dependence on expert judgment which, however well-intentioned, inevitably introduces subjectivity that undermines the reproducibility and comparability of results. Rather than replacing the CMM framework, the model extends it by grounding maturity evaluations in empirical, operationally derived data. As illustrated in Figure 1, the model receives three

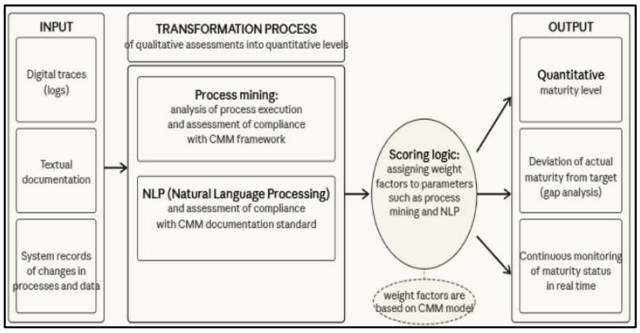


Figure 1. Hybrid model concept for objectifying IS maturity assessments using AI and CMM frameworks

categories of input — digital traces captured in process execution logs, formal textual documentation describing process specifications, and system records of operational changes — and processes them through a structured transformation pipeline that converts qualitative process characteristics into a quantitative maturity score.

Table 4. CMM Maturity Level Mapping [3]

ML Level	Name (CMMI)	Range (%)	Status Description in Your Model
ML 0	Incomplete	0% – 10%	Processes do not exist or are not monitored.
ML 1	Initial	11% – 30%	Ad-hoc work, complete dependence on individuals.
ML 2	Managed	31% – 50%	Basic project planning and tracking.
ML 3	Defined	51% – 75%	Standardized processes across the entire organization.
ML 4	Quantitatively Managed	76% – 90%	Management based on statistics and measurement.
ML 5	Optimizing	91% – 100%	Continuous independent improvement and innovation.

To illustrate this transformation, consider the emergency department triage process. Process mining yielded a CMM compliance estimate of 70%, while NLP analysis of triage protocols and role definitions produced 50%. Applying the scoring formula — Score = (0.6 × 70) + (0.4 × 50) = 62.0 — places this process at ML 3 (Defined), consistent with the organization's target maturity level. Every component of the score is traceable to concrete observable evidence, ensuring that the final maturity determination is anchored in data that can be independently verified, continuously updated, and compared across processes and time periods.

4.2 Sample and Data Sources

The practical part of the study was conducted in a medium-sized healthcare organization with 892 employees, between January and March 2026. The organization manages over forty essential business processes, all supported by its information systems. The dataset consisted of process logs from the organization's business process management (BPM) system, project documentation archives, and previous CMM assessments carried out by certified assessors over a span of two years. Six key processes were selected for analysis: emergency department triage, regular patient reception, laboratory services, radiology, patient discharge, and invoicing. Each process was characterized using a combination of quantitative metrics from system logs and qualitative insights from process documentation.

4.3 Operationalisation of Variables

To decrease the subjectivity of conventional CMM evaluations, the qualitative criteria of CMM are made operational through a set of quantitative variables covering key dimensions of process maturity, such as predictability, stability, formalization, documentation completeness and formalization and level of management control [13]. The standard deviation of process duration (σT) calculated from actual process executions logged to measure process variability. The variability is defined as:

$$\sigma_T = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(t_i - \bar{t})^2} \qquad (1)$$

where $\bar{t}$ denotes the average duration of the process. Higher values suggest that processes are ad hoc and poorly controlled and are characteristic of the lower CMM levels. The coefficient of variation (CV) of process results operationalizes stability; lower CV values indicate a stable process which is statistically controlled and at higher levels of CMM. Stability is defined as:

$$CV = \frac{\sigma_O}{\mu_O} \qquad (2)$$

where σ_O represents the standard deviation of process outcomes, and μ_O their arithmetic mean.

The level of automation (A) is measured as the share of automated activities in the total number of process activities:

$$A = \frac{N_{auto}}{N_{total}} \qquad (3)$$

where is N_{auto} the number of automated activities, and is N_{total} the total number of process activities. An increase in automation has a direct correlation with an increase in formalization, and this is the case for CMM Levels 3 and 4. Documentation completeness and formalization (D) were assessed through NLP-based identification of predefined process-documentation elements, including process steps, roles, inputs, outputs,

and control points. The indicator was calculated as the proportion of explicitly documented elements relative to the total number of required elements. Therefore, D measures documentation coverage and formalization rather than overall documentation coverage. Visual layout, general readability, and user-perceived usefulness were not evaluated in the present study. These dimensions could be incorporated in future versions of the model through readability metrics, structural-consistency indicators, and user-based document usability assessment.Documentation formalization is defined as:

$$D = \frac{1}{k}\sum_{j=1}^{k} s_j \qquad (4)$$

where s_j denotes individual semantic components (e.g., coverage of process steps, clarity of roles, definition of inputs and outputs), and is k the total number of analyzed components.

Finally, the level of control (C) is measured by how often formal control points and feedback occur in the process. Higher values mean there are more monitoring steps, which you usually see in the top CMM levels. The level of control is defined as:

$$C = \frac{N_{ctrl}}{T_{proc}} \qquad (5)$$

where N_{ctrl} denotes the number of formal control points, and T_{proc} is the observed time period. Higher values of the variable C show that there are more developed systems for monitoring and ongoing improvement, which are typical of the top CMM levels.

These variables together create a set of measurable indicators that allow for ongoing, consistent, and predictive evaluation of how mature information systems are within an AI-based CMM framework.

Equations (1)-(5) represent the authors' operationalisation of selected CMM-related maturity dimensions. They are conceptually grounded in previous maturity-model research, particularly [13], but the specific formulas and indicator combinations were developed for this study.

Table 5. Features Used in the Hybrid Maturity Assessment Model

Feature	**Data source**	**Formula / measurement**	**Interpretation**
Process duration variability	Event logs	Standard deviation of process duration	Lower variability indicates higher process stability
Coefficient of variation	Event logs	CV = σ / μ	Lower CV indicates more controlled execution
Automation ratio	System records / logs	Automated activities / total activities	Higher value indicates greater formalization
Conformance rate	Event logs	Conforming cases / total cases	Higher value indicates stronger process standardization
Documentation coverage	Process documents	Covered required steps / total required steps	Higher value indicates more complete documentation
Role clarity	Process documents	Explicitly defined roles / required roles	Higher value indicates clearer responsibility allocation
Control-point density	Logs and documentation	Number of control points / process or time period	Higher value indicates stronger management control

4.4 Weighted Factor Calculation for Key Processes

The hybrid model transforms three categories of empirical input — process execution logs, formal process documentation, and system change records — into a single quantitative maturity score through a two-channel analytical pipeline. Process mining (PM) is assigned a weight of $w_1 = 0.6$, reflecting its status as a direct, objective measurement of operational reality. NLP is assigned a complementary weight of $w_2 = 0.4$. The final weighted score is computed as:

$$Score = (w_1 \times PM\%) + (w_2 \times NLP\%) = (0.6 \times PM\%) + (0.4 \times NLP\%)$$

The differential weighting reflects a deliberate epistemological choice. Process mining draws on timestamped event logs generated automatically and independently of any evaluator, lending them greater informational value for maturity assessment than documentation, which remains a human-authored representation of intended rather than actual process behaviour.

4.5 Measurement Outcomes

The model was applied to six selected processes and revealed meaningful maturity variation. The laboratory process recorded the highest maturity profile (ML 4), driven by low turnaround time variability, high automation through laboratory information systems, and complete documentation.

Table 6. Weighted Factor Calculation for All Six Key Processes (w_1=0.6, w_2=0.4)

Process	PM% (w=0.6)	NLP% (w=0.4)	Weighted Score	Calculation	ML (AI Model)
Laboratory Services	88%	82%	**85.6**	0.6×88 + 0.4×82	**ML 4**
Invoicing	92%	85%	**89.2**	0.6×92 + 0.4×85	**ML 4–5**
Emergency Dept. Triage	70%	50%	**62.0**	0.6×70 + 0.4×50	**ML 3**
Patient Discharge	63%	55%	**59.8**	0.6×63 + 0.4×55	**ML 3**
Radiology	78%	68%	**74.0**	0.6×78 + 0.4×68	**ML 3–4**
Regular Patient Reception	65%	58%	**62.2**	0.6×65 + 0.4×58	**ML 3**

The invoicing process received a comparable score (ML 4–5) due to extensive robotic process automation (RPA) coverage and minimal variability. The emergency department triage process displayed characteristics of ML 3, reflecting standardized workflows with active control mechanisms. The patient discharge process exhibited defined but variable characteristics (ML 3). Radiology reflected stable and controlled processes with optimization potential (ML 3–4). Regular patient reception scored at ML 3, characterized by repeatable but not yet quantifiably optimized processes.

Across all six processes, the AI model consistently mapped results to CMM levels in a continuous, transparent, and empirically verifiable manner. The actual flow of hospital processes was reconstructed from operational data rather than from formal documents and descriptions, enabling a more grounded evaluation. Notably, the process mining channel revealed execution patterns — such as recurring path deviations and automation gaps — that were not captured in the formal documentation reviewed by expert assessors. These findings underscore the added diagnostic value of data-driven assessment, particularly for processes where formal documentation does not accurately reflect operational reality.

4.6 Comparative Analysis: Classical CMM Assessment vs. AI-Supported Hybrid Model

The main limitation of classical CMM assessment does not lie in the CMM framework itself, but in the way maturity levels are determined in practice. A conventional assessment relies on structured interviews, document reviews, and expert interpretation — a procedure that incorporates valuable contextualunderstanding but also introduces evaluator-dependent subjectivity. The AI-supported hybrid model addresses this limitation by grounding assessment in traceable empirical evidence. In this study, the available expert-based CMM assessments were used as reference labels. The observed Cohen's kappa of $\kappa = 0.81$ indicates substantial model-to-reference agreement and suggests that the hybrid model can reproduce expert-level maturity classifications with greater transparency and traceability.

Several patterns emerge from the comparison. First, in cases where classical assessment produced an ambiguous maturity range, such as Laboratory Services and Regular Patient Reception, the AI-supported model provides a more precise classification through continuous scoring. Second, where both approaches produce the same maturity level, as in Emergency Department Triage and Patient Discharge, the AI-supported model adds diagnostic transparency by decomposing the score into process mining and NLP components. Third, in the cases of Invoicing and Radiology, the AI-supported model identifies signs of upward maturity progression less visible in the classical assessment.

The proposed model does not eliminate subjectivity entirely. Subjective decisions remain present in the election and operationalization of indicators, the assignment of PM and NLP weights, the definition of maturity-level thresholds, data-preprocessing choices, and the use of expert assessments as reference labels. The principal contribution of the model is therefore not the complete removal of subjectivity, but its reduction, formalization, and increased transparency. In contrast to conventional assessments, where evaluator judgement may remain implicit, the assumptions embedded in the hybrid model are explicitly documented, consistently applied, and open to validation and sensitivity analysis. the model provides a more data-anchored, transparent, and reproducible assessment.

Taken together, these findings suggest that the AI-supported hybrid model should not be interpreted as a replacement for classical CMM assessment. Rather, it functions as a data-supported extension of the CMM approach that strengthens assessment by making the evidence base more explicit, the scoring logic more reproducible, and the resulting maturity diagnosis more detailed. At the same time, the model does not eliminate subjectivity entirely. Subjective elements may still enter through the selection of indicators, the definition of weights, preprocessing decisions, and the quality of expert reference labels. For this reason, the model is best understood as a mechanism for reducing and documenting subjectivity, while preserving the role of expert judgement in final interpretation and validation.

5 Discussion

The results support the main assumption underlying this research: that AI can overcome some of the epistemological and methodological weaknesses of classic CMM assessments [12].

Table 7. Classical vs AI Hybrid Model

Process	Classical CMM (Expert)	AI Hybrid Model	Key Difference / Observation
Laboratory Services	ML 3–4	**ML 4 (85.6)**	AI resolves ambiguity between ML 3 and ML 4 with a precise score, enabling unambiguous level assignment.
Invoicing	ML 4	**ML 4–5 (89.2)**	AI detects optimization potential approaching ML 5 through RPA coverage quantification, invisible to expert assessment.
Emergency Dept. Triage	ML 3	**ML 3 (62.0)**	Agreement confirmed, but AI adds traceability: the score exposes a documentation gap (NLP 50%) that the expert assessment did not formally record.
Patient Discharge	ML 3	**ML 3 (59.8)**	Agreement. AI quantifies process variability (σT) providing an improvement target absent from the expert assessment report.
Radiology	ML 3	**ML 3–4 (74.0)**	AI identifies progression toward ML 4, which the expert classified as solidly ML 3; score reveals optimization potential not captured in the static expert judgement.
Regular Patient Reception	ML 2–3	**ML 3 (62.2)**	AI upgrades from ambiguous ML 2–3 to a clear ML 3 determination; process log data confirms repeatability that the expert had assessed only tentatively.

One structural limitation of maturity models — the subjectivity of evaluation — has long been recognized [4, 5], and the findings indicate that data-driven analytical methods can significantly reduce this problem. For H1, regular expert CMM evaluations showed a Cohen's kappa coefficient of 0.62, indicating moderate inter-rater agreement. The AI-assisted model achieved a substantially higher value of 0.81 when compared against expert reference labels, confirming greater internal consistency. For H2, combining process analytics and NLP in the evaluation model enabled CMM level classification with 87% accuracy and an F1-score of 0.84, substantially exceeding the performance of expert assessment alone. For H3, the longitudinal analysis demonstrated that the AI-based model could detect maturity level changes three to four months before these changes appeared in official CMM assessments, confirming a shift from retrospective to predictive assessment.

The scientific contribution of the paper is reflected in the development of a new model that integrates the Capability Maturity Model with AI techniques — process mining and natural language processing — in order to achieve objectivity in assessments of information system maturity within healthcare organizations. Practically, the implementation of automated data analysis capabilities enables organizations to reduce evaluation costs, increase assessment frequency, and deliver more timely and accurate information to managers.

From a theoretical standpoint, the study also advances the conceptualization of CMM as a dynamic analytical framework that can receive support from continuously generated empirical data, rather than being treated as a strictly normative framework applied through periodic, expert-mediated evaluations. The introduction of a quantitative scoring layer that is fully traceable to input data and whose logic is explicitly defined creates a new class of maturity assessment artefact that can be independently verified, refined over time, and extended to additional process domains. It is important to note that the model's higher consistency does not necessarily imply that it is always more accurate in an absolute sense. Expert assessors bring contextual understanding, organizational knowledge, and professional judgement that cannot be fully captured by operational logs or documentation analysis. The hybrid model is therefore most appropriately positioned not as a substitute for expert assessment, but as a complementary instrument that provides a data-anchored perspective alongside expert interpretation. The combination of both approaches is likely to yield assessments that are simultaneously more consistent and more contextually informed than either approach could achieve in isolation.

6 Limitations and Future Work

The research was conducted within a single organization, which limits the generalizability of the results. The healthcare sector provides a rigorous testing environment for process maturity research, yet its specific characteristics may not transfer directly to other business sectors or organizational types. Future research should include a larger number of organizations across different sectors to establish the broader applicability of the proposed model. Although the proposed model reduces evaluator-dependent subjectivity, it does not eliminate subjectivity entirely. Subjective elements may enter through the selection of indicators, the definition of weights, preprocessing decisions, and the quality of

expert reference labels. To mitigate these risks, the model uses explicit feature definitions, transparent weighting, traceable input data, expert validation, and sensitivity analysis of the PM/NLP weighting scheme. The model should therefore be understood as a mechanism for reducing and documenting subjectivity, rather than as a fully objective replacement for expert judgement. Future work will expand the dataset and train a supervised classifier across multiple organizations and sectors.

A further limitation concerns the weighting scheme itself. The assigned weights of $w_1 = 0.6$ for process mining and $w_2 = 0.4$ for NLP were derived from expert consensus and validated against historical assessments, but they may not be universally optimal. Different organizational contexts or process types may warrant different weight distributions. Future research should investigate whether adaptive weighting schemes — for example, weights calibrated to the specific data quality characteristics of each organization — could further improve the consistency and accuracy of model output.
The empirical evaluation was limited to one healthcare organization and six selected business processes. Therefore, the findings should be interpreted as an initial design science evaluation rather than as evidence of universal generalizability. Broader validation across organizations, sectors, and larger datasets is required before the model can be considered generally applicable. Questions surrounding data privacy in AI-driven maturity assessments and the ethical implications of automated process monitoring also warrant further investigation.

7 Conclusion

The research began with a defined problem in information systems evaluation: that traditional maturity assessment methods, particularly those using the CMM framework, depend on expert evaluations and scheduled assessments that are difficult to standardize across different evaluators. The proposed model addresses this problem by using AI to analyze actual operational data. The model applies process analytics, NLP, and machine learning to transform part of the maturity assessment process from subjective judgment into concrete, traceable evidence. The empirical findings show that the AI-supported model compares favorably to traditional expert-based CMM assessments, producing more consistent results (Cohen's kappa 0.81 vs. 0.62) and better classification accuracy (87%, F1 = 0.84). The model enables organizations to track maturity progress through ongoing assessments capable of forecasting changes three to four months before scheduled evaluation dates. This enables organizations to use CMM as an active, data-driven assessment instrument rather than a fixed periodic diagnostic tool. The overarching contribution of this study lies not only in the technical development of the hybrid model, but in the reconceptualization of how maturity assessments are produced and used. By embedding the assessment process within the organization's information infrastructure — drawing continuously on event logs, documentation repositories, and system records — the model transforms maturity evaluation from an episodic intervention into an ongoing monitoring function. This shift has practical significance: organizations can receive early signals of deteriorating process performance, identify maturity bottlenecks at a finer granularity than traditional assessments allow, and allocate improvement resources based on quantitatively grounded evidence rather than evaluator impressions.

Table 8. Classification Results of the AI-Supported Model

Metric	Value	Explanation
Accuracy	87%	Share of process observations where predicted ML matched the expert reference
Macro-F1	0.84	Average F1 across maturity-level classes
Cohen's κ	0.81	Agreement beyond chance between AI model and expert reference

8 Data, Code, and Supplementary Material Availability

Due to organizational confidentiality and data-protection requirements, the raw healthcare event logs and organization-specific integration scripts cannot be made publicly available. The original data and technical materials are stored in the hospital's secure institutional environment. To support methodological transparency and partial reproducibility, a supplementary package has been prepared containing the model pseudocode, complete feature definitions, normalization procedures, process-mining and NLP aggregation rules, weighting formulas, CMM classification thresholds, anonymized input examples, and step-by-step calculation examples. A README file describes the structure of the supplementary materials and the procedure for reproducing the scoring process and maturity-level mapping reported in this study. The non-confidential parts of the supplementary package may be made available to qualified researchers upon reasonable request to the corresponding author, subject to institutional approval and applicable data-protection requirements. These materials enable independent verification and reproduction of the scoring logic and aggregate calculations; however, complete end-to-end replication using the original hospital event logs is

restricted by confidentiality and data-protection requirements.

References

de Brun, T., Freeze, R., Kaulkarni, U., & Rosemann, M. (2005). Understanding the main phases of developing a maturity assessment model. *Australasian Conference on Information Systems Proceedings*, 8, 1–10.

Röglinger, M., Pöppelbuß, J., & Becker, J. (2012). Maturity models in business process management. Business *Process Management Journal*, 18(2), 328–346. https://doi.org/10.1108/14637151211225225

Kenett, R. S., & Baker, E. R. (2010). *Improvement and CMMI for Systems and Software*. CRC Press.

Mettler, T. (2011). Maturity assessment models: A design science research approach. *International Journal of Society Systems Science*, 3(1–2), 81–98. https://doi.org/10.1504/IJSSS.2011.038934

Becker, J., Knackstedt, R., & Pöppelbuß, J. (2009). Developing maturity models for IT management. *Business & Information Systems Engineering*, 1(3), 213–222. https://doi.org/10.1007/s12599-009-0044-5

Pöppelbuß, J., & Röglinger, M. (2011). What makes a useful maturity model? A framework of general design principles for maturity models. European *Conference on Information Systems Proceedings*.

Van der Aalst, W. (2016). *Process Mining: Data Science in Action* (2nd ed.). Springer. https://doi.org/10.1007/978-3-662-49851-4

Russell, S. J., & Norvig, P. (2021). *Artificial Intelligence: A Modern Approach* (4th ed.). Pearson.

Shrestha, Y. R., Ben-Menahem, S. M., & von Krogh, G. (2019). Organizational decision-making structures in the age of artificial intelligence. *California Management Review*, 61(4), 66–83.

Tarafdar, M., Beath, C. M., & Ross, J. W. (2019). Using AI to enhance business operations. MIT *Sloan Management Review*, 60(4), 37–44.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, 28(1), 75–105.

Helfat, C. E., & Raubitschek, R. S. (2018). Dynamic and integrative capabilities for profiting from innovation in digital platform-based ecosystems. *Research Policy*, 47(8), 1391–1399.

Van Looy, A., De Backer, M., & Poels, G. (2014). Defining business process maturity. *Total Quality Management & Business Excellence*, 25(11–12), 1239–1256.

The Adoption of Vibe Coding and the Transformation of Programming Practices Induced by Generative Models: A SonarQube-Based Analysis

Maja Benkus Kumek, Neven Vrček
University of Zagreb Faculty of Organization and Informatics
Pavlinska ul. 2, 42000 Varaždin, Croatia
{mbenkus, neven.vrcek}@foi.unizg.hr

Abstract. *This paper studies how vibe coding – specifying behaviour in natural language for generative models to implement – affects code quality. It combines a review of mostly qualitative research with a preliminary, exploratory case study using SonarQube static analysis. Three implementations of the same feature are compared: one manual and two vibe-coded with GPT-5.2-Codex and Gemini 3.1 Pro. The aim is to demonstrate a metric-based comparison and generate hypotheses, not to confirm them. Both vibe-coded versions showed fewer code smells, lower complexity, and less technical debt than the manual baseline, while a runtime defect missed by static metrics reveals the approach's limits.*

Keywords. vibe coding, programming, software engineering, artificial intelligence, large language models, generative models, technical debt, SonarQube

1 Introduction

Large language models (LLMs) are a type of deep learning model trained on large amounts of text data, which allows them to understand and generate natural language across a wide range of tasks. Their versatility has made them one of the most impactful technologies in recent years. In addition to common uses such as artificial intelligence (AI) assistants and text processing, LLMs are increasingly being used to generate and review source code, as well as to detect bugs (Hou et al., 2024; Jiang et al., 2025).

As generative AI becomes more widespread in software development, a new paradigm has emerged known as *vibe coding*. In vibe coding, the developer describes the desired behaviour of a software product in natural language, and a generative model produces or refines the corresponding code (Sapkota et al., 2025; Sarkar and Drosos, 2025). While this approach can make software development faster and more accessible, it also raises concerns about code quality, readability, and long-term maintainability. These concerns motivate a closer scientific examination of how generative models influence programming practices and the quality of the code they produce.

This paper investigates the impact of vibe coding on programming practices and code quality. It combines a review of existing literature on generative models in software development with an empirical analysis of code produced using the vibe coding approach. The primary objective is to compare code quality between vibe coding and conventional development, using static code quality metrics from SonarQube as the basis for comparison. A secondary objective is to assess how well the empirical findings align with the claims made in existing research.

In accordance with the stated research objectives, the following research questions are addressed:

- **RQ1:** What research methodologies and changes in developer work practices are identified in the existing scientific literature on the application of generative models in programming?
- **RQ2:** How does the existing literature address the impact of generative models and vibe coding on code quality and technical debt?
- **RQ3:** Are there measurable differences in static code quality metrics between manually developed and vibe coding implementations, and what advantages and limitations of the vibe coding approach can be identified through SonarQube analysis?

The empirical component reported here is deliberately framed as a preliminary, exploratory case study rather than a controlled experiment; its contribution lies in establishing and demonstrating a measurable, metric-based comparison approach that future work can scale to multiple developers, tasks, and projects.

2 Related Work

The literature review covers three aspects of generative language model use in programming: the experience of developers working with these tools, changes in the software development process, and the potential effects

on code quality and long-term maintainability. Particular attention is given to studies that view vibe coding not just as a tool for boosting productivity, but as a broader shift in how responsibility is divided between human developers and AI-based systems.

2.1 Developer Role, Comprehension, and Maintainability

Research on generative tools in programming extends beyond the speed and quality of produced code to the developer's changing role and, increasingly, to the downstream consequences for code comprehension and maintenance. A body of work grounded in developer-experience surveys and self-reported assessments (Brandebusemeyer, 2025; Sampson, 2025) examines the subjective side of working with generative AI – cognitive load, control, and the pressure to adopt new tools. While these well-being concerns lie outside the scope of the present study, the same studies consistently observe that working with generated code introduces additional verification and correction effort. Weisz et al. (2025) report this pattern directly in an industrial deployment of an AI coding assistant: perceived productivity rises, but the gains are uneven across developers and are partly offset by the extra steps required to check and fix generated output, so an initial impression of higher productivity may not be sustained.

Sankaranarayanan (2026) conducted a controlled experiment with 78 participants to examine whether unrestricted AI use leads to a decline in developer competence. The study compared three groups: developers working manually, those using AI without restrictions, and those using AI with a structured scaffolding mechanism. While both AI groups performed similarly on the initial coding task, the unrestricted AI group had a 77% failure rate on a subsequent maintenance task completed without AI assistance, compared to 39% in the scaffolded group. The author attributes this gap to the accumulation of *epistemic debt* – a situation where developers own code functionally but do not understand it well enough to maintain it independently. This finding is relevant to the present study because it provides behavioural evidence for the same phenomenon that SonarQube-based static analysis aims to capture at the codebase level: that AI-generated code may appear functional while concealing underlying quality and maintainability problems.

Maes (2026) provides a broad empirical synthesis of AI-assisted software development practices in early 2026, drawing on data from developer surveys, code complexity analyses, and production environment observations. The study introduces the concept of the *Great Toil Shift*, describing how the time saved during AI-assisted code generation tends to be entirely consumed by the increased effort required for code review, security auditing, and maintenance. Quantitative findings indicate that AI-assisted pull requests contained 1.7 times more structural and logical issues than those authored by humans, while cognitive complexity in repositories with heavy AI use increased by 39%. These results directly support the motivation for the empirical component of the present study and offer an external reference point for interpreting the SonarQube metrics reported here.

2.2 Vibe Coding as a New Development Practice

Several studies have sought to characterise vibe coding as an emerging development practice and to examine methodologies for its investigation. Chou et al. (2025) apply Straussian Grounded Theory analysis to twenty YouTube videos dedicated to vibe coding, identifying recurring behavioural patterns, prompt formulation strategies, and error resolution approaches in AI-mediated software development. Their contribution lies primarily in conceptualising vibe coding as a new development paradigm and in describing how generative models affect the cognitive and Organisational dimensions of programming.

A methodologically more elaborate approach is adopted by Sarkar and Drosos (2025), who reconstruct complete development workflows from multi-hour recordings. Their results demonstrate that the use of generative models does not eliminate the need for programming expertise; rather, it transforms the way that expertise is exercised. Emphasis shifts increasingly from writing code toward evaluating the quality of generated solutions, managing task context, and deciding when to rely on AI and when to intervene manually. These findings support the understanding of vibe coding as a qualitative paradigm shift rather than mere automation of programming – and it is precisely this perspective that motivates the research questions of the present study.

Sapkota et al. (2025) examine the distinction between vibe coding and agentic coding, arguing that both approaches alter the structure of developer work by displacing focus from implementation details toward supervision, evaluation, and high-level decision-making. While both can reduce routine cognitive effort and accelerate development, the authors warn that the increased need for continuous evaluation and management of unpredictable model behaviour may simultaneously lead to cognitive fatigue.

Horvat (2026) investigates how the structure of prompting affects the quality of code generated through vibe coding, comparing zero-shot prompting with prompt chaining in a web front-end development task using GPT-4o and Gemini 2.5 Pro. The results show that prompt chaining – where the task is broken into sequential steps, each building on the previous output – consistently produces code with better architectural consistency, cleaner separation of concerns, and fewer

layout errors compared to zero-shot prompting, which submits all requirements in a single prompt. Zero-shot prompting was faster but resulted in lower maintainability and less adherence to design specifications. This study is relevant to the present work as it demonstrates that the quality of AI-generated code is not only a property of the model used, but also of how the developer interacts with it – a dimension that static analysis tools such as SonarQube can help quantify across different development workflows.

2.3 Technical Debt and Code Quality

One of the first detailed studies directly addressing the relationship between vibe coding and technical debt is the work of Waseem et al. (2025), who analyse how the use of generative language models affects long-term software sustainability. The study employs an experiential and qualitative methodology based on the authors' own development projects and observations from industrial practice. A key finding is that rapid iterations and the sensation of fast progress in early development phases tend to conceal the accumulation of architectural, security, testing, and maintenance problems – a dynamic the authors describe as the tension between development flow and technical debt. The authors propose guidelines for more sustainable vibe coding practice; however, the absence of standardised quantitative code quality metrics leaves room for empirical investigations such as the present study.

Van Hurne (2025) approaches technical debt in the context of generative AI more systematically, proposing a framework aimed at preventing its accumulation in early development phases. The work connects existing theoretical frameworks of technical debt with the ways in which generative models are typically employed in software development, and emphasises the need for clear development guidelines, early identification of quality requirements, and continuous maintainability assessment. While the study does not include empirical code quality measurement via tools such as SonarQube, it explicitly highlights the need for measurable technical debt indicators – a gap that the present study directly addresses through comparative static analysis.

Casserini et al. (2026) introduce the concept of *agentic entropy* to describe the process by which autonomous coding agents gradually drift away from the intended architectural design of a software system. Because agents typically operate within a limited context window, they tend to optimise for local correctness without maintaining awareness of the broader system structure, which leads to the accumulation of what the authors call *agentic technical debt* – structural misalignments, duplicated logic, and fragile refactoring that compound over time. The authors propose a process-oriented explainability framework built around architectural seeding, reasoning monitoring, and a causal reasoning graph to make agent behaviour more transparent to human reviewers. While the proposed framework addresses the problem at the level of agent oversight, the present study approaches the same concern through its measurable codebase consequences, using SonarQube to detect the structural degradation that agentic entropy produces.

Gandhi (2026) presents a systematic survey of security and quality risks associated with vibe coding and autonomous development workflows, synthesising findings from empirical studies and vulnerability reports published between 2024 and 2025. Among the key quantitative findings cited in the survey is a CodeRabbit analysis of 470 GitHub pull requests, which found that AI-generated code contains 1.7 times more defects than human-authored code overall, with security issues up to 2.74 times higher and readability issues more than three times higher. The survey also highlights the *comprehension gap* as a central mechanism behind these quality problems, arguing that developers who rely on AI without reviewing the generated code lose the ability to identify structural flaws. These findings are consistent with the theoretical motivation of the present study and provide a broader empirical context for the SonarQube-based results reported here.

Taken together, the reviewed works reveal a notable absence of studies with a strong empirical component focused on measuring technical debt and code quality through quantitative metrics. Although numerous authors point to the potential accumulation of technical debt and changes in development practices associated with vibe coding, few studies have tested these claims through measurable indicators. The empirical component of the present study directly addresses this gap.

3 Methodology

The research is based on a comparison of source code developed through three distinct implementation approaches for the same functionality within the same application. To ensure comparability of results, all three scenarios share the same code repository, the same functional requirements, the same development environment, and identical evaluation criteria.

The *BlazorWebAppMovies* repository, part of Microsoft's official Blazor samples on GitHub, serves as a reference application for learning and demonstrating web application development with Blazor and Entity Framework (EF) Core ("BlazorWebAppMovies Sample Application", 2024). Since the objective of the research was to examine the impact of vibe coding, the only differentiating factor across implementations was the development approach – manual or GitHub Copilot with two distinct generative models. GitHub Copilot provides developer support through automatic code suggestion and generation based on context and source code comments. It is founded on generative language models and is designed to increase developer produc-

tivity, while responsibility for the correctness and quality of generated code remains with the user ("GitHub Copilot Documentation", 2024). For static code quality analysis, SonarQube Community Edition was employed, enabling static source code analysis and calculation of metrics such as bugs, vulnerabilities, code smells, and technical debt ("SonarQube Community Edition", 2024).

3.1 Implementation Scenarios

The study encompasses three implementation scenarios realised through separate Git branches: `implementation/without_vibe_coding`, `implementation/vibe_coding_gpt52codex`, and `implementation/vibe_coding_gemini31pro`. Three separate SonarQube projects were created due to the limitations of the Community Edition, which does not support cross-branch comparison within a single project. Table 1 provides an overview of the three scenarios.

Table 1. Implementation scenario overview

Scenario	Approach	Model Used
Scenario 1	Manual development	—
Scenario 2	Vibe Coding	GPT-5.2-Codex
Scenario 3	Vibe Coding	Gemini 3.1 Pro

In all three scenarios, the same set of functional requirements was implemented, focused on extending the existing system by associating each film entity with an actor entity – specifically, implementing a many-to-many (M:N) relationship between films and actors with appropriate user interface (UI) support. The requirements included:

- adding a new `Actor` class,
- explicit implementation of the M:N relationship via a `MovieActor` join entity,
- seeding initial data such that every film has associated actors,
- displaying a list of all actors,
- displaying each actor's filmography,
- enabling actor assignment on the film editing component,
- displaying the actor list on the film details component.

In all implementations, the explicit join class for the M:N relationship was deliberately used to avoid differences arising from Entity Framework's implicit mechanisms.

3.1.1 Manual Development (Scenario 1)

In the first scenario, the entire functionality was implemented manually without working with generative language models. The author possesses foundational knowledge of the Blazor framework and has limited prior experience with this framework. The development workflow involved manual creation of the `Actor` and `MovieActor` classes, manual addition of database migrations, seeding logic, creation of new Blazor components for actor display and assignment, and adaptation of existing components.

3.1.2 Vibe Coding (Scenarios 2 and 3)

The remaining two scenarios employed the vibe coding approach using GitHub Copilot in Agent mode. Agent mode enables the autonomous execution of complex development tasks, allowing the agent to analyse existing code, propose and implement changes, and iteratively resolve issues with minimal developer intervention. Unlike standard Copilot usage, Agent mode extends the tool's role from a coding assistant to a partially autonomous participant in the development process ("About GitHub Copilot Coding Agent", 2024).

A notable limitation of this study concerns the model selection within GitHub Copilot. As the research was conducted using an educational GitHub Copilot licence, the full range of available models was not accessible; therefore, the models chosen for this study represent the best options available within that constrained environment. GPT-5.2-Codex was selected as OpenAI's most capable coding-focused model available, purpose-built for long-horizon agentic tasks with support for high-effort reasoning and a 400,000-token context window ("GPT-5.2-Codex Model", 2025). Gemini 3.1 Pro was selected as a representative model from Google – a different major AI vendor – to enable cross-vendor comparison, as it delivers substantially improved core reasoning over its predecessor, achieving more than double the reasoning performance of Gemini 3 Pro on complex problem-solving benchmarks ("Gemini 3.1 Pro: A smarter model for your most complex tasks", 2026). In both vibe coding scenarios, identical prompts were provided to GitHub Copilot, instructing it to extend the solution with actor functionality while keeping changes minimal and code clean, adhering to EF Core conventions, and maintaining UI consistency with the existing Blazor application. The role of the author was limited to confirming proposed commands and changes until a successful project build was achieved.

3.2 Evaluation and Code Analysis

Evaluation was conducted using SonarQube Community Edition with the aim of objectively comparing code quality across the three development approaches. The analysis relies exclusively on static code quality

metrics, without additional runtime testing – an acknowledged limitation of the study. The only inclusion criterion was a successful project build, meaning the analysed code had to constitute a deliverable software artefact, which eliminated the need for subjective manual testing from the evaluation criteria.

For each of the three implementations, the following metrics were collected: number of bugs, number of vulnerabilities, number of code smells, cognitive complexity, cyclomatic complexity, number of duplicated blocks and lines, duplicated lines density, number of classes, functions, and files, total lines of code and non-commented lines of source code (NCLOC) and technical debt expressed as the SQALE (Software Quality Assessment based on Lifecycle Expectations) index. Metrics with a value of zero across all implementations (bugs, coverage, line coverage, vulnerabilities) were excluded from the comparative analysis. Due to SonarQube Community Edition's limitations, metric data for each project was retrieved via the SonarQube API and consolidated into a CSV file using a Python script.

4 Results

4.1 Implementation Results by Scenario

In **Scenario 1**, functionality was implemented entirely manually without generative models. Development required a considerable number of iterations, largely due to the author's limited prior experience with the Blazor framework. Multiple errors came up, requiring consultation of documentation and exploration of alternative solutions. The overall development process in this scenario was the longest among all three approaches.

In **Scenario 2**, implementation was carried out using the vibe coding approach with the GPT-5.2-Codex model in GitHub Copilot Agent mode. The agent began by planning out 37 tasks before executing them one by one. Some tasks failed and were retried, and the process involved frequent developer confirmations, making it surprisingly lengthy for a project of this scope. The generated output was nonetheless detailed and required no manual code modifications after the agent finished. A UI issue was observed in the film editing component during manual application launching; however, as runtime testing was outside the scope of the evaluation criteria, this was not further investigated.

In **Scenario 3**, the vibe coding approach was employed with the Gemini 3.1 Pro model. Compared to Scenario 2, the code generation process was faster and required fewer interruptions from the developer, with the agent planning only 5 tasks. The model independently organised the implementation steps, including file creation, modifications to existing code, and addition of accompanying documentation. The project built successfully; however, the application could not be launched immediately, as the code responsible for applying migrations at runtime was missing. The author used GitHub Copilot's debugger mode to identify the issue, which was then resolved with a fix suggested by the agent.

4.2 Static Code Analysis Results

The results of static code analysis obtained with SonarQube Community Edition are presented in Table 2. Metrics with a value of zero across all three implementations were excluded from the analysis, as described in Section 3.2.

Table 2. Static code analysis results (SonarQube)

Metric	S1	S2	S3
Classes	12	12	11
Code smells	29	15	15
Cognitive complexity	70	64	57
Cyclomatic complexity	219	206	199
Duplicated blocks	2	0	0
Duplicated lines	39	0	0
Dup. lines density (%)	2.2	0	0
Files	34	33	33
Functions	129	122	116
Lines	1751	1719	1681
NCLOC	1140	1107	1084
SQALE index (tech. debt)	172	140	136

All three scenarios produced a comparable amount of code in terms of files, lines, and functions. Scenario 1 recorded the highest values for code smells (29), cognitive complexity (70), cyclomatic complexity (219), and technical debt index (172), as well as the only instances of code duplication, with 2 duplicated blocks across 39 lines at a density of 2.2%. Scenarios 2 and 3 produced identical code smell counts (15) and no duplication, with Scenario 3 showing slightly lower complexity (57 cognitive, 199 cyclomatic) and a marginally lower technical debt index (136) compared to Scenario 2 (64, 206, and 140 respectively).

To complement the static metrics with an indication of development effort, Table 3 reports the approximate development time for each scenario alongside three interaction indicators: the number of tasks the agent planned, the manual code edits required after generation, and the post-build runtime issues encountered. Development time is measured as wall-clock time from the start of the task to the first successful build, including research and debugging. These figures are coarse, single-run observations and are reported as descriptive context rather than as precise productivity measurements.

Table 3. Development effort and interaction indicators per scenario

Indicator	S1	S2	S3
Development time in minutes (approx.)	120-180	30-40	15-20
Agent-planned tasks	—	37	5
Manual code edits after generation	—	0	1
Post-build runtime issues	—	1	1

5 Discussion

The objective of this study was to examine how the application of generative language models and the vibe coding approach affects software code quality in comparison with traditional software development. In addressing RQ1 and RQ2, the literature review demonstrates that existing research largely employs qualitative methodologies and frames vibe coding as a paradigm shift in developer work rather than merely a productivity tool. The empirical results support this perspective: in both vibe coding scenarios, the developer's role changed substantially from that in manual development, with focus shifting from writing code to defining requirements, supervising, and evaluating generated artefacts. Scenarios 2 and 3 are thus consistent with the claim found in the literature that developer work shifts from code authorship toward high-level decision-making, although a single observation per scenario cannot establish this as a general pattern.

Addressing RQ2, the literature warns of the risk of technical debt accumulation caused by accelerated development, insufficiently defined architecture, and vague functional requirements, but prior studies have not used quantitative metrics (van Hurne, 2025; Waseem et al., 2025). The empirical component of this study helps fill that gap through SonarQube analysis. The results show that both vibe coding scenarios produced fewer code smells, lower complexity, and lower technical debt than the manual implementation, with no code duplication present in either. This does not align with the literature's concern about accelerated development leading to increased technical debt – at least not at the level of static metrics on a project of this scope.

Addressing RQ3 directly, for this task and at the level of static metrics, the vibe coding implementations did not exhibit inferior code quality: both Scenarios 2 and 3 scored better than the manual baseline on most measured metrics. This result is specific to the studied task and cannot be generalised to vibe coding as a whole. Scenario 3 (Gemini 3.1 Pro) recorded the lowest values for cognitive complexity, cyclomatic complexity, and technical debt index of all three scenarios, despite requiring a runtime fix after the initial build. This suggests that the quality of the generated code is not strictly tied to the smoothness of the generation process. However, the runtime issue encountered in Scenario 3 – where migration application code was missing – illustrates that static metrics alone do not capture the full picture of software quality, which is an acknowledged limitation of this study. This aligns with the literature's assertion that generative models do not eliminate the need for programming knowledge, but alter the form in which it is applied — particularly in the phases of evaluation and quality control (Sapkota et al., 2025; Sarkar and Drosos, 2025).

The two vibe coding scenarios also differed in agent behaviour. GPT-5.2-Codex decomposed the task into 37 planned steps and produced a buildable result requiring no manual edits, but left a UI defect; Gemini 3.1 Pro planned only 5 steps, completed faster with fewer interruptions, and achieved marginally better static metrics, yet shipped code that failed at runtime until a migration fix was applied. This contrast suggests a possible trade-off between planning thoroughness and generation speed across vendors, but with a single run per model it is an observation rather than a benchmark, and no claim about the relative capability of the two models is warranted on this evidence.

5.1 Threats to Validity

Before interpreting the results, it is worth being clear about what a study of this design can and cannot establish, since the empirical component is a preliminary, exploratory case study rather than a controlled experiment. The most serious threat concerns internal validity. All three implementations were produced by a single developer with limited prior Blazor experience, so the differences between the manual and vibe coding scenarios cannot be cleanly separated from that developer's learning curve; because the manual scenario was also the first pass through the task, its higher metric values may reflect unfamiliarity as much as the development approach itself. A larger study would need several developers of varying expertise to isolate the effect of the approach from the effect of the person.

External validity is limited in a comparable way. The study rests on a single reference repository (*BlazorWebAppMovies*), a single technology stack (Blazor and Entity Framework Core), and a single, relatively self-contained feature – an M:N extension between films and actors – so the results say little about other languages, frameworks, paradigms, or larger and more architecturally complex systems. The choice of models was further constrained by an educational GitHub Copilot licence, which restricted the comparison to the two available models rather than the most capable models on the market. We worked in this controlled, deliberately simple setting because it lets the three scenarios share identical requirements and a clean baseline; the cost of that control is that the findings do not yet gen-

eralise beyond it.

Construct validity is limited by the decision to operationalise code quality entirely through SonarQube static metrics. These metrics do not capture functional correctness, runtime behaviour, test adequacy, security, or user-facing quality, and the missing-migration defect in Scenario 3 – clean static numbers alongside an application that would not run – makes that gap concrete. With only a single observation per scenario no statistical inference is possible either, so every difference reported here is descriptive rather than confirmatory. Taken together, these considerations are the reason the study is presented as a first, hypothesis-generating step rather than as settled evidence about vibe coding, even as it offers a new empirical perspective on a domain that has so far been investigated largely through qualitative approaches.

6 Conclusion

This paper examined the impact of vibe coding on source code quality through a combination of literature review and static code analysis using SonarQube. The literature review showed that vibe coding is seen in existing research as a shift in how software is developed, with particular emphasis on changes in the developer's role, working practices, and responsibility for code quality. A clear lack of quantitative empirical research was identified as a gap in the current body of knowledge.

The empirical component shows that, for the task studied, the vibe coding implementations did not produce worse code quality in terms of basic static metrics: in both vibe coding scenarios, fewer code smells, lower complexity, and lower technical debt were recorded than in the manual implementation. At the same time, the runtime issue encountered in Scenario 3 shows that good static metrics do not guarantee a fully working solution, and that the developer must remain actively involved in reviewing and validating the generated output.

Based on these results, vibe coding cannot be simply labelled as better or worse than traditional software development. It represents a different way of working that requires new skills and new approaches to quality control. These conclusions are drawn from a single task on a single project and are offered as preliminary, hypothesis-generating evidence rather than as a general finding about vibe coding. The value of this paper lies in providing empirical data where most existing research has relied on qualitative observation, and in connecting theoretical insights with measurable code quality indicators. Among the quality dimensions not captured by static analysis, several stand out as priorities for future work. Functional correctness and runtime behaviour are the most significant: the missing-migration defect in Scenario 3 demonstrates that favourable static metrics can coexist with a non-functioning application, so verifying that the software actually builds, runs, and behaves as specified is a precondition for any quality claim. Testing quality is a closely related priority, since the present study collected no test-coverage or test-adequacy data, and the degree to which generated code is accompanied by meaningful tests is itself a maintainability concern. Future research should therefore incorporate dynamic evaluation – runtime, integration, and unit testing – together with test-coverage, and extend the analysis to larger and more complex projects and a wider range of generative models.

References

About GitHub copilot coding agent. (2024). https://docs.github.com/en/copilot/concepts/agents/coding-agent/about-coding-agent (Accessed: 28.04.2026).

BlazorWebAppMovies sample application. (2024). https://github.com/dotnet/blazor-samples/tree/main/10.0/BlazorWebAppMovies (Accessed: 28.04.2026).

Brandebusemeyer, C. (2025). Interactions with generative AI: Wearables to measure developer experience and productivity objectively. *Proceedings of the 47th IEEE/ACM International Conference on Software Engineering: Companion Proceedings (ICSE-Companion)*, 148–150. https://doi.org/10.1109/ICSE-Companion66252.2025.00043

Casserini, M., Facchini, A., & Ferrario, A. (2026). Beyond the 'diff': Addressing agentic entropy in agentic software development. https://arxiv.org/abs/2604.16323

Chou, Y.-H., Jiang, B., Chen, Y. W., Weng, M., Jackson, V., Zimmermann, T., & Jones, J. A. (2025). Building software by rolling the dice: A qualitative study of vibe coding. https://arxiv.org/abs/2512.22418

Gandhi, D. (2026). Survey on systemic security risks in vibe coding and autonomous development workflows. *TechRxiv*, *2026*(0110). https://doi.org/10.36227/techrxiv.176800890.09196406/v1

Gemini 3.1 Pro: A smarter model for your most complex tasks. (2026). https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-pro/ (Accessed: 28.04.2026).

GitHub copilot documentation. (2024). https://docs.github.com/en/copilot/get-started/what-is-github-copilot (Accessed: 28.04.2026).

GPT-5.2-Codex model. (2025). https://developers.openai.com/api/docs/models/gpt-5.2-codex (Accessed: 28.04.2026).

Horvat, M. (2026). Which prompting technique is better? Improving vibe coding code quality with efficient prompt design for web front-End development [University of Zagreb, Faculty of Electrical Engineering and Computing]. https://doi.org/10.5281/zenodo.18165128

Hou, X., Zhao, Y., Liu, Y., Yang, Z., Wang, K., Li, L., Luo, X., Lo, D., Grundy, J., & Wang, H. (2024). Large language models for software engineering: A systematic literature review. *ACM Transactions on Software Engineering and Methodology*, *33*(8), 1–79. https://doi.org/10.1145/3695988

Jiang, J., Wang, F., Shen, J., Kim, S., & Kim, S. (2025). A survey on large language models for code generation. *ACM Transactions on Software Engineering and Methodology*. https://doi.org/10.1145/3747588

Maes, S. H. (2026, March). *Evaluating the efficacy of artificial intelligence in software engineering: A post-February 2026 analysis* (tech. rep.). https://doi.org/10.5281/zenodo.19103818

Sampson, M. (2025). Enhancing developer experience: Burnout, resilience, and the digital environment in software engineering. *Proceedings of the 38th International BCS Human-Computer Interaction Conference*, 369–374. https://doi.org/10.14236/ewic/BCSHCI2025.41

Sankaranarayanan, S. (2026). Mitigating "epistemic debt" in generative ai-scaffolded novice programming using metacognitive scripts. https://arxiv.org/abs/2602.20206

Sapkota, R., Roumeliotis, K. I., & Karkee, M. (2025). Vibe coding vs. agentic coding: Fundamentals and practical implications of agentic AI. https://arxiv.org/abs/2505.19443

Sarkar, A., & Drosos, I. (2025). Vibe coding: Programming through conversation with artificial intelligence. https://arxiv.org/abs/2506.23253

SonarQube community edition. (2024). https://docs.sonarsource.com/sonarqube-cloud/ (Accessed: 28.04.2026).

van Hurne, M. H. (2025). *A technical debt-aware prompting framework for sustainable vibe coding: Addressing the production readiness crisis in AI-assisted software development* (tech. rep.). TechRxiv. https://doi.org/10.36227/techrxiv.175459417.76916566/v1

Waseem, M., Ahmad, A., Kemell, K.-K., Rasku, J., Lahti, S., Mäkelä, K., & Abrahamsson, P. (2025). Vibe coding in practice: Flow, technical debt, and guidelines for sustainable use. https://arxiv.org/abs/2512.11922

Weisz, J. D., Kumar, S., Muller, M., Browne, K.-E., Goldberg, A., Heintze, E., & Bajpai, S. (2025). Examining the use and impact of an ai code assistant on developer productivity and experience in the enterprise. https://arxiv.org/abs/2412.06603

UIKit-to-SwiftUI Migration in an iOS E-Voting Application: A Comparative Case Study of Presentation-Layer Structure

Sandra Melovic, Dusanka Dakic, Darko Stefanovic, Teodora Vuckovic
University of Novi Sad
Faculty of Technical Sciences
Trg Dositeja Obradovića 6, 21102 Novi Sad, Serbia
{sandra.melovic, dusanka.dakic, darko.stefanovic, teodora.vuckovic}@uns.ac.rs

Abstract. *SwiftUI is increasingly adopted as a declarative alternative to UIKit, but evidence from existing applications remains limited. This comparative case study examines a feature-wise UIKit-to-SwiftUI migration in an iOS e-voting application while retaining the backend, data model, business rules, and workflow scope. The SwiftUI implementation contained 71.3% fewer UI-layer lines of code and 64.2% fewer presentation-layer lines. Function-level analysis showed the same median cyclomatic complexity, a slightly higher mean, and a lower maximum in SwiftUI. The findings demonstrate structural differences in this case but do not establish universal improvements in maintainability, productivity, performance, or usability.*

Keywords. iOS migration, Model-View-ViewModel, software modernization, SwiftUI, UIKit

1 Introduction

Modernizing an established mobile application involves more than replacing one user-interface framework with another. In UIKit applications, interface rendering, navigation, lifecycle handling, validation, and state coordination are frequently implemented through UIViewController-based structures. As applications evolve, these responsibilities may become concentrated in controller classes, increasing coupling and making presentation behavior more difficult to isolate. SwiftUI introduces a declarative and state-driven model in which the rendered interface is derived from application state rather than updated through explicit sequences of interface operations.

Although SwiftUI has been increasingly adopted in iOS development, existing evidence does not yet support universal claims that migration automatically improves software quality. Previous studies include small implementation comparisons, broad comparisons of mobile frameworks, analyses of framework-adoption patterns, and research on declarative user-interface generation. These studies indicate potential reductions in boilerplate and changes in architectural organization, but they do not fully explain what happens when an existing UIKit application is reimplemented while its backend and functional scope are held constant.

A further methodological problem is that framework migration is often accompanied by architectural refactoring. A UIKit-to-SwiftUI migration may simultaneously introduce Model-View-ViewModel separation, feature-based organization, reusable components, and revised navigation. Consequently, a smaller or more modular implementation cannot automatically be attributed to SwiftUI alone. The effects of the declarative framework must be distinguished, as far as the available evidence permits, from the effects of restructuring the presentation architecture.

This study examines an iOS e-voting case application implemented in UIKit and subsequently reimplemented feature by feature in SwiftUI. The backend contract, domain data, business rules, and targeted user workflows were retained. The study focuses on the presentation layer and does not evaluate backend redesign, blockchain mechanisms, runtime performance, user experience, or long-term development outcomes.

The research addresses the following questions:

RQ1: What differences in source-code size and file organization are observed between the UIKit and SwiftUI presentation-layer implementations of the same case application?

RQ2: How are interface rendering, state handling, navigation, and reusable presentation components redistributed during the migration?

RQ3: Which observed changes can plausibly be associated with SwiftUI's declarative mechanisms, which are associated with the accompanying architectural refactoring, and what limitations constrain the interpretation of the results?

The findings are intended to support architectural decision-making for teams considering feature-wise user-interface modernization. They should be interpreted as evidence from one case rather than as proof that SwiftUI is universally smaller, more maintainable, or more productive than UIKit.

The remainder of the paper is organized as follows. Section 2 presents the conceptual and empirical background. Section 3 describes the case study, migration procedure, metric definitions, and functional-equivalence protocol. Section 4 reports the structural results. Section 5 discusses their interpretation, practical applicability, and threats to validity. Section 6 concludes the paper.

2 Theoretical Concepts

2.1 Imperative and Declarative UI Paradigms

UIKit takes an imperative approach to building interfaces. It provides for building everything step by step - screens, view hierarchies, constraints - and wiring up how the app reacts to user input. This approach provides developers with detailed control over interface construction and lifecycle behavior. The trade-off is that developers ultimately manage the details of the lifecycle and updates to the UI themselves.

In most real projects, everything gravitates toward UIViewController. Over time, those controllers turn into dumping grounds, mixing UI rendering, business logic, and state handling (Fowler, 2013).

SwiftUI flips that model. Instead of telling the UI how to update, the developer describes what interface should look like for a given state, and the framework handles the rest. As the underlying state changes, the interface recomputes itself. This shift changes not only coding style but also the way developers think about screen logic, navigation, and state ownership (Wiertel & Skublewska-Paszkowska, 2021; Zhou et al., 2025).

2.2 Software Migration Patterns

Software modernization literature commonly emphasizes that systems should rarely be replaced in a single step. Instead, gradual and controlled migration patterns are recommended in order to reduce risk, preserve stability, and enable validation throughout the transition process (Bragagnolo et al., 2023; Kemerer & Slaughter, 1999). Incremental migration and strangler-style modernization are especially appropriate when legacy systems remain operational and when direct comparison between old and new implementations is beneficial.

In the context of mobile UI migration, a phased approach is particularly relevant. The presentation layer is closely tied to user workflows, and abrupt replacement may introduce behavioral inconsistencies that affect user trust. A parallel migration strategy, in which the legacy UIKit version remains available while a SwiftUI version is developed feature by feature, provides a practical basis for comparison and reduces the risk of regressions.

2.3 Architectural Modularity and MVVM

Framework migrations are commonly motivated by maintainability and architectural concerns rather than runtime performance alone. In user-interface-intensive systems, maintainability depends partly on how clearly rendering, state management, navigation, and data access responsibilities are separated. When these concerns are concentrated in the same classes, the resulting coupling may make the presentation layer more difficult to test, extend, and refactor safely.

MVVM is compatible with SwiftUI's observable-state and one-way data-flow mechanisms. Views define rendering, ViewModels expose presentation state and process user actions, domain models represent application data, and services encapsulate backend interaction. This separation establishes clearer boundaries between interface rendering, state management, and infrastructure concerns.

2.4 Empirical Evidence and Research Gap

Existing studies provide complementary but incomplete evidence on SwiftUI adoption. Wiertel and Skublewska-Paszkowska (2021) compared UIKit and SwiftUI through implementation-level examples and reported differences in code structure and performance, but their study did not examine the migration of a complete existing application. Souha et al. (2024) compared several mobile application development frameworks from the perspective of developer choice; however, such cross-framework comparisons do not isolate the effects of migration between the two native iOS interface paradigms.

Baresi et al. (2024) examined the evolution of 140 native open-source iOS applications and found multiple patterns of SwiftUI adoption, including complete migrations, hybrid use of UIKit-SwiftUI, and adoption of SwiftUI for new components. Their work shows that migration occurs in practice, but does not measure whether migration improves maintainability or structural quality within functionally equivalent implementations. Zhou et al. (2025) investigated automated declarative UI code generation, illustrating a broader movement toward declarative development, although their research focuses on design-to-code generation rather than legacy UIKit migration.

Recent 2026 work supports both the architectural and implementation perspectives. Pankiv (2026) argues that modular iOS architectures facilitate component-wise UIKit-to-SwiftUI migration by reducing the system-wide impact of framework changes. Complementing this architectural

perspective, Gondi (2026) reported approximately 36.8% lower development time and 40% lower code volume in two functionally equivalent SwiftUI modules, although the limited experimental scope restricts generalization. Sinha et al. (2026) further showed that iOS runtime performance depends on measurable launch-time factors such as main-thread blocking, dependency initialization, and concurrency, reinforcing that performance cannot be inferred from structural metrics alone.

Taken together, these studies use different research techniques and therefore support different types of conclusions. Implementation-level examples are useful for explaining framework differences, adoption studies show how SwiftUI is used in practice, design-to-code studies evaluate declarative UI generation, and performance studies measure runtime behavior. However, none of these approaches alone establishes how the complete presentation layer of the same application changes when UIKit is reimplemented in SwiftUI while the backend and functional scope are preserved. Table 1 summarizes this positioning.

Table 1. Comparison of related research approaches and the present study

Study	Method	Focus	Limitation
Wiertel, 2021	Implementation comparison	UIKit vs. SwiftUI examples	Not a full migration case
Baresi, 2024	Repository mining	SwiftUI adoption patterns	Does not measure same-system structural change
Zhou, 2025	Declarative UI generation evaluation	Compilation, functionality, visual similarity	Not UIKit-to-SwiftUI migration
Gondi, 2026	Small experimental module comparison	Development time and code volume	Only two modules
Sinha, 2026	Runtime performance analysis	Launch-time measurement	Not a structural migration study
This study	Same-application case study	Presentation-layer structure	Single case, descriptive results

This study addresses a smaller gap by comparing the complete presentation layers of two implementations of the same e-voting application. Its contribution is not to make a universal claim about performance or maintainability. Instead, it provides a transparent structural comparison and examines how the simultaneous introduction of SwiftUI and the Model-View-ViewModel architecture affects the interpretation of the observed differences.

3 Methodology

3.1 Research Design

This research uses a descriptive comparative case study design. The unit of analysis is the complete presentation layer of an iOS e-voting application implemented in two forms: the original UIKit implementation and a feature-wise SwiftUI reimplementation of the same functional scope. The study compares the implementations, not the theoretically optimal implementations of either framework. Therefore, the results represent the properties of these two codebases under the specified design and measurement conditions. They do not determine the maximum capabilities of UIKit or SwiftUI.

The backend contract, domain model, business rules, and targeted user workflows were retained in both implementations. Backend redesign, data model redesign, blockchain redesign, and voting rule changes were outside the scope of the migration.

The SwiftUI implementation also introduced a feature-based Model-View-ViewModel organization. Consequently, the study does not treat all observed differences as effects of SwiftUI itself. It evaluates a combined modernization intervention consisting of a framework transition and presentation layer refactoring. The discussion distinguishes framework-related mechanisms, such as declarative rendering and reduced layout boilerplate, from refactoring-related mechanisms, such as dedicated ViewModels and feature-based separation of responsibilities.

Since one implementation pair from a single case was examined, the analysis is descriptive. Percentage differences are shown to summarize the observed codebases, but no inferential statistical tests or claims of statistical significance are provided.

3.2 Case System

The case system is eVote, an iOS application that simulates an electronic voting process. Both versions were built using Xcode 15.3, Swift 5.10, and a minimum iOS deployment target of 16.2.

The application uses Parse through Back4App as its backend. User-related data are stored in the _User collection, including fields such as username, password, name, surname, gender, dateOfBirth, regionName, regionCity, jmbg, vote, and role. Because the goal of this paper is to analyze UI modernization rather than backend evolution, the same data model and backend contract were preserved in the SwiftUI implementation.

3.3 Migration Procedure

The migration was performed feature-by-feature in a separate SwiftUI implementation, while the UIKit version was retained as a baseline for comparison. Fig. 1 summarizes the sequence used to preserve functional scope during the UIKit-to-SwiftUI migration and subsequent evaluation.

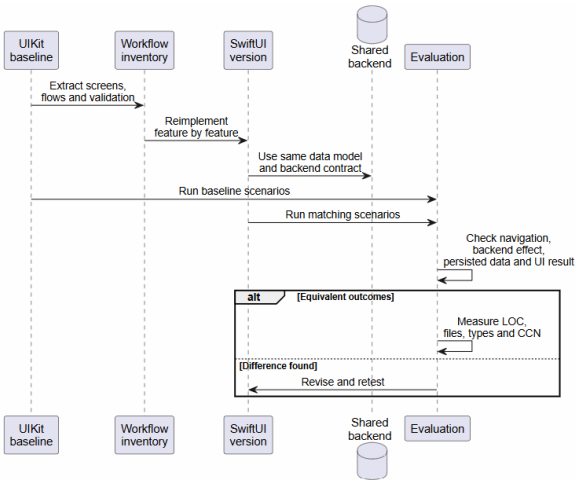


Figure 1. UML sequence diagram of the feature-wise migration and evaluation workflow

The migration was not performed as a line-by-line translation. UIKit UIView and UIViewController responsibilities were mapped to SwiftUI View structures, observable presentation state, and dedicated ViewModels. Backend access remained behind the existing service boundary. Repetitive interface patterns were extracted into reusable SwiftUI views or modifiers, and navigation was represented through SwiftUI state-driven navigation rather than controller-driven transitions.

3.4 Evaluation Metrics and Classification Rules

The evaluation measures the structural properties of the presentation layer. These properties can serve as indicators of code size and distribution of responsibilities, but are not direct measures of maintainability, testability, defect proneness, developer productivity, or usability.

The following metrics were collected:

- **UI-layer code lines**: physical source lines in production files whose primary responsibility was to render the interface, configure the interface, handle user input, or navigate the user interface.
- **Presentation-layer code lines**: UI-layer lines plus lines contained in dedicated ViewModel files.
- **UI implementation files**: source files classified as directly implementing the behavior of the rendered interface.
- **Presentation-layer files**: UI Implementation Files plus dedicated ViewModel files.
- **Screen-level interface types**: specific UIKit or SwiftUI types that represent complete application screens.
- **Reusable view component types**: interface types that are reused across multiple screens and are responsible for rendering visible content.
- **Controller types**: UIViewController subclasses responsible for orchestrating presentation.
- **Dedicated ViewModel types**: types whose primary responsibility was to expose presentation state and handle user-initiated presentation actions independently of the specific view being rendered.
- **Function-level cyclomatic complexity number (CCN):** the number of linearly independent control-flow paths detected within each function. The comparison considers the number of analyzed functions, mean and median function-level CCN, maximum CCN, and the distribution of functions across complexity bands.

The UIKit UI category included UIView subclasses, UIViewController subclasses, table view or collection view cells, and interface-specific navigation or rendering helpers. The SwiftUI UI category included View structures, ViewModifier implementations, navigation structures, theme-related rendering code, and reusable rendered components. A file is assigned according to its primary responsibility. It is not counted in more than one file category.

Generated code, third-party dependencies, test files, preview-only files, backend services, persistence code, domain models, resource catalogs, and build configuration files are excluded. Empty lines and comment-only lines are excluded. The same rule was applied to both codebases.

Percentage change is calculated only for metrics with directly comparable units and a non-zero baseline, including lines of code and number of files. Percentages are not calculated for the introduction or removal of architectural element types. Specifically, a change from zero to four ViewModel files is reported as the introduction of four ViewModels, not as a 100% increase.

3.5 Data Collection

Lines-of-code and file-count metrics were collected using a custom Python 3.9.6 script and cross-checked with ripgrep 15.1.0. The presentation-layer inventory contained 21 UIKit files and 12 SwiftUI files. Generated code, tests, preview-only files, models, backend services, persistence code, resources, and build-configuration files were excluded. Files with mixed responsibilities were classified according to their primary role.

Function-level cyclomatic complexity was measured using Lizard 1.23.0 with Swift analysis enabled. The same file inventories were used for both the size and complexity analyses. Mean, median, and maximum CCN values were calculated, together with

complexity-band and threshold distributions. Because some SwiftUI behavior is represented through computed properties and closures rather than conventional functions, the Lizard results are treated as supplementary structural indicators.

A second author independently reviewed the classifications and aggregate results, with disagreements resolved using the primary-responsibility rule.

4 Results

This section reports on the measured structural differences between the two presentation layer implementations. The results are organized around source code size, file organization, and redistribution of presentation responsibilities. They should not be interpreted as direct measurements of maintainability, productivity, error rate, runtime performance, or usability.

4.1 Code-Size and File-Organization Comparison

The UIKit and SwiftUI implementations differ substantially in measured presentation-layer size. The UIKit user-interface layer contained 2,832 lines of code, whereas the SwiftUI user-interface layer contained 814 lines, corresponding to an observed reduction of 71.3%. When the four SwiftUI ViewModel files were included, the SwiftUI presentation layer contained 1,013 lines, which was 64.2% smaller than the UIKit presentation-layer baseline.

The user-interface implementation also used fewer source files: 21 in UIKit and 8 in SwiftUI. Including ViewModels, the SwiftUI presentation layer contained 12 files, compared with 21 in UIKit.

The function-level complexity results were more nuanced than the size-based results. Lizard detected 112 functions in the UIKit presentation layer and 22 in the SwiftUI presentation layer. The mean CCN was 1.33 for UIKit and 1.73 for SwiftUI, while both implementations had a median CCN of 1.00. Therefore, the migration did not produce a reduction in average detected function complexity.

The upper end of the distribution was lower in the SwiftUI implementation. The maximum detected CCN was 9 in UIKit and 4 in SwiftUI. In UIKit, 109 of 112 functions (97.3%) had a CCN between 1 and 5, while three functions had a CCN between 6 and 10. All 22 detected SwiftUI functions had a CCN between 1 and 5. Neither implementation contained a detected function with a CCN greater than 10. These findings indicate similarly low typical function complexity in both implementations, but a narrower upper complexity range in the SwiftUI version.

4.2 Redistribution of Presentation Responsibilities

The migration changed the way presentation responsibilities were represented. The UIKit implementation used eight UIViewController types as the main units for coordinating rendering, input handling, navigation, and presentation state. The SwiftUI implementation did not include UIViewController types in the measured layer and introduced four dedicated ViewModels along with declarative View structures.

This result represents a redistribution, not an elimination, of complexity. SwiftUI reduced the explicit code for the lifecycle, layout, and interface updates, while the accompanying Model-View-ViewModel refactoring moved some responsibilities for state and action handling to dedicated types.

Thus, the disappearance of controllers is primarily related to the framework transition, while the introduction of dedicated ViewModels is an architectural design decision that could have been applied to UIKit as well.

The observed changes cannot be attributed exclusively to SwiftUI. The reduction in explicit layout, lifecycle, and interface-update code, together with the removal of UIViewController types, is primarily associated with SwiftUI's declarative presentation model. In contrast, the introduction of dedicated ViewModels and feature-based source organization resulted mainly from architectural refactoring. Reusable declarative components and state-driven navigation reflect the combined effects of SwiftUI and the redesigned presentation architecture. Therefore, the study evaluates the overall modernization intervention rather than isolating the independent causal effect of each change.

4.3 Implications for Maintainability, Testability, and Error Risk

Maintainability, testability, and defect proneness were not directly measured. Nevertheless, the structural results identify mechanisms that may influence these qualities.

The smaller interface layer reduces the amount of explicit layout and lifecycle code that must be inspected during interface changes. Dedicated ViewModels also create potential unit-testing boundaries for presentation state and user actions. Reusable views may reduce repeated visual configuration and inconsistent changes across screens.

These potential benefits are counterbalanced by SwiftUI-specific risks. Incorrect state ownership can cause unexpected interface updates, broad observable objects can introduce hidden coupling, and environment-provided dependencies may be less explicit than constructor-injected dependencies. Excessive view recomputation, complex bindings, and

state-driven navigation can also make behavior difficult to trace when responsibilities are poorly separated. Consequently, the structural changes create opportunities for improved maintainability and testability, but they do not prove that these outcomes occurred.

4.4 Functional Equivalence

Functional equivalence was assessed manually across eight core scenarios: registration with JMBG-based data derivation, valid login, role-based navigation, duplicate-vote prevention, candidate browsing and details, vote confirmation and persistence, display for users who had already voted, and administrator access to the vote overview. For each scenario, the navigation outcome, backend interaction, persisted-data effect, and user-visible result were compared with the UIKit baseline.

All eight scenarios produced equivalent outcomes in both implementations, and no unresolved discrepancies remained before structural measurement.

5 Discussion

The case study shows that a feature-wise SwiftUI reimplementation can produce a substantially smaller presentation layer and a clearer separation between rendered views and dedicated presentation state types. The strongest supported conclusion concerns structural compactness: according to the specified measurement rules, the SwiftUI version required fewer source lines and files.

The complexity results do not show a uniform reduction across all measures. Although the SwiftUI implementation had a lower maximum CCN and no functions in the 6–10 range, its mean detected function-level CCN was slightly higher than that of UIKit, and the median was identical. This suggests that the primary observed benefit was not a universal reduction in branching complexity, but the absence of the comparatively complex controller methods present in the UIKit implementation. The substantial difference in detected function counts must also be interpreted cautiously because declarative SwiftUI views and closures are not always represented as conventional functions by the analyser.

The results are consistent with Gondi's (2026) module-level comparison, which also associated SwiftUI implementations with reduced code volume, although that study examined only two modules rather than a complete presentation layer. They also complement research that shows that real-world applications adopt SwiftUI through multiple strategies, including full migration, hybrid coexistence, and use for newly developed UI components. However, unlike adoption studies, this study compares two complete implementations of a single functional scope.

The results should not be interpreted as implying that SwiftUI itself caused every architectural difference. UIKit can be combined with Model-View-ViewModel architecture, reusable components, coordinators, and service abstraction. In contrast, a poorly structured SwiftUI application may place business logic directly in views or concentrate state in large observable objects. In this case, the reduced layout and lifecycle code are likely related to SwiftUI, while the dedicated ViewModels and feature-based organization are mostly the result of architectural refactoring.

The results also differ in scope from studies that assess development time, runtime behavior, accessibility, or developer productivity. Since these outcomes were not measured here, the current results neither confirm nor refute claims related to these dimensions. This limitation is important because, as Sinha et al. (2026) show, iOS runtime behavior depends on launch-time factors such as main-thread blocking, dependency initialization, and concurrency rather than on source-code size alone.

Compared with module-level experiments, the present study provides broader internal coverage of one application but weaker external generalizability. Compared with repository-mining studies, it offers deeper measurement of a controlled same-system migration but lacks the scale of multi-project evidence. Compared with runtime-performance studies, it isolates structural presentation-layer changes but cannot support claims about launch time, responsiveness, or energy efficiency. Therefore, the contribution is best understood as a controlled structural case study rather than a general evaluation of SwiftUI as a framework.

5.1 Threats to Validity

5.1.1 Construct validity

The number of lines of code and files measures the structural size, not the maintainability or quality of the software as a whole. Classifying a file as UI or presentation code can also simplify files that contain multiple responsibilities. These risks are mitigated through explicit definitions and documented file inventory, but the metrics chosen remain proxies.

Cyclomatic complexity captures control-flow branching within detected functions but does not measure coupling, cohesion, state-management difficulty, declarative view complexity, or developer comprehension. It should therefore be interpreted as one structural indicator rather than a complete maintainability measure.

5.1.2 Internal validity

The SwiftUI implementation combined framework migration with Model-View-ViewModel refactoring, feature-based refactoring, and the ability to redesign existing code. The study, therefore, cannot isolate

SwiftUI as the sole cause of the observed differences. The authors' familiarity with both the application and the study's goals may also have influenced the implementation choices in favor of the migrated version. No implementation is considered globally optimal.

5.1.3 External validity

The study examines an iOS e-voting application with predominantly form-based, state-driven, and multi-stage workflows. Results may differ in applications dominated by complex animations, graphics rendering, extensive customization of collection views, multimedia processing, large hybrid code bases, or support for older iOS versions. The results support analytical comparisons with similar applications rather than statistical generalizations to all iOS software.

5.1.4 Reliability validity

Metric values depend on source boundaries, file classification, exclusion rules, and line-counting configuration. Fixed source revisions, documented scripting, and a reviewable file inventory improve repeatability. Independent replication is still needed to determine whether the same results were obtained by other developers or on other applications.

Lizard uses language-specific source parsing, and some SwiftUI computed properties and closure-based behaviors may not be represented as function-level records. Consequently, function counts and CCN distributions may partly reflect differences between imperative and declarative syntax as well as differences in application structure. The same Lizard version, configuration, threshold, and source-file boundaries were applied to both implementations to improve comparative consistency.

5.2 Practical Applicability

The migration approach is most directly applicable to applications whose presentation layers contain form input, validation, role-based navigation, stateful lists, and reusable interface patterns, while relying on a stable backend contract. For such systems, feature-based migration can allow teams to modernize presentation code without simultaneously replacing domain and backend functionality. This interpretation is also consistent with Pankiv's (2026) argument that modular iOS architectures can reduce the system-wide impact of component-wise framework migration.

The results should not be used as a numerical forecast for another project. Before migrating, teams should evaluate operating system support, existing UIKit dependencies, navigation complexity, test coverage, developer experience, accessibility requirements, and the cost of maintaining an interim hybrid architecture.

In practical terms, migrating from UIKit to SwiftUI is most defensible when the existing presentation layer contains duplicate layout code, state management that requires many controllers, and stable workflows that can be validated against the base implementation. It is less suitable as an immediate full-replacement strategy when the application depends on highly customized UIKit components, complex animation pipelines, unsupported older iOS versions, or insufficient regression testing. For such systems, a hybrid or feature-by-feature migration strategy is more realistic than a complete rewrite because it allows teams to gradually validate functional equivalence and structural impact.

6 Conclusion

This paper presented a comparative case study of a feature-wise migration from UIKit to SwiftUI in an iOS e-voting application. The UIKit and SwiftUI implementations retained the same backend contract, data model, business rules, and evaluated workflow scope.

According to the aforementioned classification and measurement rules, the SwiftUI implementation contained 71.3% fewer lines of code in the UI layer and 64.2% fewer total lines of code in the presentation layer. It also replaced the controller-centric presentation organization with declarative views and four dedicated ViewModels.

The defensible conclusion is that the examined SwiftUI implementation was structurally smaller and had more explicitly distributed presentation responsibilities. Some changes, notably the reduced explicit layout and lifecycle code, are likely related to SwiftUI, while others, including dedicated ViewModels and feature-based organization, are the result of accompanying architectural refactoring.

Future research should examine multiple applications and independent development teams, measuring cyclomatic or cognitive complexity, coupling, duplication, test coverage, effort spent on changes, error rate, programming time, runtime behavior, accessibility, and user task outcomes. Longitudinal studies would also be needed to determine whether the observed structural differences translate into lower maintenance effort over time.

Acknowledgments

This research has been supported by the Ministry of Science, Technological Development and Innovation (Contract No. 451-03-34/2026-03/200156) and the Faculty of Technical Sciences, University of Novi Sad through project "Scientific and Artistic Research Work of Researchers in Teaching and Associate Positions at the Faculty of Technical Sciences, University of Novi Sad 2026" (No. 01-3609/1).

References

Baresi, L., Di Penta, M., Quattrocchi, G., & Tamburri, D. A. (2024). How have iOS Development Technologies Changed over Time? A Study in Open-Source. *Proceedings of the IEEE/ACM 11th International Conference on Mobile Software Engineering and Systems*, 33–42. https://doi.org/10.1145/3647632.3647988

Bragagnolo, S., Anquetil, N., Ducasse, S., Seriai, A., & Derras, M. (2023). *Software Modernisation: Theoretical Framework*. SSRN. https://doi.org/10.2139/ssrn.4411437

Fowler, M. (2013). *Patterns of enterprise application architecture* (Nineteenth printing). Addison-Wesley.

Gondi, M. L. (2026). A comparative study on the adoption of SwiftUI over UIKit in modern iOS application development. *International Journal of Research and Scientific Innovation*, *13*(1). https://doi.org/https://doi.org/10.51244/IJRSI.2026.13010214

Kemerer, C. F., & Slaughter, S. (1999). An empirical approach to studying software evolution. *IEEE Transactions on Software Engineering*, *25*(4), 493–509. https://doi.org/10.1109/32.799945

Pankiv, O. (2026). *Methodology for designing fault-tolerant modular SDK systems for scalable mobile platforms*. Internauka Publishing House. https://doi.org/https://doi.org/10.25313/Pankiv-2026-353

Sinha, P., Karanjkar, R., & Dutta, A. (2026). iOS app start-time performance: A comprehensive analysis and optimization framework. *International Journal of Computer Applications*, *187*(72).

Souha, A., Benaddi, L., Ouaddi, C., & Jakimi, A. (2024). Comparative analysis of mobile application Frameworks: A developer's guide for choosing the right tool. *Procedia Computer Science*, *236*, 597–604. https://doi.org/10.1016/j.procs.2024.05.071

Wiertel, P., & Skublewska-Paszkowska, M. (2021). Comparative analysis of UIKit and SwiftUI frameworks in iOS system. *Journal of Computer Sciences Institute*, *20*, 170–174. https://doi.org/10.35784/jcsi.2662

Zhou, T., Zhao, Y., Hou, X., Sun, X., Chen, K., & Wang, H. (2025). DeclarUI: Bridging Design and Development with Automated Declarative UI Code Generation. *Proceedings of the ACM on Software Engineering*, *2*(FSE), 219–241. https://doi.org/10.1145/3715726

A State-Indexed Execution Model for Embedded Systems

Matúš Nečas
University of Žilina
Faculty of Electrical Engineering and Information Technology
Univerzitná 8215/1, 010 26 Žilina, Slovakia
matus.necas@feit.uniza.sk

Roman Budjač
University of Žilina
Research Centre
Univerzitná 8215/1, 010 26 Žilina, Slovakia
roman.budjac@uniza.sk

Juraj Ďudák
Slovak University of Technology in Bratislava
Faculty of Materials Science and Technology in Trnava
Jána Bottu 25, 917 24 Trnava, Slovakia,
University of Žilina
Research Centre
Univerzitná 8215/1, 010 26 Žilina, Slovakia
juraj.dudak@stuba.sk

Abstract. *This work introduces the State-Indexed Application Loop (SIAL), a lightweight implementation pattern for microcontroller-based systems. SIAL decomposes application logic into a fixed set of indexed states, where each state encapsulates both computation and successor selection through a single activation function. Rather than introducing a fundamentally different finite-state-machine formalism, SIAL organizes state-oriented application logic by embedding successor selection directly into individual states, while the Scheduler performs indexed state selection and invocation. The model preserves deterministic sequential execution* over a statically allocated state space while remaining compatible with interrupt-driven asynchronous inputs via a shared system context. The resulting organization provides a uniform structure for state-oriented control flow with constant-time indexed state selection. The presented properties are analytical and structural; quantitative performance advantages require implementation-specific experimental validation.

Keywords. control flow design, embedded systems, finite-state machines, software architecture, state-based execution

1 Introduction

Microcontrollers have become an integral part of modern electronic systems and represent a fundamental computing platform for a wide range of applications (Hee et al., 2021). Their usage spans from simple data acquisition and processing devices, through wearable electronics and body sensing systems, industrial automation applications, to complex embedded solutions equipped with user interfaces, wireless communication, and cloud connectivity (Lestariningsih et al., 2019; Madhu and Vyjayanthi, 2018; Zavitsanou et al., 2016). With the increasing performance of modern hardware, the complexity of software deployed on these platforms continues to grow.

The development of embedded applications requires knowledge across multiple domains. In addition to low-level programming and hardware peripheral handling, it involves analog and digital signal processing, communication protocol implementation, data flow management, user interface design, and coordination of concurrent activities. In recent years, machine learning techniques, including computer vision and neural networks, have also been integrated into embedded systems, further increasing the demands on modularity, maintainability, and scalability of application firmware (Sanchez-Iborra and Skarmeta, 2020).

Applications with user interfaces represent a particularly challenging category. Such systems must ensure smooth interaction between the user and the device through physical controls, touch screens, or their combination (Necas et al., 2026). They typically operate in multiple modes or screens between which the user dynamically navigates. Each mode requires dedicated logic for input processing, validation and update of application data, control of displayed content, and handling of events generated by both the user and the system.

The increasing complexity of embedded applications places high demands on software architecture design (Hofmeister et al., 2000). Poorly designed program structures may reduce clarity of control logic, increase maintenance costs, and limit extensibility (Ganssle, 2008). For this reason, various architectural

approaches are used in practice, ranging from simple main-loop-based designs (superloop), through finite-state machine (FSM) implementations, to event-driven frameworks and real-time operating systems (White, 2024).

Each of these approaches has specific advantages and limitations. Finite-state machines provide an explicit representation of operational modes and transitions, but their implementation may take different forms, including *centralized switch statements*, *transition tables*, and *function-pointer-based state handlers*. The choice of implementation therefore affects how state behavior and transition logic are organized in the source code. Event-driven architectures provide a different approach by distributing execution across event handlers and dispatching mechanisms, while real-time operating systems introduce scheduling and concurrency abstractions that are not required by all embedded applications (Michelson, 2006).

For a class of microcontroller applications characterized by a fixed set of operational modes, sequential execution, and statically allocated resources, a lightweight organization of state-oriented application logic can be preferable to a more general execution framework. In such systems, the main design concern is not necessarily concurrency or dynamic event dispatching, but maintaining a clear relationship between the currently active application mode, the code executed in that mode, and the selection of the subsequent mode.

This work presents the *State-Indexed Application Loop* (SIAL), a lightweight implementation pattern addressing this organization problem. SIAL structures application logic as a statically allocated collection of indexed states. Each state exposes a uniform `activate()` operation that performs state-specific processing and returns the identifier of the next state. The scheduler is intentionally kept minimal: it selects the active state by index, invokes its activation operation, and stores the returned state identifier for the next execution step.

The contribution of this work is not a new finite-state-machine semantics. Instead, SIAL defines a specific implementation organization for state-oriented embedded applications, emphasizing static allocation, uniform state activation, and local successor selection. The paper describes the resulting execution model formally and discusses its relationship to conventional FSM implementation techniques, as well as its applicability, advantages, and limitations in resource-constrained microcontroller systems.

2 Related Work

Existing approaches for managing software complexity in embedded systems can be broadly categorized according to their execution model, including event-driven frameworks, real-time operating systems, and state-oriented or hybrid architectures. The main objective of these approaches is to improve modularity and scalability of application logic compared to the classical superloop paradigm.

Event-driven frameworks represent one of the dominant design strategies. QP™ Real-Time Embedded Frameworks (RTEFs) implement a hierarchical state machine model combined with an event-driven execution mechanism. System behavior is modeled through active objects that communicate via event queues, enabling explicit representation of system dynamics and separation of concerns. While this improves modularity and scalability, it introduces an additional event management layer that can obscure direct execution flow and increase architectural complexity (Quantum Leaps, LLC, 2014).

A similar event-driven principle is widely used in graphical user interface libraries such as LVGL (Light and Versatile Graphics Library). In this case, application logic is decomposed into callback functions bound to user interface events (LVGL Kft., 2024). This enables flexible and portable UI development across platforms. However, distributing logic across callbacks can reduce traceability of execution flow and make global reasoning about system behavior more difficult.

Another class of solutions is represented by real-time operating systems such as FreeRTOS and Zephyr. These systems adopt a scheduler-based execution model in which application logic is decomposed into concurrent tasks. Execution is coordinated using synchronization primitives and inter-task communication mechanisms. While this improves modularity and enables concurrency, it also introduces complexity related to scheduling, synchronization, and shared resource management (Barry and FreeRTOS Team, 2023).

In addition to these approaches, state-machine-based extensions integrated into RTOS environments, such as the Zephyr State Machine Framework, combine explicit state modeling with operating system scheduling. This enables structured representation of application behavior while maintaining real-time capabilities (The Zephyr Project Contributors, 2025). However, application logic is distributed across both the state-machine layer and the underlying RTOS infrastructure, which increases conceptual fragmentation.

Overall, existing approaches represent different strategies for addressing the limitations of the superloop architecture. Event-driven systems emphasize modularity through indirect execution, RTOS-based systems introduce concurrency and scheduling abstractions, while state-machine frameworks provide structured behavioral modeling at the cost of additional architectural layers. SIAL instead focuses on the organization of state-oriented application logic at the implementation level, without requiring an event dispatcher, event queue, or operating system scheduler.

Recent work continues to explore lightweight software architectures and state-machine-based ap-

proaches for resource-constrained embedded systems. These approaches emphasize structured state representation, modular control flow, and reduced runtime infrastructure, particularly for microcontroller-oriented applications. Such developments demonstrate the continued relevance of state-oriented execution models in embedded software, while also highlighting the variety of implementation techniques available for realizing them (Ali and Vitek, 2022; Babiuch and Foltynek, 2024; Ball et al., 2024).

SIAL is positioned within this context as a lightweight implementation pattern that focuses specifically on state-local execution and statically indexed control flow rather than introducing a fundamentally new finite-state-machine formalism.

2.1 Conventional FSM Implementation Techniques

Finite-state machines can be implemented using several conventional software structures. A switch-based implementation commonly represents the active state using an enumerated value and evaluates the corresponding state behavior through a centralized `switch` statement. A table-driven implementation stores transition information explicitly in a table indexed by the current state and input condition. Another common approach uses function pointers to associate each state with a handler function, allowing the active handler to be selected through an indexed structure.

SIAL is structurally related to the function-pointer approach. The distinction is primarily organizational rather than semantic: SIAL defines a uniform state abstraction in which the state activation operation performs both state-specific processing and successor selection, while the scheduler remains responsible only for indexed state selection and invocation. SIAL is therefore positioned as an implementation pattern for state-oriented control flow rather than as a fundamentally different finite-state-machine formalism.

3 Materials and Methods

The *State-Indexed Application Loop* (SIAL) is a lightweight implementation pattern for microcontroller-based systems. It is derived from the classical superloop architecture, which represents one of the most widely used execution paradigms in embedded applications.

3.1 Baseline Execution Model: Superloop

In the superloop model (as shown in Fig. 1), the application is structured into an initialization phase (*Setup Logic*) followed by an infinite cyclic execution loop (*Cyclic Logic*). Within this loop, all application logic is executed sequentially, while asynchronous behavior is typically handled through interrupts.

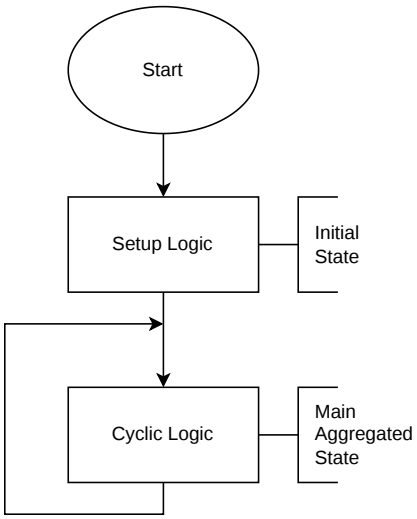


Figure 1. Superloop execution model

Although this model is widely adopted due to its simplicity, deterministic execution flow, and low runtime overhead, it does not scale well in systems with multiple operational modes. As system complexity increases, the application logic becomes concentrated within a single execution block, leading to a monolithic structure with strong coupling between functional components. From a software engineering perspective, this corresponds to a single aggregated state without explicit decomposition into independent behavioral units.

To address these limitations, SIAL introduces a decomposition of the monolithic execution loop into a set of explicitly defined states with localized runtime logic and transition behavior.

3.2 State-Driven Decomposition and RACe Execution Rule

SIAL is based on the idea that the monolithic control logic of a superloop-based system can be decomposed into a set of explicitly defined states. Each state represents a distinct operational mode of the system and encapsulates the functionality associated with that mode.

This decomposition transforms the application structure from a single centralized control flow into a collection of self-contained functional units. As a result, system behavior is no longer expressed as a single cyclic logic, but as an interaction between explicitly defined states.

To realize this decomposition, SIAL uses the RACe (*Run And Choose Next*) execution rule, a term introduced in this work to describe the state activation and successor-selection mechanism used by SIAL.

Within the RACe execution rule, each state executes its associated functionality and determines its successor based on the current system context. Computation and successor selection are therefore integrated into a single state-local operation.

This organization avoids the need for a separate centralized transition-handling layer and places application-specific successor selection within the corresponding state implementation, while the scheduler remains responsible for selecting and invoking the active state.

3.3 System Architecture

The SIAL architecture consists of two tightly connected components: the *Scheduler* and the *State* abstraction, as illustrated in Fig. 2.

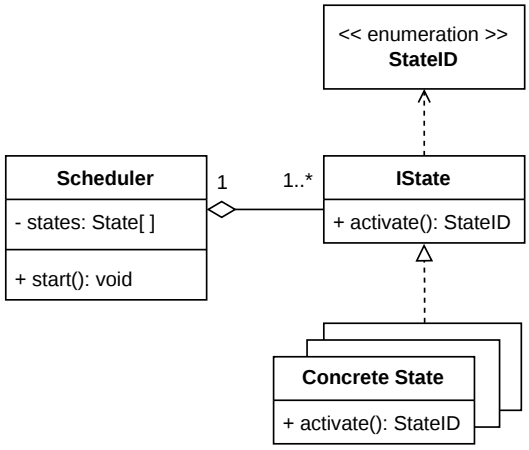


Figure 2. Architecture of the SIAL framework

The State abstraction is defined through the interface `IState`. Each state encapsulates a self-contained portion of application logic and exposes a single activation method, denoted as `activate()`. This method serves as the execution entry point of the state and provides a unified operation for state-specific processing and successor selection

The activation method processes the current system context and returns the identifier of the next state, thereby implementing the RACe execution rule at the software level. In this sense, each state contains its own successor-selection logic as part of its execution.

The Scheduler is responsible for controlling the execution flow of the system. It maintains the identifier of the currently active state and invokes the corresponding state through the `IState` interface (as shown in Fig. 3). After each execution, the Scheduler updates its internal state index based on the value returned by the active state. This creates a strict execution cycle in which sequencing is centralized in the Scheduler, while application-specific successor selection is distributed across states.

3.4 State Lifecycle Semantics

Each state may be activated repeatedly while it remains the current state. A self-transition therefore corresponds to repeated activation of the same state without leaving its operational mode. A transition to a different state terminates the current state activation sequence and makes the target state the next active state. One-time initialization and cleanup operations are not implicitly provided by the `activate()` interface; they are explicitly implemented by the application, for example by detecting the first activation or the transition into a state. Similarly, re-entry into a previously visited state is treated as a new activation sequence, allowing the application to explicitly perform any required re-initialization.

This lifecycle definition intentionally avoids imposing separate *entry*, *do*, and *exit* methods. Their functionality can instead be represented by application-level conditions within `activate()`, or by auxiliary state-local mechanisms when required. Consequently, SIAL defines the execution structure but does not prescribe a specific policy for initialization, cleanup, or re-entry behavior.

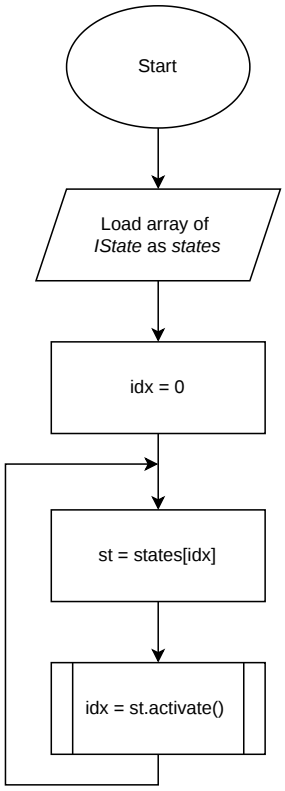


Figure 3. SIAL execution flow

3.5 Illustrative Example: Comparative Execution Models

To illustrate the organization of state-oriented application logic in SIAL and its relationship to conventional FSM implementations, consider a typical multi-state embedded application with human interaction (Fig. 4). The system operates in four distinct modes: Booting (B), Calibration (C), Measurement (M), and Failure (F).

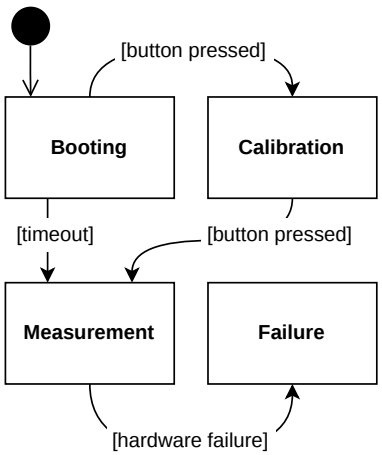


Figure 4. An example of typical embedded four-state system

The system reacts to a set of discrete events, namely timeout (tm), button press (bp), and hardware failure (hf). The resulting transition structure can be represented by the partial transition relation in Table 1.

Each state in the system corresponds to a specific operational mode and defines its internal behavior and possible successor states.

The Booting state (B) represents the system initialization phase, during which hardware and software components are brought into a valid operational configuration. Upon successful initialization, the system transitions either to the Calibration state (C) or directly to the Measurement state (M), depending on the system context and startup conditions.

The Calibration state (C) is responsible for establishing reference parameters required for accurate operation. Once calibration is completed, the system proceeds to the Measurement state (M), which represents the primary functional mode of the application.

The Measurement state (M) implements the core application logic, continuously processing incoming data and performing system-specific computations. During operation, it monitors for abnormal conditions; in the case of a detected hardware failure (hf), the system transitions to the Failure state (F).

The Failure state (F) represents an error-handling or recovery-oriented mode in which normal operation is suspended. Depending on the implementation, it may remain active until external intervention or recovery conditions are satisfied.

In a conventional FSM, the transition relation can be represented explicitly through a transition function or table, while the state-specific behavior may be organized separately or combined with transition handling depending on the implementation technique. SIAL follows the latter organization by associating state processing and successor selection with the active state through its `activate()` operation. The resulting structure is typically represented by a transition relation such as the one shown in Table 1.

state	B	C	M	F
B		bp	tm	
C			bp	
M				hf
F				

Table 1. Partial transition relation between system states

Event-driven architectures, such as GUI-based systems, implement transitions through event callbacks. While this reduces the need for an explicit transition table, it introduces a separation between event handling and cyclic application logic (e.g., measurement loops), often requiring additional coordination mechanisms to maintain consistent system behavior, typically realized through shared or global state variables.

FSM implementations can organize transition logic and state-specific processing in different ways depending on the selected implementation technique. Switch-based and table-driven approaches commonly provide an explicit representation of transition logic, while function-pointer-based implementations can combine state processing and successor selection within individual state handlers. As the number of states and transitions grows, these different representations involve different trade-offs in terms of code organization, traceability, and separation of concerns. In event-driven designs, application behavior is distributed across event handlers and other execution mechanisms, which may require additional coordination to maintain consistent state behavior.

In contrast, the SIAL implementation pattern integrates state behavior and successor selection into a single state-local activation operation. Each state evaluates the system context and determines its successor directly, resulting in a uniform per-state abstraction regardless of the number of possible successor states. In the illustrative four-state example, this corresponds to four state implementations, each exposing the same `activate()` interface, while the Scheduler provides a common mechanism for selecting and invoking the active state.

3.6 Formal Model

The organization of SIAL can be described using the classical finite-state machine model as a reference. The purpose of the formalization is to describe how state-specific processing and successor selection are combined within the SIAL implementation pattern.

The behavior of classical deterministic finite-state machines is defined as

$$FSM = (S, E, \delta, s_0), \tag{1}$$

where S is a finite set of states, E is a set of input

events, $\delta : S \times E \to S$ is a centralized transition function, and s_0 is the initial state (Lee and Yannakakis, 1996).

The system evolution is given by

$$s_{k+1} = \delta(s_k, e_k), \tag{2}$$

where $e_k \in E$ denotes the event observed at step k.

SIAL distributes the state-specific successor selection across the state implementations. The system is defined as

$$SIAL = (S, A, s_0), \tag{3}$$

where $A = \{a_0, a_1, \ldots, a_{n-1}\}$ is a set of state-specific activation functions, and each function a_i is realized in software through the corresponding `activate()` method of a state.

Each activation function is defined as

$$a_i : X \to X \times S, \tag{4}$$

where X denotes the system context. The output therefore consists of the evolved system context and the identifier of the successor state.

In this model, the system context X represents the complete set of observable runtime information available to the application at a given moment. This includes external inputs (e.g., user events or sensor readings), internal state variables, and values updated asynchronously, typically through interrupt service routines or peripheral interfaces.

The global execution semantics are defined as

$$(X_{k+1}, s_{k+1}) = a_{s_k}(X_k), \tag{5}$$

where X_k is the system context at step k.

The resulting transition relation between states is given by

$$T = \{(s_i, s_j) \mid \exists X, X' : a_i(X) = (X', s_j)\}, \tag{6}$$

which defines a directed graph over S induced by the possible successor states of the activation functions.

4 Results

The *State-Indexed Application Loop* (SIAL) was evaluated from a structural and execution-semantics perspective. The resulting properties follow directly from the organization of the implementation pattern and describe state selection, execution structure, and memory organization. They should be interpreted as analytical properties rather than experimentally measured performance results.

The SIAL implementation uses a statically allocated indexed state space

$$S = \{0, 1, \ldots, n-1\}. \tag{7}$$

Each state corresponds directly to an entry in a fixed array structure, where the array index serves as the unique state identifier. System execution is driven by the currently active state s_k. At each step, the corresponding activation function processes the current system context and returns the successor state (5),

The system evolution is therefore governed by the discrete update rule

The resulting execution trace forms a discrete sequence in the state space S

$$s_0, s_1, s_2, \ldots \tag{8}$$

where each successor state is selected by the activation function of the currently active state.

A direct consequence of the indexed state organization is that selecting the active state does not require searching through the state set. The scheduler therefore performs indexed state selection in constant time,

$$\tau_{\text{scheduler}} = O(1). \tag{9}$$

This result applies to the state-selection operation performed by the Scheduler and does not describe the execution time of `activate()`, which depends on the application-specific processing implemented by each state.

The analysis also shows that SIAL provides a uniform implementation structure for states. Each state exposes one`activate()` operation that combines state-specific processing with successor selection. The number of possible successor states therefore does not change the interface or the scheduler structure. However, the corresponding decision logic remains part of the state implementation and may increase its internal complexity as the number of possible transitions grows.

From a memory-organization perspective, the state collection can be allocated statically as a fixed array. The scheduler requires only the current state index and the state collection, without requiring dynamically allocated state objects or a runtime registration mechanism. This provides compatibility with embedded applications where static memory allocation is preferred.

The resulting properties are summarized in Table 2. The table describes structural characteristics derived from the respective execution approaches; the qualitative entries are therefore design assessments rather than experimentally measured performance values.

5 Discussion and Conclusions

The *State-Indexed Application Loop* (SIAL) provides a state-centric implementation pattern for embedded systems in which both computation and successor selection are encapsulated within individual states. This organization shifts application-specific control-flow decisions from a separate transition-handling structure to individual state implementations, while the Scheduler

Table 2. Structural characteristics of different execution approaches

Property	FSM	Event-driven	SIAL
Explicit state representation	Yes	Partial	Yes
State-local processing	Depends	Depends	Yes
State-local successor selection	Depends	Depends	Yes
Indexed state selection	Depends	No	Yes
Static memory allocation compatibility	Yes	Depends	Yes
Separate transition representation	Depends	Depends	No
External event dispatcher	No	Typically	No
Scheduler-based concurrency	No	Depends	No
Asynchronous event handling	Depends	Yes	Via shared context
Dynamic runtime extensibility	Depends	Yes	No

remains responsible for selecting and invoking the active state.

In contrast to event-driven architectures, which rely on external dispatching mechanisms such as callback registration, event queues, or observer-based models, SIAL does not require an intermediate event-handling layer for its basic execution. While event-driven systems provide flexibility in highly asynchronous environments by decomposing behavior into event-triggered handlers, this may distribute application logic across multiple components and execution mechanisms.

By embedding successor selection directly into state execution, SIAL establishes a direct relationship between the current system context and the next state. This reduces the separation between state-specific processing and transition selection and provides a uniform organization of state-oriented control flow.

A key structural property of the SIAL implementation pattern is its reliance on static memory allocation and a fixed state space, which aligns well with microcontroller-based systems where dynamic allocation is typically avoided due to fragmentation risks and non-deterministic memory behavior. Unlike classical object-oriented State design patterns, which may depend on dynamic polymorphism and heap-based allocation, SIAL can be implemented using statically defined state structures.

These design choices introduce an inherent trade-off: while state-local organization can simplify the distribution of application logic, encapsulating both computation and successor selection within states can increase per-state complexity, particularly in systems with dense or highly interconnected transition structures. In such cases, complexity is shifted from a separate transition representation to individual state implementations, which may reduce separation of concerns compared to FSM implementations where transitions are explicitly externalized.

On the level of execution semantics, SIAL follows a strictly sequential scheduling model and does not provide hardware-supported concurrency mechanisms such as preemptive multitasking or RTOS-based execution. Nevertheless, asynchronous inputs can be incorporated through interrupt-driven updates of the shared system context, provided that access to shared data satisfies the required atomicity and synchronization assumptions.

Although the implementation presented in this work uses an object-oriented abstraction based on state interfaces, the underlying organization is not bound to a specific programming paradigm and can also be implemented in procedural environments.

The results presented in this work establish the main structural and execution properties of SIAL, including indexed state selection, static state organization, and state-local successor selection. These properties describe the intended characteristics of the implementation pattern; quantitative performance advantages such as reduced runtime overhead, memory consumption, or execution time require validation through implementation-specific measurements.

The scalability of the indexed execution mechanism should also be distinguished from the scalability of the application structure itself. Even for systems containing hundreds of states, indexed state selection remains $O(1)$; however, the number of state implementations and the amount of state-local control logic increase with system size. Consequently, large and densely connected state spaces may become more difficult to debug and reason about, particularly when transitions occur between states that are distant in the overall application structure. In such systems, explicit state documentation, naming conventions, and transition tracing mechanisms may be required to maintain readability and debugging efficiency.

Overall, SIAL is positioned as a lightweight implementation pattern for embedded applications that emphasizes static structure, sequential execution, and locality of control flow. It is particularly applicable to systems with well-defined operational modes, where predictable sequential execution and static resource organization are important design considerations. Its main limitation is the reduced suitability for highly asynchronous or concurrent applications, where event-driven frameworks or real-time operating systems may provide more appropriate execution mechanisms.

Acknowledgments

This research was supported by the project VEGA through the Research on motion data analysis methods for applications in diagnosis and therapy of gnostically relevant symptoms under Grant 1/0147/25.

References

Ali, K., & Vitek, J. (2022). *Lipics, volume 222, ecoop 2022, complete volume* (Vol. 222). Schloss Dagstuhl – Leibniz-Zentrum für Informatik. https://doi.org/10.42302022

Babiuch, M., & Foltynek, P. (2024). Implementation of a universal framework using design patterns for application development on microcontrollers. *Sensors*, *24*(10). https://doi.org/10.3390/s24103116

Ball, T., Halleux, J. " d., Devine, J., Hodges, S., & Moskal, M. (2024). Jacdac: Service-based prototyping of embedded systems. *Programming Language Design and Implementation (PLDI)*. https://www.microsoft.com/en-us/research/publication/jacdac-pldi2024/

Barry, R., & FreeRTOS Team. (2023). *Mastering the freertos™ real time kernel: A hands-on tutorial guide* (Version 1.1.0) [Available online]. Amazon Web Services. https://www.FreeRTOS.org

Ganssle, J. (2008). *The art of designing embedded systems*. Newnes.

Hee, Y. H., Ishak, M. K., Asaari, M. S. M., & Seman, M. T. A. (2021). Embedded operating system and industrial applications: A review. *Bulletin of Electrical Engineering and Informatics*, *10*(3), 1687–1700. https://doi.org/10.11591/eei.v10i3.2526

Hofmeister, C., Nord, R., & Soni, D. (2000). *Applied software architecture*. Addison-Wesley Professional.

Lee, D., & Yannakakis, M. (1996). Principles and methods of testing finite state machines-a survey. *Proceedings of the IEEE*, *84*(8), 1090–1123. https://doi.org/10.1109/5.533956

Lestariningsih, T., Artono, B., Hidayatullah, N., & Kusbandono, H. (2019). Microcontroller and android hmi based water level and control system. *EAI Endorsed Transactions on Internet of Things*, *5*(17).

LVGL Kft. (2024). *Lvgl — light and versatile graphics library documentation* (v9.0) [Dostupné online; navštívené jún 2026]. LVGL Kft. https://docs.lvgl.io/9.0/

Madhu, G. M., & Vyjayanthi, C. (2018). Implementation of cost effective smart home controller with android application using node mcu and internet of things (iot). *2018 2nd International Conference on Power, Energy and Environment: Towards Smart Technology (ICEPE)*, 1–5. https://doi.org/10.1109/EPETSG.2018.8659128

Michelson, B. M. (2006). Event-driven architecture overview. *Patricia Seybold Group*, *2*(12), 10–1571.

Necas, M., Dudak, J., Gaspar, G., Husar, P., & Gurin, D. (2026). Embedded digital instrumentation for automated assessment of sensory perception. *IEEE Access*, *14*, 75933–75944. https://doi.org/10.1109/ACCESS.2026.3693773

Quantum Leaps, LLC. (2014, May). *Qp/c++ and the qt gui framework* (Application Note No. Document Revision K). Quantum Leaps, LLC. Chapel Hill, NC. http://www.state-machine.com/qt

Sanchez-Iborra, R., & Skarmeta, A. F. (2020). Tinyml-enabled frugal smart objects: Challenges and opportunities. *IEEE Circuits and Systems Magazine*, *20*(3), 4–18. https://doi.org/10.1109/MCAS.2020.3005467

The Zephyr Project Contributors. (2025). *Zephyr project documentation (latest), version 4.2.99* [Accessed: 2025-09-02]. The Linux Foundation. https://docs.zephyrproject.org/latest/index.html

White, E. (2024). *Making embedded systems*. " O'Reilly Media, Inc."

Zavitsanou, S., Chakrabarty, A., Dassau, E., & Doyle, F. J. (2016). Embedded control in wearable medical devices: Application to the artificial pancreas. *Processes*, *4*(4). https://doi.org/10.3390/pr4040035

Security technical debt and scanner-context artefacts in SME GitHub repositories: A paired pre/post-2023 SonarQube study

Tara Nikolić, Dario Ljubas, Goran Nikšić
University of Zagreb Faculty of organization and informatics
Department of information systems development
Pavlinska 2, 42000 Varaždin, Croatia
{tnikolic, dljubas}@foi.hr, gniksic21@student.foi.hr

Abstract. *This study examines how scanner scope affects temporal measurements of security technical debt in GitHub repositories maintained by small and medium-sized enterprises. A paired design compares repository states before and after 1 January 2023 using SonarQube, size-normalised debt measures, non-parametric tests, and manual classification of high-severity findings. Because AI use cannot be verified, the temporal comparison is descriptive rather than causal. Results indicate stable aggregate debt density and show that non-production files can substantially distort apparent severity changes. These findings demonstrate the importance of separating production from non-production code when comparing SAST results over time.*

Keywords. security technical debt, static analysis, SME, SonarQube, software security, AI-assisted development

1 Introduction

Temporal comparisons of static application security testing (SAST) output are increasingly employed to assess changes in software security risk. However, the validity of these comparisons depends on the scope defined by the scanner. Non-production artefacts such as test fixtures, localisation files, and vendored code often change more rapidly than production code and can disproportionately influence perceived shifts in severity. This study examines security technical debt in paired states of repositories maintained by small and medium-sized enterprises (SMEs) and investigates how the composition of production code affects the interpretation of longitudinal SAST results. The primary contribution is methodological: we evaluate what can be inferred from repository-level scanner output before attributing changes to developer behaviour or tooling.

The rapid adoption of large language model (LLM) coding assistants raises questions regarding potential changes in security debt over time. Nevertheless, this study neither observes nor infers AI usage, and no reliable ground truth for AI authorship is available in the public repositories analysed. The 1 January 2023 threshold serves solely to delineate two temporally distinct repository states, rather than to distinguish between exposed and unexposed groups. Consequently, all period results are descriptive rather than causal and cannot be used to estimate AI-induced security debt. Prior research on AI-assisted code is referenced as motivation, but it is not the construct measured in this study (Pearce et al., 2022; Sandoval et al., 2023).

Small and medium-sized enterprises (SMEs) are particularly vulnerable to this uncertainty. SMEs differ fundamentally from large organisations, operating within distinct economic and managerial contexts. They rarely maintain dedicated IT functions and often centralise decision-making in a single owner-manager who may lack the resources or expertise for systematic security governance (Devos et al., 2012). In contrast to large firms, which distribute verification across peer review, application-security teams, automated pipelines, and penetration testing, SMEs typically rely solely on the developer who accepted the suggestion (Chidukwani et al., 2022). If AI assistance accelerates the accumulation of security weaknesses, SMEs constitute the environment where this effect is likely to be most pronounced and least mitigated.

This paper makes three principal contributions. First, we present a paired, size-normalised assessment of security technical debt (STD) across 28 SME repositories, each analysed in two distinct states (56 analyses in total), encompassing 13.7 million lines of code. We detail the SonarQube configuration required to interpret the results and explain why the aggregate design is predisposed toward a null period difference. Second, manual classification of every high-severity finding reveals that the most prominent raw difference is attributable primarily to scanner context rather than to a genuine security signal. Third, we demonstrate that hotspot states observed in new research instances cannot be used to infer organisational triage behaviour and specify the additional evidence necessary to support such claims.

2 Background

This study is informed by two bodies of literature. The first conceptualizes security weaknesses as a form of technical debt, characterized by a principal, an interest rate, and a repayment mechanism. The second examines whether AI-assisted development alters the rate at which such debt accumulates. Evidence regarding this latter question remains inconclusive.

2.1 From technical debt to security debt

The technical debt (TD) metaphor conceptualizes expedient implementation choices as a loan, with interest repaid through increased maintenance costs (Avgeriou et al., 2016; Cunningham, 1993; Li et al., 2015). However, security does not fully correspond to this framework. An unresolved vulnerability remains dormant until exploited, and the resulting cost may not correspond to the original shortcut. Rindell et al. (2019) address this asymmetry by way of the ACT-S framework, which reinterprets security risk as technical debt and enables established management practices such as identification, measurement, prioritisation, and repayment. Martinez et al. (2021) define security debt as the incomplete or incorrect set of security properties required by a system in its operational context, with interest measured in accumulated risk rather than maintenance effort.

AI-induced security debt is defined in the literature as accumulated and unresolved security risk that arises or is amplified when software artefacts are produced or influenced by AI tools. Rather than constituting a new category, it represents an accelerated source of the previously described debt, altering three key properties: the principal accumulates more rapidly due to code generation at machine speed, the interest is more deeply latent, as debt is often incurred by individuals unable to recognize it, and ownership becomes diffuse, since the originating author is a probabilistic model. Four mechanisms are identified as compounding factors: insecure generation under default conditions (Pearce et al., 2022; Veracode, 2025), human overconfidence, where developers using AI assistants produce less secure code while perceiving it as more secure (Perry et al., 2023), a mismatch between code generation and review rates (DeBellis et al., 2024), and the introduction of new attack surfaces in LLM-integrated systems, such as prompt-to-SQL injections demonstrated by Pedro et al. (2025). From an architectural perspective, Recupito et al. (2024) identify Undeclared Consumers-components that silently use model outputs without access control-as having the highest security impact among the AI-specific anti-patterns examined (Sculley et al., 2015).

2.2 The verification gap

Functionally correct code does not guarantee security, however, the evidence regarding this relationship is mixed, and measurement methodologies vary. Scenario-based prompts reveal high vulnerability rates (Pearce et al., 2022). Conversational models may identify issues when prompted but often still produce non-robust default outputs (Khoury et al., 2023). Repository mining has identified weaknesses across 43 CWE classes in Copilot-generated code that has already been merged into GitHub projects (Fu et al., 2025). Controlled trials, however, do not consistently show an increase in vulnerabilities (Sandoval et al., 2023). Therefore, it is defensible to conclude that AI-generated code is not inherently less secure, but its provenance does not guarantee security, and all generated code should be subject to the same controls as any other.

For small and medium-sized enterprises (SMEs), the critical gap lies between the level of verification required and the available oversight capacity. Security reviews are often conducted by the same developer who accepted the generated code, if they occur at all (Chidukwani et al., 2022). Web applications amplify this risk because customer-facing systems combine commercial significance with direct Internet exposure. Additionally, reliance on frameworks, cloud services, and external contractors obscures responsibility for secure defaults, which may not be present (Devos et al., 2012).

Static application security testing (SAST) represents the most practical initial control in this context. SAST operates on every code change and identifies known source-to-sink patterns, dangerous application programming interfaces (APIs), and hard-coded secrets (GitHub, 2026; SonarSource, 2026a). However, SAST cannot evaluate business-logic authorization or deployment context, and it generates both false positives and false negatives (Aloraini et al., 2019; Charoenwet et al., 2024). Organisational response requires separate evidence from existing review systems or workflows, a scanner run created for research measures detections, not whether an organisation is classifying them elsewhere for analysis.

3 Methodology

Measuring security debt across two time points requires a research design that avoids confounding comparisons with differences between projects. In this study, each repository is analyzed twice, once in a pre-threshold state and once in a post-threshold state, making sure that language, domain, and team remain constant. However, this approach results in nested rather than independent states, which constrains the interpretation of results discussed in Section 4.

3.1 Research design

An empirical quantitative design is employed to examine security technical debt in GitHub repositories owned by small and medium-sized enterprises (SMEs). Measurement is based on SonarQube, which reports security findings and estimates remediation effort using a rule-based quality model (Letouzey & Ilkiewicz, 2012; SonarSource, 2026b). The estimated remediation effort serves as a relative indicator of security technical debt (STD) (Li et al., 2015).

The date 1 January 2023 serves as an operational threshold distinguishing the period before the widespread availability of code assistants from the period after, based on Copilot's general availability in mid-2022 (Pearce et al., 2022) and ChatGPT's public release in late 2022. This threshold functions as a coarse research proxy rather than a definitive treatment indicator. Pre-threshold repositories may already include assisted code, while post-threshold repositories may contain none; thus, misclassification is anticipated in both directions, which biases any period comparison toward a null result. Consequently, all findings are interpreted as associations, and the period comparison is presented as a secondary, exploratory observation.

The study employs a paired design: each selected repository is analyzed twice, once in a pre-threshold state and once in a post-threshold state, reconstructed from its commit history. Each project thus serves as its own control, eliminating between-project variance in language, domain, team composition, and coding conventions from the period comparison. This design necessitates the use of paired statistical tests throughout the analysis.

The two states of each repository are not independent samples; the post-threshold snapshot is a superset of the pre-threshold state. Consequently, any density calculated on the post-threshold state represents an average over a code base in which post-threshold code comprises only a fraction of the total, potentially biasing the period comparison toward a null result regardless of actual changes in code quality. The median repository increased in size from 122.2 to 178.9 KLOC, indicating that post-threshold code accounts for approximately one third of the later snapshot and is weighted accordingly. The aggregate is reported because it reflects the security state that an organisation actually maintains. However, a flat aggregate density does not demonstrate that recently written code is comparable in quality to older code. Commit-level attribution would directly isolate post-threshold code; the lack of such attribution represents the principal design limitation of this study.

This study addresses three research questions:

- **RQ1:** What is the level and density of STD in SME-related GitHub repositories?
- **RQ2:** Which vulnerability categories contribute most to STD?
- **RQ3:** Is there a relationship between project size, project complexity, and STD?

3.2 Repository selection and data preparation

The dataset consists of publicly available GitHub repositories owned by SMEs or SME-like organisations. Inclusion criteria required repositories to be public, contain sufficient project-specific code in a supported language, be compatible with SonarQube, have identifiable ownership, and be assignable to a specific period. Exclusion criteria comprised forks without substantial development, tutorial or demo repositories, archived or inactive projects, repositories mainly comprising third-party libraries, generated or boilerplate-only code, unclear SME connections, and those that could not be analysed for technical reasons.

A repository was classified as SME-related if the owning organisation satisfied at least two proxy criteria: (1) the public profile or linked website describes a commercial product or service, (2) the number of public contributors does not exceed 50, (3) cross-referencing via LinkedIn or a company website indicates fewer than 250 employees, consistent with the EU SME definition (European Commission, 2003), or (4) there are no clear indicators of enterprise-scale activity. These criteria serve as proxy signals rather than verified records. A single coder conducted the classification process according to set decision rules. Ambiguous cases were excluded rather than resolved by subjective judgment. Inter-rater agreement was not evaluated.

A target sample of approximately 60 repositories per group was established, consistent with non-parametric methods commonly used in empirical software engineering (Wohlin et al., 2012). The final sample size was smaller due to the combined requirements of pairing, tool compatibility, and successful analysis; the realised figures are reported in Section 4.1. Duplicate and near-duplicate repositories were removed. Where technically feasible, third-party dependencies, vendor folders, generated and minified files, and framework boilerplate were excluded to ensure the analysis concentrated on project-specific source code.

3.3 Metrics

The primary dependent variable is STD (Li et al., 2015), which is measured as the estimated security remediation effort reported by SonarQube according to its metric definition (Letouzey & Ilkiewicz, 2012; SonarSource, 2026b). Remediation effort is considered a consistent relative metric rather than an exact measure of developer time. To account for the mechanical accumulation of findings in larger repositories, two normalised metrics are also defined:

$$\text{STD Density} = \frac{\text{Security Remediation Effort (min)}}{\text{KLOC}} \tag{1}$$

$$\text{Vulnerability Density} = \frac{\text{Detected Vulnerabilities}}{\text{KLOC}} \tag{2}$$

KLOC refers to thousands of logical lines of code, as measured by SonarQube's `ncloc` metric. Both raw and normalised metrics are reported, as they rank projects differently and address distinct questions. Raw effort quantifies the remediation burden faced by an organisation, whereas density reflects the security quality of the code regardless of its size.

SonarQube assigns a fixed remediation cost to each rule; therefore, the total remediation effort is calculated as a rule-weighted function of the number of findings, rather than as an independently observed metric. As a result, STD density and vulnerability density are inherently related by design and should be interpreted as alternative representations of a single underlying quantity, rather than as independent confirmations of one another.

In addition to the previously defined security metrics, each analysis collects project size (measured in logical lines of code) and five complexity measures: files, functions, classes, cyclomatic complexity, and cognitive complexity. These serve as independent variables for RQ3. Security hotspots and the SonarQube A-E security-review rating characterise the state of each research instance, however, they are not considered indicators of organisational process. The main programming language and period assignment are recorded as descriptive variables, with the programming language also treated as a potential confounder. To ensure analytical transparency, the SonarQube version, the quality profiles for each activated language plugin, the retrieved metric keys, and the fetch date are documented for each analysis.

3.4 Statistical analysis

The analysis commences with descriptive statistics, including repository size, language distribution, vulnerability counts and density, total STD, and STD density. Language is reported descriptively because vulnerability profiles differ across languages (Croft et al., 2022). The distributions show strong right skewness, which motivates the use of non-parametric methods throughout. Shapiro-Wilk test results are provided for completeness but are not emphasized, as the test is underpowered at $n = 28$.

Spearman's rank correlation is used to examine the relationships between size, complexity, and STD (RQ3). Given the paired design, the Wilcoxon signed-rank test is applied for period comparisons, while the unpaired Mann-Whitney U test is reported as a robustness check. The Kruskal-Wallis test is employed when more than two groups are compared, as in the language analysis. The significance level is set at $\alpha = 0.05$. Effect sizes are reported as the matched-pairs rank-biserial correlation r where applicable. For monotonically increasing quantities such as code size, a rank-biserial value near unity indicates that nearly every repository changed in the same direction, which is expected for cumulative code bases and should not be interpreted as a measure of practical significance.

Given the exploratory nature of the study and the presence of multiple comparisons, p-values in the time frame analysis are adjusted using the Benjamini-Hochberg procedure (denoted p_{BH}) and interpreted with caution. The correction family includes the period comparisons reported in Section 4.5: size, absolute remediation effort, vulnerability count, STD density, and vulnerability density. Differences in the distribution of severity levels between periods are assessed using the χ^2 test of independence.

The Spearman correlations for RQ3 are based on 56 analyses from 28 repositories. The two observations from each repository are not independent, as they share language, domain, team, and most of their code. Consequently, correlations computed on the pooled set have effective degrees of freedom below the nominal n, resulting in anti-conservative p-values. The coefficients reported in Section 4.4 represent pooled descriptive values. Their p-values should be interpreted with caution, as independent inferential tests would require a one-state-per-repository or mixed-effects specification.

3.5 Manual inspection protocol

Given that SAST output frequently contains context-dependent false positives (Aloraini et al., 2019; Charoenwet et al., 2024), all findings at the two highest severity levels (CRITICAL and BLOCKER) are manually classified rather than accepted without scrutiny. Each finding is categorized as either production or non-production code based on its file path. Non-production code includes test directories, localisation and translation resources, vendored third-party code, fixtures, and mocks. Severity analysis is reported twice: once for all findings and once restricted to production code. This approach explicitly demonstrates the influence of scanner context on newly observed effects. Rules generating large clusters of findings are examined individually and traced to their source repositories. The assignment rule is strictly mechanical and was applied by a single coder: no inter-rater agreement statistic is reported. Because the rule depends solely on directory names and requires no subjective judgment regarding the findings, the risk of misclassification is considered low, although this has not been empirically verified.

3.6 Validity and limitations

Construct validity. This study cannot confirm whether specific code was AI-generated. The 2023 cut-off serves only as a coarse temporal proxy rather than a treatment indicator, and period assignment relies on commit history. The central construct identified in the motivating literature, namely AI-induced security debt, is not directly measured in this analysis; therefore, none of the subsequent results should be interpreted as estimates of this construct. Static Application Security Testing (SAST) tools are known to produce both false positives and false negatives (Aloraini et al., 2019; Charoenwet et al., 2024). To address this, the SonarQube configuration is fully reported, and all high-severity findings are manually inspected as described in Section 3.5. The exclusive dependence on a single tool constitutes an additional limitation (Charoenwet et al., 2024), but this approach was adopted to maintain analytical consistency and guarantee a comparable remediation-effort metric across all repositories.

Internal validity. The nested-snapshot problem described in Section 3.1 represents the primary threat to internal validity. The programming language functions as a potential confounder, as vulnerability profiles vary across languages (Croft et al., 2022). Language is recorded, reported descriptively, and its effect on both normalised metrics is tested; however, it is not statistically controlled for in the main models, which is acknowledged as a limitation. Additional factors, including developer experience, security practice maturity, and repository activity, may also influence the results and are incorporated into the interpretation of findings. Estimates obtained from 28 matched pairs are imprecise; therefore, a small true difference remains plausible given the data. The period comparison backs the conclusion that no aggregate change was detected, rather than providing evidence of equivalence.

External validity. Public repositories may not accurately represent private SME projects, and SME classification is based on public signals rather than verified data. The sample is mainly composed of PHP projects, with PHP representing 31 out of 56 analyses. The language inclusion criterion was revised during data collection to accommodate this. Consequently, the findings primarily reflect characteristics of SME-maintained PHP web projects rather than SME software development as a whole. Additionally, the Kruskal-Wallis language tests reported in Section 4.3 lack sufficient statistical power to exclude programming language as a confounding factor. Therefore, extrapolation beyond this specific population is not justified.

Reliability. Remediation effort is estimated and utilized as a relative metric. The selection criteria, exclusion logic, key SonarQube configuration, and analysis decisions are reported to enhance transparency. Analysis scripts, repository lists, and complete profile exports are retained internally. As no public replication package is provided and profile version identifiers were not recorded, the study does not assert independent replicability.

4 Results

The results are structured to ensure that the analysis configuration is clearly established prior to any comparative assessment. Aggregate quantities are presented initially, followed by manual inspection, as the raw data and the inspected results may lead to different interpretations. Sections 4.2, 4.3, and 4.4 address RQ1, RQ2, and RQ3, respectively, while Sections 4.5 and 4.6 present the exploratory period comparison and the manual inspection of high-severity findings.

4.1 Configuration and realised sample

All 56 analyses were conducted on a single SonarQube instance (v26.6.0.123539, metrics retrieved 22 July 2026) using the built-in Sonar way profiles across all 27 activated language plugins, including the dedicated secrets profile. Profiles and metric keys were identical for every analysis, so ruleset drift cannot account for any difference reported below. The profile name and server version were recorded, but profile version identifiers were not. Because the contents of the Sonar way profiles change across SonarQube releases, exact replication would require the profile definitions shipped with v26.6.0 rather than the profile name alone.

After applying the criteria in Section 3.2, 28 repositories remained, each analysed in a pre- and post-threshold state, resulting in 28 matched pairs. The realised sample is not language-balanced, and this is its principal external-validity constraint. PHP constitutes the majority (31 of 56 analyses), followed by TypeScript (7), JavaScript (6), Python (5), CSS (3), Java (2), and XML (2); no C# repository met the remaining criteria. The language criterion stated in Section 3.2 was revised during data collection. PHP was included after it became evident that the originally listed languages would not provide a sufficient number of SME-owned repositories that met the pairing and tool-compatibility requirements, and no C# repository qualified. Consequently, the final language composition shows repository availability rather than initial study design. Language is further examined as a potential confounder in Section 4.3.

Table 1. Descriptive statistics by period ($n = 28$ per group). Medians with totals in parentheses; densities are ratios and have no total.

	Pre-2023	**Post-2023**
Size (KLOC)	122.2 (5,267)	178.9 (8,451)
Effort (min)	267.5 (10,551)	275.5 (26,982)
Vulnerabilities	8.5 (577)	10.0 (906)
STD density (min/KLOC)	1.617	1.619
Vuln. density (/KLOC)	0.065	0.062

4.2 RQ1: Level and density of STD

Across all 56 analyses, SonarQube identified 1,483 vulnerabilities and 6,803 security hotspots, corresponding to an estimated 37,533 minutes of remediation effort over 13.7 million lines of code (Table 1). All variables show strong right skewness, and the Shapiro-Wilk test rejected normality in all cases ($W = 0.28$–0.72, $p < 0.001$), supporting the use of non-parametric tests.

The median repository requires 272 minutes of security remediation effort, corresponding to 1.62 minutes per thousand lines of code (min/KLOC) and 0.06 vulnerabilities per KLOC. In absolute terms, this represents approximately four and a half hours of estimated work for a typical project. However, security debt is highly concentrated. Five repositories require no security remediation effort, while the two largest debtors account for 36.7% of the total. Raw effort and density yield different project rankings, which is why both metrics are reported. For example, the densest repository in the sample (90.7 min/KLOC) is a 717-line project with only 65 minutes of debt, making it undetectable in a raw-effort ranking.

This concentration imposes a constraint on all aggregate statistics presented in the paper. When one third of the total debt resides in two of 28 repositories, medians, totals, and pooled correlations may be disproportionately influenced by a small number of high-leverage projects. Consequently, the correlations reported in Section 4.4 should be interpreted descriptively.

Among the 56 analyses, 54 received the lowest security-review rating. The remaining two analyses contained no hotspots and were assigned an A rating by default. All 6,803 hotspots were marked as `TO_REVIEW` on the SonarQube instances created for this study.

These states describe the research instances and must not be interpreted as evidence of organisational review behaviour. Detection and disposition were not measured symmetrically: detections were actively generated by the researchers by scanning repositories on newly created SonarQube instances, whereas dispositions were only passively observed on those same instances. Detection was therefore expected for repositories of this size, while an existing disposition could not appear unless prior review history had been imported or a new review had been performed after the research scan. Organisations may review equivalent findings in another SonarQube deployment, an issue tracker, a pull-request workflow, or another security tool. Consequently, the observed `TO_REVIEW` states neither demonstrate an absence of organisational triage nor permit estimation of an organisational triage rate. No empirical detection-versus-disposition comparison is therefore reported.

Table 2. Rules ranked by contribution to STD (all 56 analyses).

Rule	n	**Effort**	**Share**
S5122 Permissive CORS	51	12,240	32.6%
S5443 Temp file, writable dir	260	7,800	20.8%
S2068 Hard-coded credentials	149	4,470	11.9%
S6418 Hard-coded secrets	98	2,940	7.8%
S5725 Missing integrity check	509	2,545	6.8%
S6865 Service account perms	23	1,380	3.7%
S6268 Sanitization disabled	44	1,320	3.5%
S6358 Android backup unrestricted	8	720	1.9%
All other rules	341	4,118	11.0%
Total	**1,483**	**37,533**	**100%**

4.3 RQ2: Categories contributing most to STD

The findings are concentrated within a small set of rules, though their distribution changes depending on whether they are ranked by count or remediation effort. By count, missing integrity checks on remote artefacts (S5725) are most common, with 509 findings, followed by temporary files in publicly writable directories (S5443, 260), hard-coded credentials (S2068, 149), and hard-coded secrets (S6418, 98). Together, these four rules represent 68.5% of the 1,483 total vulnerability findings. Since RQ2 addresses contribution to security technical debt (STD) rather than finding counts, Table 2 presents rules ranked by remediation effort. This results in a substantially different ordering. S5725, despite being the most frequent rule, contributes only 6.8% of total remediation effort because SonarQube assigns it five minutes per finding. In contrast, permissive CORS configuration (S5122) accounts for only 51 findings, but each requires four hours to address. This issue represents 32.6% of the total measured debt, making it the largest single contributor in the sample. Ranking by count would obscure this impact. Therefore, a remediation queue ordered by finding count versus one ordered by estimated effort would direct a subject matter expert's attention to almost entirely different sets of rules.

Among identified hotspots, weak cryptography is the most prevalent, with (1,925) findings. This is followed by denial of service (1,348), a residual 'others' category (1,326), unencrypted data (999), authentication (511), remote code execution (330), and permissions

(217). Hotspots do not include remediation estimates and are therefore reported solely by count. Fig. 1(a) presents the distribution by period.

Language did not confound the results. Among the four languages with at least four analyses (PHP, TypeScript, JavaScript, Python), the Kruskal-Wallis test found no significant difference in STD density ($H = 0.95$, $p = 0.812$) or vulnerability density ($H = 0.35$, $p = 0.950$). However, these tests are underpowered, and a null result provides only limited evidence. Therefore, while the dominance of PHP in the sample does not solely account for the reported density figures, it cannot be concluded that programming language is irrelevant.

4.4 RQ3: Size, complexity and STD

The result depends on the specific construct of security technical debt (STD) under consideration, as the two constructs produce divergent results.

Larger and more complex projects exhibit substantially greater absolute security technical debt. Project size demonstrates a moderate correlation with remediation effort ($\rho = 0.45$) and vulnerability count ($\rho = 0.49$), both statistically significant ($p < 0.001$, see Fig. 1(b)). All complexity indicators show notable correlations with both outcomes: number of files ($\rho = 0.45$ / 0.44), classes ($\rho = 0.42$ / 0.45), functions ($\rho = 0.37$ / 0.42), cyclomatic complexity ($\rho = 0.33$ / 0.41), and cognitive complexity ($\rho = 0.31$ / 0.44), all $p < 0.05$. These coefficients are pooled and should be interpreted in light of the non-independence discussed in Section 3.4 and the leverage check described in Section 4.2.

After normalisation, the correlations become small, negative, and statistically non-significant: project size versus STD density yields $\rho = -0.16$ ($p = 0.233$), and versus vulnerability density $\rho = -0.14$ ($p = 0.321$). This result indicates a lack of evidence for a positive relationship between size and density, rather than a reversal; two non-significant coefficients of this magnitude do not demonstrate that density decreases with size. The data support the more cautious conclusion that, in this sample, security debt accumulates approximately in proportion to code volume. Furthermore, complexity indicators do not outperform raw size as predictors, suggesting that the total amount of code, rather than its complexity, primarily determines the extent of security debt in these repositories.

4.5 Exploratory period comparison

The comparison presented is secondary and does not imply causality (Section 3.1). The size of the repository increased substantially, with the median repository gaining approximately 46% in lines of code ($r = +0.91$, $p_{BH} < 0.001$), accompanied by parallel growth in files, functions, and complexity. As noted in Section 3.4, near-universal growth is expected for cumulative code bases; therefore, the effect size should not be interpreted as substantively meaningful. In contrast, the security metrics do not exhibit similar trends. Absolute remediation effort ($r = +0.24$, $p_{BH} = 0.343$) and vulnerability count ($r = +0.09$, $p_{BH} = 0.683$) increased only in the median and not significantly, while all normalized metrics remained stable: STD density was essentially unchanged (1.617 vs. 1.619 min/KLOC, $p_{BH} = 0.523$), and vulnerability density was marginally lower ($r = -0.30$, $p_{BH} = 0.255$). The unpaired Mann-Whitney test verified these findings throughout ($p = 0.63$-0.99). Fig. 2 presents the paired trajectories, where the size panel diverges sharply, while both density panels remain level.

4.6 Manual inspection of high-severity findings

The inspection protocol described in Section 3.5 substantially alters the interpretation of the measured period difference.

The proportion of blocker and critical vulnerabilities increases from 13.3% to 47.5% ($\chi^2(3) = 206.5$, $p < 0.001$), representing the largest period difference in the dataset and, if interpreted without further analysis, appearing to strongly support the acceleration hypothesis. However, classification of all 507 high-severity findings by file path indicates that this increase is primarily an artefact: 15.6% of pre-threshold high-severity findings are located outside production code, compared to 78.4% post-threshold. Two sources account for the majority of these findings. Rule S5443 (temporary file in a world-writable directory) generated 260 findings, with 257 located in `__tests__` directories and 257 originating from a single repository, primarily flagging scratch files written to `/tmp` in unit tests. Rule S6418 generated 56 findings in translation files, triggered by words such as `token` and `secret` used as user-interface labels.

When analysis is restricted to production code, the observed shift decreases to 12.2% versus 17.3% (Fig. 3). Although this difference remains detectable ($\chi^2(3) = 22.1$, $p < 0.001$), it represents only about one quarter of the initial apparent magnitude. The difference is no longer statistically significant when tested per repository ($p = 0.079$), and it disappears entirely when the single dominant repository is excluded ($p = 0.143$). The χ^2 test treats findings as independent observations, despite clustering within repositories and rules, which inflates its significance. The per-repository test is therefore more reliable, and it does not reach significance. Some genuine production findings persist, including hard-coded credentials, unverified TLS certificates, and unauthenticated endpoints, indicating that the residual effect is not zero. However, the primary conclusion is that the observed severity shift is mainly a scanner-context artefact. This anal-

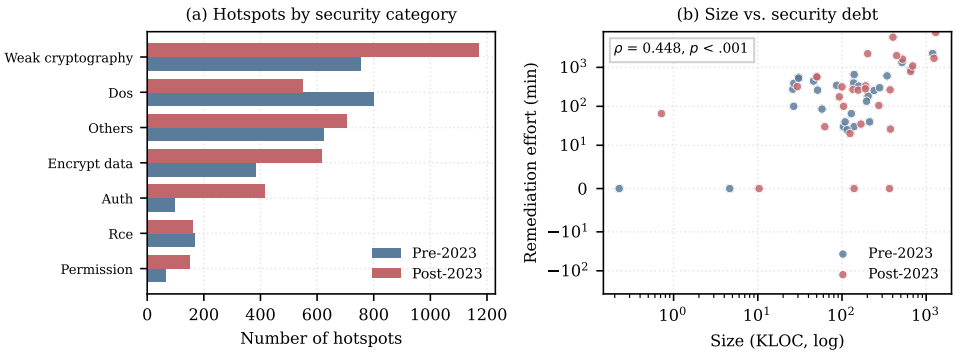


Figure 1. (a) Security hotspots by category and period. (b) Repository size versus estimated remediation effort, with pooled Spearman's ρ.

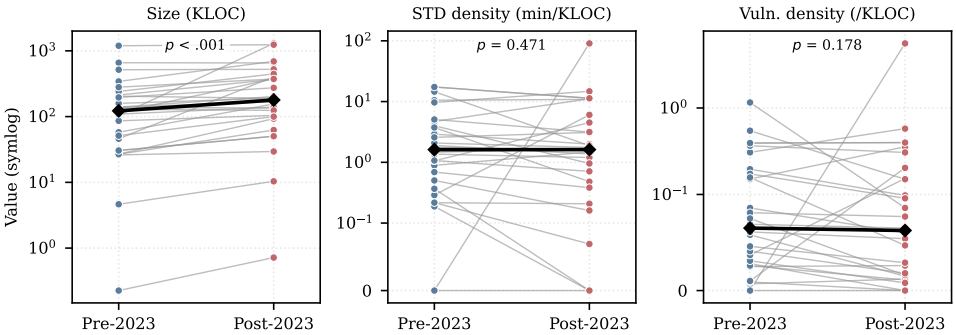


Figure 2. Paired repository trajectories (symmetric-log ordinate). Grey lines connect the two states of each repository; the black line joins group medians.

ysis also illustrates the practical cost of unfiltered scanner output: a subject matter expert acting on the raw findings would have allocated most of the remediation budget to test fixtures and translation strings.

5 Discussion

Taken together, these findings redirect attention from evaluating whether the assistant era has increased the density of security weaknesses to examining how scanner scope affects measurements of technical debt in expanding code bases. The period comparison supports only a limited descriptive claim, whereas the scanner-context finding supports a broader methodological conclusion. The hotspot states observed on the newly created research instances do not support conclusions about organisational triage behaviour.

5.1 Aggregate debt density did not change, the code base grew

The initial hypothesis that development in the assistant era would accelerate the accumulation of security debt is not supported by the data in this sample. Although repositories increased substantially in size and their absolute security debt grew proportionally, the density measures remained stable. An STD density of 1.617 compared to 1.619 min/KLOC represents a result that is effectively null for this type of measurement.

This null result should be considered in the context of the dilution described in Section 3.1. Degradation restricted to new code would appear as only a minor aggregate change. Therefore, the most defensible interpretation is more limited than a general null for AI-assisted code: the density of these repositories, as maintained artefacts, did not change across the threshold, even though their size increased by nearly fifty percent. Three explanations remain consistent with the data: assistance may not degrade shipped code in these contexts, the threshold may be too coarse to detect changes, or any effect in new code may be obscured by aggregation. Therefore, this result should be considered provisional.

This result does not contradict the scenario-based literature, but rather measures a different aspect. While Pearce et al. (2022) and Veracode (2025) assess model outputs in isolation, the present measurement captures the aggregate state of a repository following organisational review, testing, and iteration, with results affected by historical code. The finding corresponds more closely to Sandoval et al. (2023), who also observed no significant increase under controlled conditions, although their study design identifies a treatment, whereas this study does not.

5.2 The artefact problem is the practical finding

The severity analysis presented in Section 4.6 represents the most consequential finding for practitioners and is the least affected by the aforementioned design limitations. A 34-percentage-point increase in high-severity findings could be interpreted as evidence of an AI-driven security crisis. However, manual inspection reduces this effect to a residual level that is not statistically significant when assessed at the repository level and disappears entirely when a single repository is excluded. This phenomenon happens because expanding projects tend to add test suites and localisation files at a faster rate than production code, leading scanners to flag patterns in these directories that are benign in context but would be serious if found elsewhere.

This evidence has methodological implications that go beyond the scope of this study. Temporal comparisons of static application security testing (SAST) output that do not account for the composition of production versus non-production code risk conflating test-suite expansion with actual security degradation. Future research on AI and code security should routinely report severity distributions limited to production code, conduct analyses at the repository level rather than the individual finding level, and examine the rules underlying any substantial cluster of findings prior to drawing conclusions.

5.3 Implications for SME governance

The primary practical implication supported by this study is the need for context-aware interpretation of scanner output. Most post-threshold high-severity findings were located outside production code. Organisations that use raw scanner counts as a remediation queue risk misallocating limited resources. Distinguishing test, vendor, fixture, and localisation paths from production-security reporting can enhance signal quality without fully suppressing these findings.

Remediation priorities should reflect both exposure and estimated remediation effort. Absolute measured security debt was associated with code volume and concentrated in a small subset of repositories. Prioritisation should therefore consider customer exposure and remediation effort rather than distribute attention uniformly or rank rules solely by finding count. Generated code, when present, should be subject to the same controls as other contributions because provenance alone does not guarantee security.

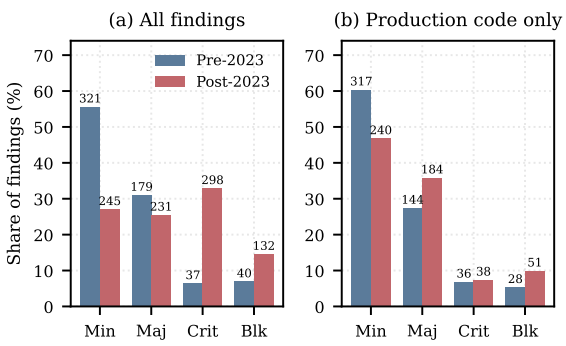


Figure 3. Severity distribution. (a) All findings. (b) Production code only. Bar labels are counts.

The hotspot results support recommendations for future measurement rather than conclusions about SME governance. A study of organisational disposition would require observation of organisations' existing scanner deployments or triangulation of scanner states with issue trackers, pull requests, security tickets, and interviews. Detection and disposition would need to be defined and observed within the same observation window. The newly created SonarQube instances used in this study cannot reveal whether, where, or how the sampled organisations review security findings.

6 Conclusion and future work

Analysis of 28 SME repositories in matched pre- and post-2023 states indicates that aggregate size-normalised security debt remained unchanged. Code bases expanded by a median of 46%, yet both security technical debt (STD) density and vulnerability density were stable. Absolute security debt correlated with code volume rather than code complexity. The identified vulnerabilities primarily reflected common web weaknesses, rather than issues unique to generated code.

The only substantial period difference in the raw data, a rise in high-severity findings from 13.3% to 47.5%, was determined through manual inspection of all 507 of those findings to be primarily a scanner-context artefact, driven by test and localisation directories. This difference disappears when assessed at the repository level or when a single repository is excluded. Therefore, period comparisons of static application security testing (SAST) output that do not control for production-code composition are unreliable.

Subsequent research should directly address the principal limitations of this study. Attributing findings to their introducing commit would replace the diluted aggregate with a measurement focused on newly written code, representing the most impactful improvement to the study design. Repositories with verified labels of AI usage, obtained through commit-level attribution, telemetry from assistant vendors, or alliances with organisations willing to disclose their tooling, would allow replacement of the temporal proxy with a measured treatment variable. Employing a multi-tool design would reduce dependence on a single ruleset and enable estimation of tool-specific false-positive rates in the production versus non-production code dimension identified here. Finally, research on organisational disposition should follow findings through organisations'

existing scanners and work-management systems over a shared observation window. Newly created research instances alone cannot establish whether, where, or how organisations review scanner output.

References

Aloraini, B., Nagappan, M., German, D. M., Hayashi, S., & Higo, Y. (2019). An empirical study of security warnings from static application security testing tools. *Journal of Systems and Software*, *158*, 110427. https://doi.org/10.1016/j.jss.2019.110427

Avgeriou, P., Kruchten, P., Ozkaya, I., & Seaman, C. (2016). *Managing technical debt in software engineering (dagstuhl seminar 16162)*. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. https://doi.org/10.4230/DAGREP.6.4.110

Charoenwet, W., Thongtanunam, P., Pham, V.-T., & Treude, C. (2024). An empirical study of static analysis tools for secure code review. *Proceedings of the 33rd ACM SIGSOFT International Symposium on Software Testing and Analysis*, 691–703. https://doi.org/10.1145/3650212.3680313

Chidukwani, A., Zander, S., & Koutsakis, P. (2022). A survey on the cyber security of small-to-medium businesses: Challenges, research focus and recommendations. *IEEE Access*, *10*, 85701–85719. https://doi.org/10.1109/ACCESS.2022.3197899

Croft, R., Xie, Y., Zahedi, M., Babar, M. A., & Treude, C. (2022). An empirical study of developers' discussions about security challenges of different programming languages. *Empirical Software Engineering*, *27*(1), 27. https://doi.org/10.1007/s10664-021-10054-w

Cunningham, W. (1993). The WyCash portfolio management system. *ACM SIGPLAN OOPS Messenger*, *4*(2), 29–30. https://doi.org/10.1145/157710.157715

DeBellis, D., Storer, K., Lewis, A., Good, B., Villalba, D., Maxwell, E., & Castillo, K. (2024). *Dora accelerate state of devops 2024 report* (tech. rep.). DORA, Google. Retrieved August 18, 2026, from https://dora.dev/research/2024/dora-report/2024-dora-accelerate-state-of-devops-report.pdf

Devos, J., Van Landeghem, H., & Deschoolmeester, D. (2012). Rethinking IT governance for SMEs. *Industrial Management & Data Systems*, *112*(2), 206–223. https://doi.org/10.1108/02635571211204263

European Commission. (2003). Commission recommendation of 6 may 2003 concerning the definition of micro, small and medium-sized enterprises.

Fu, Y., Liang, P., Tahir, A., Li, Z., Shahin, M., Yu, J., & Chen, J. (2025). Security weaknesses of copilot-generated code in GitHub projects: An empirical study. *ACM Transactions on Software Engineering and Methodology*, *34*(8), 1–34. https://doi.org/10.1145/3716848

GitHub. (2026). *Code scanning with CodeQL*. GitHub. Retrieved July 20, 2026, from https://docs.github.com/code-security/code-scanning/introduction-to-code-scanning/about-code-scanning-with-codeql

Khoury, R., Avila, A. R., Brunelle, J., & Camara, B. M. (2023). How secure is code generated by ChatGPT? *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2445–2451. https://doi.org/10.1109/SMC53992.2023.10394237

Letouzey, J.-L., & Ilkiewicz, M. (2012). Managing technical debt with the SQALE method. *IEEE Software*, *29*(6), 44–51. https://doi.org/10.1109/MS.2012.129

Li, Z., Avgeriou, P., & Liang, P. (2015). A systematic mapping study on technical debt and its management. *Journal of Systems and Software*, *101*, 193–220. https://doi.org/10.1016/j.jss.2014.12.027

Martinez, J., Quintano, N., Ruiz, A., Santamaria, I., De Soria, I. M., & Arias, J. (2021). Security debt: Characteristics, product life-cycle integration and items. *2021 IEEE/ACM International Conference on Technical Debt (TechDebt)*, 1–5. https://doi.org/10.1109/TechDebt52882.2021.00009

Pearce, H., Ahmad, B., Tan, B., Dolan-Gavitt, B., & Karri, R. (2022). Asleep at the keyboard? assessing the security of GitHub copilot's code contributions. *2022 IEEE Symposium on Security and Privacy (SP)*, 754–768. https://doi.org/10.1109/SP46214.2022.9833571

Pedro, R., Castro, D., Carreira, P., & Santos, N. (2025, January 27). From prompt injections to SQL injection attacks: How protected is your LLM-integrated web application? https://doi.org/10.48550/arXiv.2308.01990

Perry, N., Srivastava, M., Kumar, D., & Boneh, D. (2023). Do users write more insecure code with AI assistants? *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, 2785–2799. https://doi.org/10.1145/3576915.3623157

Recupito, G., Pecorelli, F., Catolino, G., Lenarduzzi, V., Taibi, D., Di Nucci, D., & Palomba, F. (2024). Technical debt in AI-enabled systems: On the prevalence, severity, impact, and management strategies for code and architecture. *Journal of Systems and Software*, *216*, 112151. https://doi.org/10.1016/j.jss.2024.112151

Rindell, K., Bernsmed, K., & Jaatun, M. G. (2019). Managing security in software: Or: How i learned to stop worrying and manage the security technical debt. *Proceedings of the 14th International Conference on Availability, Reliability and Security*, 1–8. https://doi.org/10.1145/3339252.3340338

Sandoval, G., Pearce, H., Nys, T., Karri, R., Garg, S., & Dolan-Gavitt, B. (2023). Lost at c: A user study on the security implications of large language model

code assistants. *Proceedings of the 32nd USENIX Conference on Security Symposium*.

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. (2015). Hidden technical debt in machine learning systems. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 28). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2015/file/86df7dcfd896fcaf2674f757a2463eba-Paper.pdf

SonarSource. (2026a). *Managing security hotspots*. SonarSource. Retrieved July 20, 2026, from https://docs.sonarsource.com/sonarqube-server/user-guide/security-hotspots

SonarSource. (2026b). Metric definitions: Understanding measures and metrics [Accessed 2026].

Veracode. (2025). *Genai code security report* (tech. rep.).

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., & Wesslén, A. (2012). *Experimentation in software engineering*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-29044-2

Publication Ethics And Malpractice Statement

Central European Conference on Information and Intelligent Systems – CECIIS is committed to upholding the highest standards of publication ethics and takes all possible measures against any publication malpractices for its double-blind peer-reviewed Conference Proceedings.

1. Duties of Editors and Track Directors

Editors are responsible for maintaining the integrity of the academic record, for having processes in place to assure the quality of the material they publish and for precluding business needs from compromising intellectual and ethical standards.

Track Directors are responsible that appropriate reviewers are selected for submissions (i.e. individuals who are able to judge the work and are free from disqualifying competing interests).

Publication decisions

Editors and track directors are responsible for deciding which manuscripts submitted to CECIIS will be accepted for presentation and for publication of the final paper in the Conference Proceedings. This decision is based on the recommendation of reviewers. Main selection criteria are the contribution's importance, originality and clarity, and the study's validity and its relevance. Decisions will not be influenced by the authors' race, gender, sexual orientation, religious belief, ethnic origin, citizenship, or political philosophy. Editors and track directors may be guided by the policies of the CECIIS Program Committee and constrained by such legal requirements as shall then be in force regarding libel, copyright infringement and plagiarism. Editors and track directors may confer with reviewers in making this decision.

Confidentiality

Editors and track directors must not disclose any information about a manuscript submitted to anyone other than the corresponding author, reviewers, potential reviewers, other editorial advisers, and the publisher, as appropriate.

Disclosure and conflicts of interest

Unpublished materials disclosed in a submitted manuscript will not be used in an editors´ or track directors' own research without the express written consent of the author.

2. Duties of Reviewers

Contribution to editorial decisions

Papers will be published in the Conference Proceedings after double-blind peer-reviewed process. The reviewers advise track directors and editors. Track directors communicate with authors, as required, and help them in improving the quality of their manuscript. The reviewers don't know the author's identity and reviewers' comments to track directors and editors are confidential and before passing on to the author will be made anonymous. The names of the reviewers remain strictly confidential; with their identities known only to editors and corresponding track directors.

Promptness

Any selected reviewer who feels unqualified to review the research reported in manuscript or knows that its prompt review will be impossible, should notify the track director and withdraw from the review process.

Confidentiality

Any manuscripts received for review must be treated as confidential documents. They must not be disclosed to or discussed with others except as authorized by editors.

Standards of objectivity

Reviews should be conducted objectively. Personal criticism of the author is inappropriate. Reviewers should express their views clearly with supporting arguments, if necessary with explanation.

Acknowledgement of sources

Reviewers should identify relevant published work that has not been cited by authors. They should point out whether observations or arguments derived from other publications are accompanied by the respective source. Reviewers will notify track directors or editors of any substantial similarity or overlap between the manuscript under consideration and any other published paper of which they have personal knowledge.

Disclosure and conflict of interest

Privileged information or ideas obtained through the review process must be kept confidential and not used for personal advantage. Reviewers should not consider manuscripts in which they have conflicts of interest resulting from competitive, collaborative, or other relationships or connections with any of the authors, companies, or institutions associated with the papers.

3. Duties of Authors

Reporting standards

Authors of manuscripts should present an accurate account of the work performed as well as an objective discussion of its significance. Underlying data should be represented accurately in the manuscript. A manuscript should contain sufficient detail and references to permit others to replicate the work. Fraudulent or knowingly inaccurate statements constitute unethical behavior and are unacceptable.

Data access and retention

Authors could be asked to provide the raw data in connection with a manuscript for review and should be prepared to provide public access to such data, if practicable, and should in any event be prepared to retain such data for a reasonable time after publication.

Originality and plagiarism

Authors should ensure that they have written entirely original works, and if authors have used the work and/or words of others that this has been appropriately cited or quoted. Plagiarism in any form, including the touting of material contained in another paper (of the same authors or some other author) with cosmetic changes as a new paper; copying or paraphrasing substantial parts of another's paper (without attribution), and claiming results from research conducted by others are among the numerous forms of plagiarism. In all its forms plagiarism constitutes unethical publishing behavior and is unacceptable.

Acknowledgement of sources

Proper acknowledgment of the work of others must always be given. Authors should cite publications that have been influential in determining the nature of the reported work.

Multiple, redundant or concurrent publication

An author should not in general publish manuscripts describing essentially the same research in more than one journal or primary publication. Submitting the same manuscript to more than one journal or conference concurrently constitutes unethical publishing behaviour and is unacceptable.

Authorship of the paper

Authorship should be limited to those who have made a significant contribution to the conception, design, execution, or interpretation of the reported study. All those who have made significant contributions should be listed as co-authors. Where there are others who have participated in certain substantive aspects of the manuscript, they should be acknowledged or listed as contributors.

The corresponding author should ensure that all appropriate co-authors and no inappropriate co-authors are included on the manuscript, and that all co-authors have seen and approved the final version of the manuscript and have agreed to its submission for publication.

Fundamental errors in published work

When an author discovers a significant error or inaccuracy in his/her own published work, it is the author’s obligation to promptly notify editors or publishers and cooperate with editors to retract or correct the paper.

If editors or the publisher learns from a third party that a published work contains a significant error, it is the obligation of the author to promptly retract or correct the paper or provide evidence to editors of the correctness of the original paper.

Disclosure and conflicts of interest

All authors should disclose in their manuscript any financial or other substantive conflict of interest that might be construed to influence the results or interpretation of their manuscript. All sources of financial support for the paper should be disclosed.

Publisher's confirmation

In cases of alleged or proven scientific misconduct, fraudulent publication or plagiarism the publisher, in close collaboration with the editors, will take all appropriate measures to clarify the situation and to amend the paper in question. This includes the prompt publication of an erratum or, in the most severe cases, the complete retraction of the affected work.

Author Guidlines

Microsoft Word Template for the Central European Conference on Information and Intelligent Systems (CECIIS)

Author 1, Author 2
Author's Common Affiliation
Department/Institute
Full Address
{author1, author2}@domain.com

Author(s) Name(s)
Author Affiliation(s)
Department/Institute
Full Address(es)
E-mail(s)

Abstract*. The abstract is to be in fully-justified italicized text, at the top of the left-hand column as presented here, below the author information. Use the word "Abstract." in 10-point Times, boldface type, left positioned, initially capitalized, followed by the abstract in 10-point, single-spaced type, up to 100 words long.*
Leave one blank line after the abstract, and then begin the keywords. Use the word "Keywords." in 10-point Times, boldface type, left positioned, initially capitalized, followed by up to ten keywords in 10-point, separated by comma, as below.
Leave two blank lines after the keywords, and then begin the main text.

Keywords. CECIIS, conference paper, template

1 Introduction

This is an example of a Central European Conference on Information and Intelligent Systems (CECIIS) submission. These guidelines include complete description of the paper style including formatting, fonts, spacing, and related information for producing your proceedings manuscripts.

This template can be used to create your own paper but if you have any troubles or if in doubt do not hesitate to contact us via e-mail: ceciisoo@foi.hr or via the conference web site at http://www.ceciis.foi.hr.

2 Structure and Requirements for Research Paper (Examples)

Central European Conference on Information and Intelligent Systems (CECIIS) welcomes research papers on all topics of interest (Computer Games, Data and Knowledge Bases, Digital transformation and corporate culture, Innovation, entrepreneurship and economic growth, Intelligent Energy, Intelligent Systems, Learning analytics, AI and technologies in education, Strategic planning and decision making, Systems for Smart Mobility, Trends in Software Engineering) as well as on the Conference theme/special emphasis announced on yearly base, addressing both, research theory and practice.

The following paper structure examples regarding the research type could be used for preparing a CECIIS paper:

1. Quantitative questionnaire-based: background research (state of the art); research problem; hypothesis or research question; description of instrument used; sampling; statistical methods used for analysis; discussion of results with limitations of research.
2. Qualitative research: background research (state of the art); research problem; research question; qualitative method used (case study research, interviews, content analysis etc.); discussion of results with limitations of research.
3. Algorithms and theoretical research: theory description with references; original theoretical contribution (theorem, algorithm etc.) with proofs or confirmations; future research and conclusion
4. Systematic literature review with meta-analysis: scope of research, research questions or research purpose; sources of literature with period covered (data bases, journals etc.), methods of analysis, conclusions with future research

Research paper should be theoretically and evidence based, clearly and consistently as well as ethically prepared and written. Central European Conference on Information and Intelligent Systems (CECIIS) has zero-tolerance policy on any kind of plagiarism. Paper containing text copied and used from another author(s) work without permission or from authors previously published own work, without referencing, will be retracted.

3 Paper Organization and Formatting

All papers should be written in English, up to 8 pages long (approximately 20 000 characters), and arranged in the following order:

- Main title
- Author(s), affiliation(s), full and e-mail addresses
- Abstract
- Keywords
- Body text (Main text)
- Footnotes
- Acknowledgements
- References

All printed material, including text and figures, must be kept within a print area of 16 cm (6.3") wide by 24.7 (9.7") high. Do not write or print anything outside the print area. Paper size is A4: 21x29.7 cm (8.3x11.7"). Margins (top, bottom, left, right) are 2.5 cm (1"). All *text* must be in a two-column format. Columns are to be 7.7 cm (3") wide, with a 0.6 cm (0.24") space between them. Text must be fully justified. Indent each paragraph by 0.5 cm (0.2").

The final submission has to be submitted in a single PDF file with all fonts that are used embedded, and also the source file (Microsoft Word, LaTeX - zip original files and images).

4 Paper Submission in Croatian Language

Authors who also want to publish the translation of their paper in Croatian, need to upload both the English and Croatian version of the paper (in the same document) for review. English and Croatian versions need to be in parallel columns (side by side). Left column is the paper in English and right column is the paper in Croatian. A camera-ready full paper should be prepared according to the Instructions for Authors and submitted following all the stages of the Paper Submission Process.

5 Main Title

The main title (on the first page) should begin 3,5 cm from the top edge of the page, centred, and in Times 18-point, bold face. Capitalize the first letter of nouns, pronouns, verbs, adjectives, and adverbs; do not capitalize articles, coordinate conjunctions, or prepositions (unless the title begins with such a word). Leave a blank line after the title.

6 Author Name(s) and Affiliation(s)

Author names and affiliations are to be centred beneath the title and printed in Times 10-point type. Author names should be in bold face. Multiple authors shall be grouped by affiliation as shown in the title above. Include also e-mail addresses in 9-point `Courier/Typewriter` face. Follow the author information by two blank lines before main text.

7 Type-style and Fonts

Wherever Times is specified, Times Roman or Times New Roman may be used. If neither is available on your word processor, please use the font closest in appearance to Times that you have access to. Please avoid using bit-mapped fonts if possible. True-Type 1 fonts are preferred.

Also make sure that all fonts in the final PDF file are embedded. Thus we strongly encourage the use of Latex especially if you use non-standard fonts for formulas for your examples. You can make sure that your fonts are embedded by opening the PDF file in some PDF viewing program and taking a look at the properties of the document.

8 First-order Headings (May Extend to the Next Line)

For example, “1 Introduction”, should be Times 14-point boldface, initially capitalized, flush left, with two blank lines before, and one blank line after.

Don't use periods (“.”) after the heading number, use them only in lower order headings to separate them from higher order headings. For *long* headings use *a hanging indent* aligning the text to the right of the heading number as shown above.

All paragraphs in the text should be indented except for the first paragraph in a section.

8.1 Second-order Headings (May Extend to the Next Line)

As in this heading, they should be Times 12-point boldface, initially capitalized, flush left, with one blank line before, and one after. Use a hanging indent for long headings.

8.1.1 Third-order Headings (May Extend to the Next Line)

Third-order headings, as in this paragraph, are discouraged. However, if you must use them, use 10-point Times, boldface, initially capitalized, flush left, with one blank line before, and one after. Use a hanging indent for long headings.

9 Main Text

Type your main text in 10-point Times, single-spaced. Do not use double-spacing. All paragraphs should be indented 0.5 cm (0.2") except for the first paragraph in a section.

Be sure your text is fully justified—that is, flush left and flush right. Please do not place any additional blank lines between paragraphs.

9.1 Figures

All figures are to be included within the text. **Figure captions** are to be *below* the figures, in 10-point Times (or a similar serif font), normal face. Initially capitalize only the first word of each figure caption.

Figures are to be numbered consecutively with Arabic numerals throughout the paper, for example: "Figure 1. Database contexts", and are referred to in the text as Fig. 1, Fig. 2, etc.

9.1.1 Illustrations, Photographs and Graphs

Illustrations, photographs and graphs are considered as figures. All graphics should be centred. Your artwork must be in place in the article (preferably printed as part of the text rather than pasted up). If you are using photographs and are able to have halftones made at a print shop, use a 100- or 110-line screen. Supply the best quality photographs and illustrations possible.

Pencilled lines and very fine lines do not reproduce well. Remember, the quality of the book cannot be better than the originals provided. The conference proceedings will be printed in grey scale, but you can provide images in colour since the papers will also be published on the web site of the conference.

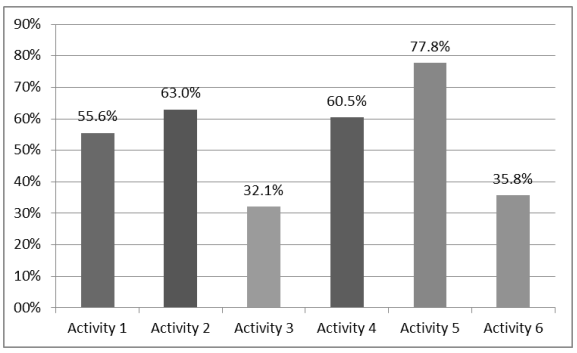


Figure 1. Times, 10 points, normal face

9.2 Tables

All tables are to be included within the text. **Table titles** are to be *above* the tables, in 10-point Times (or a similar serif font), normal face. Initially capitalize only the first word of each table title. Tables are to be numbered consecutively with Arabic numerals throughout the paper, for example: "Table 1. Input data", and are referred to in the text as Table 1, Table 2, etc. Column headings should be as brief as possible.

Table 1. Times, 10 points, normal face

	Column 1	Column 2	Column 3
Row 1	1	2	3
Row 2	4	5	6
Row 3	3	4	7
Row 4	2	6	4
Row 5	2	2	4

9.3 Equations

Equations should be numbered serially on the right-hand side by Arabic numerals in parentheses, and referred to in the text by eq. 1, eq. 2 etc.

$$3x^2 + 3x - 23 = 0 \tag{1}$$

9.4 Footnotes

Use footnotes sparingly (or not at all!) and place them at the bottom of the column on the page on which they are referenced.[1]

10 In-text Citations and Referencing

In-text citation must be used to denote all text used from another author(s) work or from authors previously published own work. When referenced in the text, enclose the author surname(s) and the publication year in brackets, for example (Surname, year) or (Surname, year, p. x) where x is the page number of the cited reference. When referenced paper has more than three authors, cite the paper as (Surname et al., year).

List and number all bibliographical references, alphabetically sorted, in 10-point Times, single-spaced, with a hanging indent, at the end of your paper. Use 6-point paragraph spacing after each reference. References should be distributed evenly in both columns as much as possible, so use a column break where appropriate.

In the References section below you will find examples of common types of bibliographical items: book (Mayer, 2009, p. 52), journal paper (Bule & Peer, 2014), journal paper with more than three authors (Shukor et al., 2014), conference proceedings paper (Steingartner & Novitzká, 2015), report or standard

[1] Use Times 8-point type, single-spaced. Avoid using footnotes altogether and include necessary peripheral observations in the text (within parentheses, if you prefer, as in this sentence).

(*ISO/IEC 25010: Systems and software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - System and software quality models*, 2011) and web page ("Web Content Accessibility Guidelines (WCAG) 2.0," 2008). For other types of bibliographical items refer the **APA style guidelines**. It should be ensured that every reference cited in the text is also listed in the References section (and vice versa).

Acknowledgments

Acknowledgments, if necessary, should appear in a separate paragraph preceding the references.

References

Bule, J., & Peer, P. (2014). Technical, Legal, Economic and Social Aspects of Biometrics for Cloud Computing. *Journal of Information and Organizational Sciences*, 38(2), 83–95.

ISO/IEC 25010: Systems and software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - System and software quality models. (2011).

Mayer, R. E. (2009). *Multimedia Learning* (Second Edi.). New York: Cambridge University Press.

Shukor, N. A., Tasir, Z., Van der Meijden, H., & Harun, J. (2014). A Predictive Model to Evaluate Students' Cognitive Engagement in Online Learning. *Procedia - Social and Behavioral Sciences*, 116, 4844–4853. doi:10.1016/j.sbspro.2014.01.1036

Steingartner, W., & Novitzká, V. (2015). A new approach to operational semantics by categories. In T. Hunjak, V. Kirinić, & M. Konecki (Eds.), *Proceedings of the 26th Central European Conference on Information and Intelligent System (CECIIS 2015)* (pp. 247–254). University of Zagreb, Faculty of Organization and Informatics Varaždin.

Web Content Accessibility Guidelines (WCAG) 2.0. (2008). Retrieved from https://www.w3.org/TR/WCAG20/

Microsoft Word Template for papers in Croatian Language for the Central European Conference on Information and Intelligent Systems

Microsoft Word predložak za radove na hrvatskome jeziku za Central European Conference on Information and Intelligent Systems

Author 1, Author 2
Author's Common Affiliation
Naziv ustanove
Department/Institute
Naziv odjela/instituta
Full Address / Adresa
{author1, author2}@domain.com

Author(s) Name(s)
Author Affiliation(s)
Naziv ustanove
Department/Institute
Naziv odjela/instituta
Full Address / Adresa
E-mail(s)

Abstract. *The abstract is to be in fully-justified italicized text, at the top of the column as presented here, below the author information. Use the word "Abstract." in 10-point Times, boldface type, left positioned, initially capitalized, followed by the abstract in 10-point, single-spaced type, up to 100 words long.*
Leave one blank line after the abstract, and then begin the keywords. Use the word "Keywords." in 10-point Times, boldface type, left positioned, initially capitalized, followed by up to ten keywords in 10-point, separated by comma, as below.
Leave two blank lines after the keywords, and then begin the main text.

Keywords. CECIIS, conference paper

Sažetak. *Tekst sažetka treba biti obostrano poravnat i pisan kurzivom, na vrhu stupca kako je prikazano ovdje ispod podataka o autoru. Upotrijebite riječ "Sažetak." u fontu Times veličine 10, podebljano, poravnato ulijevo, velikim početnim slovom. Nakon riječi "Sažetak." slijedi sažetak rada u fontu veličine 10, jednostruki prored, dužine do 100 riječi.*
Nakon sažetka jedan redak ostaje prazan, a zatim slijede ključne riječi. Upotrijebite riječi "Ključne riječi." u fontu Times veličine 10, podebljano, lijevo poravnanje, s velikim početnim slovom, nakon čega slijedi do deset ključnih riječi u fontu veličine 10, odvojenih zarezom, kao u nastavku.
Ostavite dva prazna retka nakon ključnih riječi, a zatim započnite glavni tekst.

Ključne riječi. CECIIS, konferencijski rad

1 Introduction

This is an example of a Central European Conference on Information and Intelligent Systems (CECIIS) submission. These guidelines include complete description of the paper style including formatting, fonts, spacing, and related information for producing your proceedings manuscripts.

This template can be used to create your own paper but if you have any troubles or if in doubt do

1 Uvod

Ovo je predložak rada za Central European Conference on Information and Intelligent Systems (CECIIS). Smjernice uključuju cjelovit opis stila rada, uključujući oblikovanje, fontove, razmake i srodne informacije za objavu članaka u zborniku.

Predložak možete koristiti za izradu vlastitog rada, a ako imate bilo kakvih problema ili pitanja, ne

not hesitate to contact us via e-mail: ceciis@foi.unizg.hr.

2 Structure and Requirements for Research Paper (Examples)

Central European Conference on Information and Intelligent Systems (CECIIS) welcomes research papers on all topics of interest (Computer Games, Data and Knowledge Bases, Digital transformation and corporate culture, Innovation, entrepreneurship and economic growth, Intelligent Energy, Intelligent Systems, Learning analytics, AI and technologies in education, Strategic planning and decision making, Systems for Smart Mobility, Trends in Software Engineering) as well as on the Conference theme/special emphasis announced on yearly base, addressing both, research theory and practice.

The following paper structure examples could be used for preparing a CECIIS paper:

1. Quantitative questionnaire-based: background research (state of the art); research problem; hypothesis or research question; description of instrument used; sampling; statistical methods used for analysis; discussion of results with limitations of research.
2. Qualitative research: background research (state of the art); research problem; research question; qualitative method used (case study research, interviews, content analysis etc.); discussion of results with limitations of research.
3. Algorithms and theoretical research: theory description with references; original theoretical contribution (theorem, algorithm etc.) with proofs or confirmations; future research and conclusion
4. Systematic literature review with meta-analysis: scope of research, research questions or research purpose; sources of literature with period covered (data bases, journals etc.), methods of analysis, conclusions with future research

Research paper should be theoretically and evidence based, clearly and consistently as well as ethically prepared and written. Central European Conference on Information and Intelligent Systems (CECIIS) has zero-tolerance policy on any kind of plagiarism. Paper containing text copied and used from another author(s) work without permission or from authors previously published own work, without referencing, will be retracted.

3 Paper Organization and Formatting

ustručavajte se kontaktirati nas putem e-pošte: ceciis@foi.unizg.hr.

2 Struktura i zahtjevi za istraživački rad (primjeri)

Central European Conference on Information and Intelligent Systems (CECIIS) očekuje istraživačke radove na teme od interesa (Računalne igre, Baze podataka i baze znanja, Digitalne transformacije i korporativna kultura, Inovacije, poduzetništvo i ekonomski rast, Pametna energija, Inteligentni sustavi, Analitike učenja, umjetna inteligencija i tehnologije u obrazovanju, Strateško planiranje i donošenje odluka, Sustavi za pametnu mobilnost, Trendovi u softverskom inženjerstvu), kao i na godišnju temu konferencije, a bave se kako teorijskim, tako i primijenjivim istraživanjem.

Sljedeći primjeri strukture rada mogli bi se koristiti za pripremu rada za CECIIS:

1. Kvantitativni upitnik: pregled dosadašnjih istraživanja; problem istraživanja; hipoteza ili istraživačko pitanje; opis korištenog instrumenta; uzorkovanje; statističke metode korištene za analizu; rasprava o rezultatima s ograničenjima istraživanja.
2. Kvalitativno istraživanje: pregled dosadašnjih istraživanja; problem istraživanja; istraživačko pitanje; korištena kvalitativna metoda (istraživanje slučaja, intervjui, analiza sadržaja itd.); rasprava o rezultatima s ograničenjima istraživanja.
3. Algoritmi i teorijsko istraživanje: opis teorije s referencama; izvorni teorijski doprinos (teorem, algoritam, itd.) s dokazima ili potvrdama; buduće istraživanje i zaključak
4. Sustavni pregled literature s metaanalizom: opseg istraživanja, istraživačka pitanja ili svrha istraživanja; izvori literature s obuhvaćenim razdobljem (baze podataka, časopisi itd.), metode analize, zaključci o budućim istraživanjima

Istraživački rad trebao bi biti teorijski i zasnovan na dokazima, jasno i dosljedno te etički pripremljen i napisan. Central European Conference on Information and Intelligent Systems (CECIIS) ima politiku nulte tolerancije prema bilo kojoj vrsti plagijarizma. Rad koji sadrži tekst kopiran i korišten iz djela drugog autora bez odobrenja ili iz prethodno objavljenog vlastitog djela, bez referenciranja, bit će odbačen.

3 Organizacija i formatiranje rada

Svi radovi trebaju biti predani na engleskom jeziku, dužine do 8 stranica (oko 20 000 znakova) i organizirani prema sljedećem redoslijedu:

All papers should be written in English, up to 8 pages long (approximately 20 000 characters), and arranged in the following order:

- Main title
- Author(s), affiliation(s), full and e-mail addresses
- Abstract
- Keywords
- Body text (Main text)
- Footnotes
- Acknowledgements
- References

All printed material, including text and figures, must be kept within a print area of 16 cm (6.3") wide by 24.7 (9.7") high. Do not write or print anything outside the print area. Paper size is A4: 21x29.7 cm (8.3x11.7"). Margins (top, bottom, left, right) are 2.5 cm (1"). All *text* must be in a two-column format. Columns are to be 7.7 cm (3") wide, with a 0.6 cm (0.24") space between them. Text must be fully justified. Indent each paragraph by 0.5 cm (0.2").

The final submission has to be submitted in a single PDF file with all fonts that are used embedded, and also the source file (Microsoft Word, LaTeX).

4 Paper Submission in Croatian Language

Authors who also want to publish the translation of their paper in Croatian, need to upload both the English and Croatian version of the paper (in the same document) for review. English and Croatian versions need to be in parallel columns (side by side). Left column is the paper in English and right column is the paper in Croatian. A camera-ready full paper should be prepared according to the Instructions for Authors and submitted following all the stages of the Paper Submission Process.

5 Main Title

The main title should begin 3,5 cm from the top edge of the page, centred, and in Times 18-point, bold face. Capitalize the first letter of nouns, pronouns, verbs, adjectives, and adverbs; do not capitalize articles, coordinate conjunctions, or prepositions (unless the title begins with such a word). Leave a blank line after the title.

6 Author Name(s) and Affiliation(s)

- Naslov rada
- Autor(i), pripadnost(i), adrese i e-mail adrese
- Sažetak
- Ključne riječi
- Tekst rada
- Fusnote
- Zahvale
- Reference

Sav materijal, uključujući tekst i slike, mora biti u području ispisa od 16 cm (6,3") širine i 24,7 (9,7") visine. Ne pišite i ne ispisujte ništa izvan područja ispisa. Veličina papira je A4: 21x29,7 cm (8,3x11,7"). Margine (gore, dolje, lijevo, desno) su 2,5 cm (1"). Sav tekst mora biti u formatu s dva stupca. Stupci trebaju biti široki 7,7 cm (3"), a između njih moraju biti razmaci od 0,6 cm (0,24"). Tekst mora biti obostrano poravnat. Uvucite svaki odlomak za 0,5 cm (0,2").

Konačni rad mora se poslati u jednoj PDF datoteci s ugrađenim svim fontovima koji se koriste, kao i izvornoj datoteci (Microsoft Word, LaTeX).

4 Predaja rada na hrvatskome jeziku

Autori koji žele objaviti i prijevod svog rada na hrvatskome jeziku, trebaju predati i englesku i hrvatsku verziju rada (u istom dokumentu) na recenziju. Engleska i hrvatska verzija trebaju biti u paralelnim stupcima (jedan pored drugog). Lijevi stupac je rad na engleskom jeziku, a desni stupac je rad na hrvatskome jeziku. Finalnu verziju cjelovitog rada treba pripremiti u skladu s Uputama za autore i predati slijedeći sve faze procesa predaje radova.

5 Naslov rada

Naslov rada (na prvoj stranici) trebao bi započeti 3,5 cm od gornjeg ruba stranice, centrirano, u fontu Times veličine 18, podebljano. Naslov započinje velikim početnim slovom. Ostavite jedan prazan redak nakon naslova.

6 Imena autora i pripadnost(i)

Imena autora i pripadnost trebaju biti centrirani ispod naslova i pisani fontom Times veličine 10. Imena autora trebaju biti podebljana. Više autora bit će grupirano prema pripadnosti kao što je prikazano u ovom predlošku. Uključite i adrese e-pošte u fontu `Courier/Typewriter` veličine 9. Nakon podataka

Author names and affiliations are to be centred beneath the title and printed in Times 10-point type. Author names should be in bold. Multiple authors shall be grouped by affiliation as shown in the title above. Include also e-mail addresses in 9-point `Courier/Typewriter` face. Follow the author information by two blank lines before main text.

7 Type-style and Fonts

Wherever Times is specified, Times Roman or Times New Roman may be used. Please avoid using bit-mapped fonts if possible. True-Type 1 fonts are preferred.

Also make sure that all fonts in the final PDF file are embedded. We strongly encourage the use of LaTeX, especially if you use non-standard fonts for formulas for your examples.

8 First-order Headings (May Extend to the Next Line)

For example, “1 Introduction”, should be Times 14-point boldface, initially capitalized, flush left, with two blank lines before, and one blank line after.

Don't use periods (“.”) after the heading number, use them only in lower order headings to separate them from higher order headings. For *long* headings use *a hanging indent* aligning the text to the right of the heading number as shown above.

All paragraphs in the text should be indented except for the first paragraph in a section.

8.1 Second-order Headings (May Extend to the Next Line)

As in this heading, they should be Times 12-point boldface, initially capitalized, flush left, with one blank line before, and one after. Use a hanging indent for long headings.

8.1.1 Third-order Headings (May Extend to the Next Line)

Third-order headings, as in this paragraph, are discouraged. However, if you must use them, use 10-point Times, boldface, initially capitalized, flush left, with one blank line before, and one after. Use a hanging indent for long headings.

9 Main Text

Type your main text in 10-point Times, single-spaced. Do not use double-spacing. All paragraphs should be indented 0.5 cm (0.2") except for the first paragraph in a section.

o autoru slijede dva prazna retka prije nego počne glavni tekst.

7 Stil i fontovi

Gdje god je naveden font Times, može se koristiti Times Roman ili Times New Roman. Molimo izbjegavajte upotrebu bitmapiranih fontova ako je moguće. Poželjni su fontovi True-Type 1.

Također provjerite jesu li ugrađeni svi fontovi u konačnu PDF datoteku. Pritom, snažno preporučamo upotrebu LaTeX-a, posebno ako za svoje primjere koristite nestandardne fontove za formule.

8 Naslovi prvog reda (mogu se nastaviti u sljedeći redak)

Na primjer, "1 Uvod" treba biti font Times veličine 14, podebljano, s velikim početnim slovom, poravnato ulijevo, s dva prazna retka prije i jednim praznim redom nakon naslova.
Ne upotrebljavajte točke (".") iza broja naslova, koristite ih samo u naslovima nižeg reda kako biste ih odvojili od naslova višeg reda. Za duge naslove upotrijebite “hanging indent” poravnavajući tekst udesno od broja naslova, kao što je prikazano gore.

Svi odlomci u tekstu trebaju biti uvučeni, osim prvog odlomka u odjeljku.

8.1 Naslovi drugog reda (mogu se nastaviti u sljedeći redak)

Kao i u ovom naslovu, trebali bi biti font Times veličine 12, s velikim početnim slovom, poravnati lijevo, s jednim praznim retkom prije i jednim nakon. Koristite “hanging indent” za duge naslove.

8.1.1 Naslovi trećeg reda (mogu se nastaviti u sljedeći redak)

Naslovi trećeg reda, kao u ovom paragrafu, ne preporučuju se. Međutim, ako ih morate koristiti, upotrijebite font Times veličine 10, podebljano, s velikim početnim slovom, poravnati ulijevo, s jednim praznim retkom prije i jednim nakon. Koristite “hanging indent” za duge naslove.

9 Tekst rada

Tekst rada pišite fontom Times veličine 10, s jednostrukim proredom. Ne koristite dvostruki prored. Svi odlomci trebaju biti uvučeni 0,5 cm (0,2"), osim prvog odlomka u odjeljku.

Be sure your text is fully justified—that is, flush left and flush right. Please do not place any additional blank lines between paragraphs.

9.1 Figures

All figures are to be included within the text. **Figure captions** are to be *below* the figures, in 10-point Times (or a similar serif font), normal face. Initially capitalize only the first word of each figure caption.

Figures are to be numbered consecutively with Arabic numerals throughout the paper, for example: “Figure 1. Database contexts”, and are referred to in the text as Fig. 1, Fig. 2, etc.

9.1.1 Illustrations, Photographs and Graphs

Illustrations, photographs and graphs are considered as figures. All graphics should be centred. Your artwork must be in place in the article (preferably printed as part of the text rather than pasted up). Supply the best quality photographs and illustrations possible.

Very fine lines do not reproduce well. The conference proceedings will be printed in grey scale, but you can provide images in colour since the papers will also be published on the conference web site.

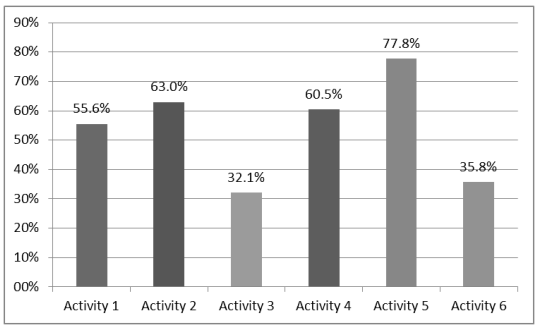


Figure 1. Times, 10 points, normal face

9.2 Tables

All tables are to be included within the text. Table titles are to be *above* the tables, in 10-point Times, normal face. Initially capitalize only the first word of each table title. Tables are to be numbered consecutively with Arabic numerals throughout the paper, for example: “Table 1. Input data”, and is referred to in the text as Table 1.

Table 1. Times, 10 points, normal face

	Column 1	Column 2	Column 3
Row 1	1	2	3
Row 2	4	5	6
Row 3	3	4	7
Row 4	2	6	4
Row 5	2	2	4

Provjerite je li vaš tekst poravnat obostrano - to jest, i lijevo i desno. Molimo ne stavljajte dodatne prazne retke između odlomaka.

9.1 Slike

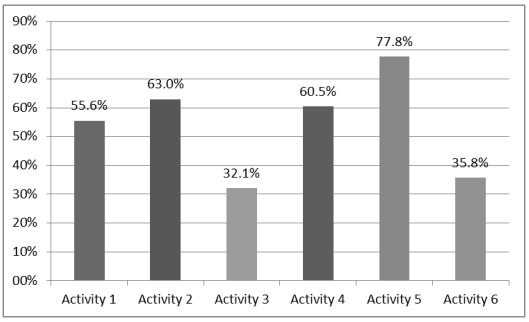
Sve slike trebaju biti uključene u tekst. Nazivi slika trebaju biti ispod slika, u fontu Times veličine 10 (ili sličnom fontu). Samo prvu riječ svakog naziva slike pišite velikim početnim slovom.

Slike se numerira uzastopno arapskim brojevima u cijelom radu, na primjer: "Slika 1. Konteksti baze podataka", a u tekstu se navodi kao Slika 1.

9.1.1 Ilustracije, fotografije i grafovi

Ilustracije, fotografije i grafikoni smatraju se slikama. Sva grafika treba biti centrirana. Svaka slika ima svoje mjesto u članku. Osigurajte najbolju moguću kvalitetu fotografija i ilustracija.

Vrlo tanke linije ne reproduciraju se dobro. Zbornik radova s konferencije bit će tiskan u nijansama sive boje, ali možete osigurati slike u boji jer će radovi biti objavljeni i na mrežnoj stranici konferencije.



Slika 1. Times, veličina 10

9.2 Tablice

Sve tablice trebaju biti uključene u tekst. Naslovi tablica trebaju biti iznad tablica, pisane fontom Times veličine 10. Velikim početnim slovom pišite samo prvu riječ svakog naslova tablice. Tablice se u radu redom numeriraju arapskim brojevima, na primjer: "Tablica 1. Ulazni podaci", a u tekstu se naziva Tablica 1.

Tablica 1. Times, 10 points, normal face

	Column 1	Column 2	Column 3
Row 1	1	2	3
Row 2	4	5	6
Row 3	3	4	7
Row 4	2	6	4
Row 5	2	2	4

9.3 Equations

Equations should be numbered serially on the right-hand side by Arabic numerals in parentheses, and referred to in the text by eq. 1, eq. 2 etc.

$$3x^2 + 3x - 23 = 0 \quad (1)$$

9.4 Footnotes

Use footnotes sparingly (or not at all!) and place them at the bottom of the column on the page on which they are referenced.[1]

10 In-text Citations and Referencing

In-text citation must be used to denote all text used from another author(s) work or from authors previously published own work. When referenced in the text, enclose the author surname(s) and the publication year in brackets, for example (Surname, year) or (Surname, year, p. x) where x is the page number of the cited reference. When referenced paper has more than three authors, cite the paper as (Surname et al., year).

List and number all bibliographical references, alphabetically sorted, in 10-point Times, single-spaced, with a hanging indent, at the end of your paper. Use 6-point paragraph spacing after each reference. References should be distributed evenly in both columns as much as possible, so use a column break where appropriate.

In the References section below you will find examples of common types of bibliographical items: book (Mayer, 2009, p. 52), journal paper (Bule & Peer, 2014), journal paper with more than three authors (Shukor et al., 2014), conference proceedings paper (Steingartner & Novitzká, 2015), report or standard (*ISO/IEC 25010: Systems and software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - System and software quality models*, 2011) and web page ("Web Content Accessibility Guidelines (WCAG) 2.0," 2008). For other types of bibliographical items refer the **APA style guidelines**. It should be ensured that every reference cited in the text is also listed in the References section (and vice versa).

Acknowledgments

9.3 Jednadžbe

Jednadžbe se numeriraju slijedno s desne strane arapskim brojevima u zagradama, a u tekstu se na njih poziva tekstom Jednadžba 1, Jednadžba 2 itd.

$$3x^2 + 3x - 23 = 0 \quad (1)$$

9.4 Fusnote

Koristite fusnote štedljivo (ili ih uopće ne koristite!) i stavite ih na dno stupca na stranici na kojoj su navedeni.[2]

10 Navodi u tekstu i reference

Citiranje u tekstu mora se koristiti za označavanje cijelog teksta korištenog iz djela drugog autora ili prethodno objavljenih vlastitih djela. Kad se citira u tekstu, prezimena autora i godinu izdanja priložite u zagrade, na primjer (Prezime, godina) ili (Prezime, godina, str. X) gdje je X broj stranice citirane reference. Kada referentni rad ima više od tri autora, navedite ga kao (Prezime i sur., Godina).

Navedite sve bibliografske reference, poredane po abecedi, u fontu Times veličine 10, s jednostrukim proredom, s "hanging indent", na kraju vašeg rada. Upotrijebite razmak paragrafa veličine 6 nakon svake reference. Reference bi trebale biti raspoređene u oba stupca što ravnomjernije, stoga upotrijebite prijelom stupca tamo gdje je to prikladno.

U odjeljku Reference u nastavku naći ćete primjere uobičajenih vrsta bibliografskih jedinica: knjiga (Mayer, 2009., str. 52.), rad u časopisu (Bule & Peer, 2014.), članak s više od tri autora (Shukor i sur., 2014), članak s konferencije (Steingartner i Novitzká, 2015), izvještaj ili standard (*ISO/IEC 25010: Systems and software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - System and software quality models, 2011*) i mrežna stranica ("Web Content Accessibility Guidelines (WCAG) 2.0," 2008). Za ostale vrste bibliografskih jedinica pogledajte **smjernice za stil APA**. Potrebno je osigurati da svaka referenca navedena u tekstu bude navedena i u odjeljku Reference (i obrnuto).

Zahvale

[1] Use Times 8-point type, single-spaced. Avoid using footnotes altogether and include necessary peripheral observations in the text (within parentheses, if you prefer, as in this sentence).

[2] Koristite font Times veličine 8, s jednostrukim proredom. Izbjegavajte potpuno koristiti fusnote i u tekst uvrstite potrebna periferna zapažanja (u zagrade, ako želite, kao u ovoj rečenici).

Acknowledgments, if necessary, should appear in a separate paragraph preceding the references.

References

Bule, J., & Peer, P. (2014). Technical, Legal, Economic and Social Aspects of Biometrics for Cloud Computing. *Journal of Information and Organizational Sciences*, 38(2), 83–95.

ISO/IEC 25010: Systems and software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - System and software quality models. (2011).

Mayer, R. E. (2009). *Multimedia Learning* (Second Edi.). New York: Cambridge University Press.

Shukor, N. A., Tasir, Z., Van der Meijden, H., & Harun, J. (2014). A Predictive Model to Evaluate Students' Cognitive Engagement in Online Learning. *Procedia - Social and Behavioral Sciences*, 116, 4844–4853. doi:10.1016/j.sbspro.2014.01.1036

Steingartner, W., & Novitzká, V. (2015). A new approach to operational semantics by categories. In T. Hunjak, V. Kirinić, & M. Konecki (Eds.), *Proceedings of the 26th Central European Conference on Information and Intelligent System (CECIIS 2015)* (pp. 247–254). University of Zagreb, Faculty of Organization and Informatics Varaždin.

Web Content Accessibility Guidelines (WCAG) 2.0. (2008). Retrieved from https://www.w3.org/TR/WCAG20/

Zahvale, ako ih ima trebaju biti u zasebnom odlomku koji prethodi referencama.

Reference

Bule, J., & Peer, P. (2014). Technical, Legal, Economic and Social Aspects of Biometrics for Cloud Computing. *Journal of Information and Organizational Sciences*, 38(2), 83–95.

ISO/IEC 25010: Systems and software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - System and software quality models. (2011).

Mayer, R. E. (2009). *Multimedia Learning* (Second Edi.). New York: Cambridge University Press.

Shukor, N. A., Tasir, Z., Van der Meijden, H., & Harun, J. (2014). A Predictive Model to Evaluate Students' Cognitive Engagement in Online Learning. *Procedia - Social and Behavioral Sciences*, 116, 4844–4853. doi:10.1016/j.sbspro.2014.01.1036

Steingartner, W., & Novitzká, V. (2015). A new approach to operational semantics by categories. In T. Hunjak, V. Kirinić, & M. Konecki (Eds.), *Proceedings of the 26th Central European Conference on Information and Intelligent System (CECIIS 2015)* (pp. 247–254). University of Zagreb, Faculty of Organization and Informatics Varaždin.

Web Content Accessibility Guidelines (WCAG) 2.0. (2008). Retrieved from https://www.w3.org/TR/WCAG20/

CECIIS 2026 Sponsors

CECIIS 2026 GOLD SPONSOR

INFODOM

CECIIS 2026 GOLD SPONSOR

ORACLE®

CECIIS 2026 GOLD SPONSOR

JANAF

CECIIS 2026 SILVER SPONSOR

HBOR

HRVATSKA BANKA
ZA OBNOVU I RAZVITAK

CECIIS 2026 BRONZE SPONSORS

VARAŽDINSKA ŽUPANIJA

Varaždin County

VARAŽDIN

City of Varaždin

SICK

Sensor Intelligence.

CECIIS 2026 Program Committee

Nina Begičević Ređep, Program Committee Chair, University of Zagreb, Croatia

Victoria Appatova, University of Cincinnati, USA
Igor Balaban, University of Zagreb, Croatia
Marko Bohanec, Jozef Stefan Institute, Slovenia
Leo Budin, University of Zagreb, Croatia
Aruna Chandra, Indiana State University, USA
Violeta Cvetkoska, University in Skopje, Macedonia
Luis de Marcos Ortega, University of Alcalá, Spain
Blaženka Divjak, University of Zagreb, Croatia
Jasminka Dobša, University of Zagreb, Croatia
Juraj Dud'ak, Slovak University of Technology in Bratislava, Slovakia
Matjaž Garms, Jozef Stefan Institute, Slovenia
Gabriel Gaspar, Slovak University of Technology in Bratislava, Slovakia
Gordan Gledec, University of Zagreb, Croatia
Andrina Granić, University of Split, Croatia
Marjan Heričko, University of Maribor, Slovenia
Dieter Hertweck, Herman Hollerith Research Centre, Germany
Aleksandar Marković, University of Belgrade, Serbia
Peter Márton, University of Žilina, Slovakia
Marta Mas-Machuca, UIC Barcelona, Spain
Oliver Moravčík, Slovak University of Technology, Slovakia
Enrique Mu, Carlow University, USA
Karlo Nyarko, University of Osijek, Croatia
Jan Paralič, Technical University of Košice, Slovakia
Elisabeth Pergler, CAMPUS 02 - University of Applied Sciences, Austria
Krzysztof Podlaski, University of Lodz, Poland
Patrizia Poščić, University of Rijeka, Croatia
Sonja Ristić, University of Novi Sad, Serbia
Markus Schatten, University of Zagreb, Croatia
Zlatko Stapić, University of Zagreb, Croatia
William Steingartner, Technical University of Kosice, Slovakia
Vjeran Strahonja, University of Zagreb, Croatia
Marko Tadić, University of Zagreb, Croatia

Honorary Members - Former Program Committee Chairs

Tihomir Hunjak, University of Zagreb, Croatia
Mirko Maleković, University of Zagreb, Croatia
Boris Aurer, University of Zagreb, Croatia
Prof. Wolf Rauch, University of Graz, Austria
Prof. Günter Müller

CECIIS 2026 Organizing Committee

Bogdan Okreša Đurić, Organizing Committee Chair, University of Zagreb, Croatia

Nikolina Dreven, University of Zagreb, Croatia
Ivona Kocijan, University of Zagreb, Croatia
Bernarda Kos, University of Zagreb, Croatia
Snježana Križanić, University of Zagreb, Croatia
Marija Kuštelega, University of Zagreb, Croatia
Lea Masnec, University of Zagreb, Croatia
Marko Matus, University of Zagreb, Croatia
Karlo Nadoveza, University of Zagreb, Croatia
Karlo Pepelnik, University of Zagreb, Croatia
Dijana Peras, University of Zagreb, Croatia
Klara Šilec, University of Zagreb, Croatia

CECIIS 2026 Research Tracks' Chairs

Computer Games: Markus Schatten, University of Zagreb, Croatia; Bogdan Okreša Đurić, University of Zagreb, Croatia ; Jaume Jordán Prunera, University Politècnica de València, Spain

Cyber-Physical Systems and Intelligent Energy: Ivan Magdalenić, University of Zagreb, Croatia; Marinko Barukčić, University of Osijek, Croatia; Danijel Topić, University of Osijek, Croatia

Data and Knowledge Bases: Sandra Lovrenčić, University of Zagreb, Croatia; Patrizia Poščić, University of Rijeka, Croatia

Digital Transformation and Corporate Culture: Katarina Tomičić-Pupek, University of Zagreb, Croatia; Gabriel Gaspar, Slovak University of Technology in Bratislava, Slovakia; Peter Márton, University of Žilina, Slovakia

Innovation, Entrepreneurship and Economic Growth: Vladimir Kovšca**,** University of Zagreb, Croatia; Martina Tomičić Furjan, University of Zagreb, Croatia

Intelligent Systems: Jasminka Dobša, University of Zagreb, Croatia; Dunja Mladenić, Jožef Stefan Institute, Slovenia

Learning Analytics, AI and Technologies in Education: Blaženka Divjak, University of Zagreb, Croatia; Igor Balaban, University of Zagreb, Croatia; Marjan Krašna, University of Maribor, Slovenia

Strategic Planning and Decision Making: Nina Begičević Ređep, University of Zagreb, Croatia; Nikola Kadoić, University of Zagreb, Croatia

The Next Wave of Intelligent Systems: Neven Vrček, University of Zagreb, Croatia; Nina Begičević Ređep, University of Zagreb, Croatia

Trends in Software Engineering: Marko Mijač, University of Zagreb, Croatia; William Steingartner, Technical University of Kosice, Slovakia; Luis de Marcos Ortega, University of Alcalá, Spain

CECIIS 2026 Student Poster and Presentation Section Chairs

Violeta Vidaček Hainš, University of Zagreb, Croatia

Victoria Appatova, University of Cincinnati, USA

Peter Márton, University of Žilina, Slovakia