

ARTIFICIAL INTELLIGENCE CONFERENCE

BOOK OF ABSTRACTS

BELGRADE, OCTOBER 8-9, 2026

PROGRAM COMMITTEE

Co-chairs:

Vladan Devedžić, SASA, Belgrade and Faculty of Organizational Sciences, University of Belgrade
Miodrag Mihaljević, SASA, Belgrade and Mathematical Institute, SASA, Belgrade
Dušan Teodorović, SASA, Belgrade

Members:

Aleksandar Kostić, SASA, Belgrade
Angelo Sifaleras, University of Macedonia, Greece
Boško Nikolić, School of Electrical Engineering, University of Belgrade
Basant Agarwal, Central University of Rajasthan, Ajmer, India
Danijela Milošević, Faculty of Technical Sciences in Čačak, University of Kragujevac
Dipak Kumar Jana, Gangarampur College, Dakshin Dinajpur, West Bengal, India
Goce Trajčevski, Iowa State University, Ames, USA
Igor Đurović, Montenegro Academy of Sciences and Arts, Podgorica, Montenegro
Jelena Jovanović, Faculty of Organizational Sciences, University of Belgrade
Jesus Medina, University of Cadiz, Cadiz, Spain
Jianxun Cui, Harbin Institute of Technology, Harbin, China
Leonid Stoimenov, Faculty of Electronic Engineering, University of Niš
Marija Mitrović Dankulov, The Institute of Physics Belgrade
Mart Laanpere, Tallinn University, Tallinn, Estonia
Mirjana Ivanović, Faculty of Sciences, University of Novi Sad
Mirko Dinulović, Faculty of Mechanical Engineering, University of Belgrade
Predrag Janjić, Faculty of Mathematics, University of Belgrade
Ranka Stanković, Faculty of Mining and Geology, University of Belgrade
Shivani Kharra, Stantec, Houston, Texas, United States

Slobodan Vukosavić, Faculty of Electrical Engineering, University of Belgrade and SASA, Belgrade
Sonja Dimitrijević, "Mihajlo Pupin" Institute, Belgrade
Tüü-Szabó Boldizsár, University of Győr, Hungary
Vladimir Srdanović, Mathematical Institute SASA, Belgrade
Weiqin Chen, Oslo Metropolitan University, Oslo, Norway
Zoran Ognjanović, Mathematical Institute SASA, Belgrade

ORGANIZING COMMITTEE

Co-chairs:

Andreja Tepavčević, Mathematical Institute SASA, Belgrade and Faculty of Sciences, University of Novi Sad
Tatjana Davidović, Mathematical Institute SASA, Belgrade

Members:

Ana Anokić, Academy of Technical and Art Applied Studies, Belgrade
Anastazia Žunić, Mathematical Institute SASA, Belgrade
Anđelka Zečević, Mathematical Institute SASA, Belgrade
Biljana Stojanović, Mathematical Institute SASA, Belgrade
Božo Vuković, Mathematical Institute SASA, Belgrade
Dejan Vukelić, Mathematical Institute SASA, Belgrade
Dragan Jočić, Mathematical Institute SASA, Belgrade
Đorđe Jovanović, Mathematical Institute SASA, Belgrade
Luka Matijević, Mathematical Institute SASA, Belgrade
Matea Lukić, Mathematical Institute SASA, Belgrade
Nataša Stević, SASA, Belgrade
Petar Vranić, Mathematical Institute SASA, Belgrade
Raka Jovanović, Mathematical Institute SASA, Belgrade
Slobodan Jelić, Faculty of Civil Engineering, University of Belgrade
Tatjana Jakšić Kruger, Mathematical Institute SASA, Belgrade
Una Stanković, Mathematical Institute SASA, Belgrade
Vanja Stepanović, Faculty of Agriculture, University of Belgrade
Željko Petrović, The Faculty of Media and Communications, Singidunum University

ORGANIZERS

Department of Technical Sciences SASA, Belgrade
Mathematical Institute SASA, Belgrade
Academy Committee for Artificial Intelligence, SASA, Belgrade



FOREWARD

We are pleased to announce the fourth edition of the Artificial Intelligence conference. It represents the aspirations of numerous researchers in Serbia working in the field of artificial intelligence to establish a gathering with a primarily research-oriented focus, under the auspices of the Serbian Academy of Sciences and Arts (SASA), Serbia's most prestigious scientific institution.

The conference aims to bring together researchers from various branches of artificial intelligence from Serbia and abroad to showcase their work and findings. Artificial intelligence is not only a highly popular field today but also a vast one. Consequently, the topics of interest at this conference are numerous and diverse.

In addition to emphasizing its research character, this conference is distinguished by several key features:

- A dedicated focus exclusively on the field of artificial intelligence
- Encouragement of discussions on the fundamental aspects of artificial intelligence
- Special attention to emerging topics in artificial intelligence
- Integration of techniques and results from different subfields of artificial intelligence
- Gathering a significant number of Serbian researchers in the field, including both experienced and emerging talents

We have received a total of 72 submissions to the conference from leading universities and institutions in Serbia and abroad, covering a wide range of topics in artificial intelligence. Out of these, 47 were selected for presentation. Each abstract underwent review by at least two Reviewers or members of the Program Committee, with some

requiring a second revision before acceptance. The conference program consists of two keynote lectures and 10 sessions divided into 2 parallel streams containing the presentations of accepted contributed talks.

Artificial intelligence, with its profound influence on modern society, necessitates ongoing study and a systematic analytical approach. This conference was conceived with precisely that objective – to facilitate the exchange of research ideas and experiences within this vast and complex field.

The conference organizers are particularly committed to fostering opportunities for young and promising researchers. Through presentations and discussions, these individuals will have the chance to share their work, gain valuable feedback from esteemed experts, and build connections with future collaborators. We believe that fostering this interaction between generations is essential for the sustained advancement of artificial intelligence research in Serbia.

The organizers of this conference edition, representing three institutions – the Serbian Academy of Sciences and Arts, the Mathematical Institute of the Serbian Academy of Sciences and Arts, and the Academy Committee for Artificial Intelligence of SASA – firmly believe in the importance of artificial intelligence for the future of various activities in our country. We are confident that research and innovation in this field will contribute to economic development, improvements in health and education, and the opening of new horizons in science, art, and other areas of life. We are convinced that through joint efforts, we can influence the trajectory of this significant field while advancing knowledge, creativity, and innovation.

Sincerely,

Conference Co-chairs

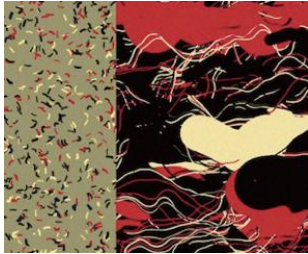
Vladan Devedžić, SASA, Belgrade, and Faculty of Organizational Sciences, University of Belgrade

Miodrag Mihaljević, SASA, Belgrade, and Mathematical Institute, SASA, Belgrade

Dušan Teodorović, SASA, Belgrade

Belgrade, October 2026

Legend:



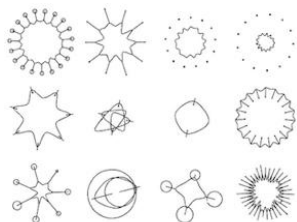
ARTIFICIAL INTELLIGENCE, ETHICS AND SOCIETY



COMPUTER VISION



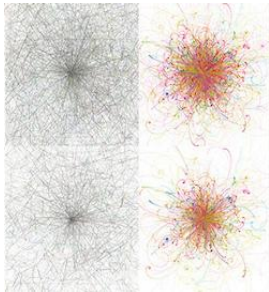
DECISION MAKING AND PROBLEM SOLVING



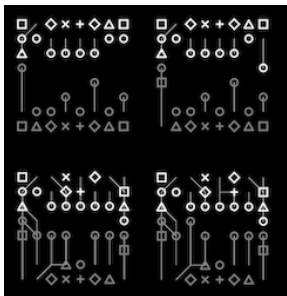
EXPLAINABLE AI



GENERATIVE AI



KNOWLEDGE REPRESENTATION, REASONING AND PLANNING



MACHINE LEARNING



NATURAL LANGUAGE PROCESSING

KEYNOTE TALKS

The Impact of Duplicate Data on Federated Learning: An Overlooked Issue in Data Quality and Privacy-Preserving Governance

Speaker:

Wei Shao, Associate Research Professor, Shandong Computer Science Center (National Supercomputing Jinan Center), Qilu University of Technology (Shandong Academy of Sciences), China

E-mail: shaow@sdas.org

Abstract. Federated learning enables collaborative model training across diverse clients without directly sharing raw data, and has become an important paradigm for privacy-preserving data collaboration in real-world applications. However, most existing research focuses on model optimization and secure aggregation, while the quality of data before training has received much less attention.

In real-world scenarios, cross-client data redundancy and overlaps are prevalent due to cross-platform synchronization or repeated user registration. Such data redundancy is far from a negligible preprocessing issue and can induce systematic drawbacks to federated learning pipelines. This talk will first discuss how duplicate data affects federated learning in terms of training efficiency and model effectiveness, and then further present our series of privacy-preserving decentralized deduplication solutions for different privacy and efficiency requirements. These methods enable secure cross-client redundancy detection while preserving data locality, and provide a practical basis for privacy-aware data quality governance before training, thereby enhancing the performance of federated learning.

Biography: Dr. Wei Shao currently works as an associate research professor at Shandong Computer Science Center (National Supercomputer Center in Jinan), Qilu University of Technology (Shandong



Academy of Sciences) in China. She leads and participates in multiple national and provincial research projects including the National Key R&D Programs, and the National Natural Science Foundation projects. She has published 20+ high-impact papers in top international journals and conferences, and possesses 10+ national invention patents. Her research focuses on applied cryptography, blockchain security, privacy-preserving digital identity and trustworthy federated learning. She has proposed multiple innovative schemes including user-controlled anonymous digital identities, auditable revocable permissioned blockchain mechanisms, and blockchain-based privacy-preserving federated learning frameworks..

Synergies Between Machine Learning and Metaheuristics: New Frontiers in Combinatorial Optimization

Speaker:

Raka Jovanović, Mathematical Institute of the Serbian Academy of Sciences and Arts

E-mail: rjovanovic@mi.sanu.ac.rs

Abstract: Combinatorial optimization problems are central to many scientific and engineering disciplines, yet their NP-hard nature often renders exact solution methods impractical. Metaheuristics have long provided effective approximate solutions, while recent advances in machine learning have opened new opportunities to enhance and transform optimization methodologies.

In this talk, I will explore the synergies between these two fields. Machine learning can guide metaheuristic search by learning problem structures and predicting promising solution regions, while metaheuristics can optimize hyperparameters and neural architectures in machine learning pipelines. Through case studies from my research — including graph neural networks for combinatorial problems, predictive models for constructive heuristics, and hybrid approaches in electric vehicle charging, smart agriculture, and logistics — I will illustrate how these synergies can lead to more efficient and adaptable optimization frameworks. I will also discuss open challenges and future research directions at the intersection of learning and optimization.

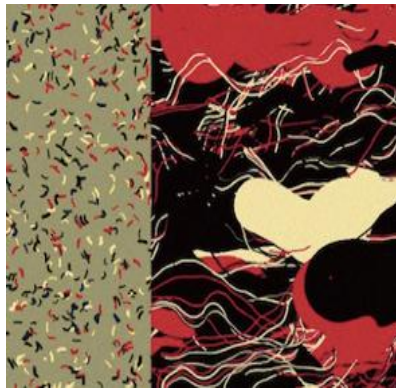
Biography: Dr. Raka Jovanović (born 1978, Belgrade, Serbia) has recently joined the Mathematical Institute of the Serbian Academy of Sciences and Arts as a Senior Research Associate. He completed his undergraduate, Master's (2007), and Doctoral (2012) studies at the



Faculty of Mathematics, University of Belgrade, under the supervision of Prof. Dr. Milan Tuba. His career includes positions at the Institute of Physics, University of Belgrade (2007-2022) and as a researcher at Texas A&M University at Qatar (2009-2011). Since 2013, he has been based in Qatar, at the Qatar Environment and Energy Research Institute (QEERI), part of Hamad Bin Khalifa University in Doha, Qatar where he teaches at the postgraduate level and leads research projects. His primary research directions include the development of novel metaheuristic methods for hard combinatorial optimization problems, mathematical modeling and optimization of real-world systems in energy and transportation, and the application of machine learning techniques to complex predictive and decision-making tasks. He has authored over 100 scientific papers published in prestigious international journals and conference proceedings. He is a recipient of best paper awards, serves as a regular reviewer for top international journals, and maintains extensive collaborations with institutions in Germany, the United Kingdom, the Netherlands, and others.

CONTRIBUTED TALKS

ARTIFICIAL INTELLIGENCE, ETHICS AND SOCIETY



The Dual Role of AI in Human Societies

Ljiljana Brankovic¹ and Olga Timcenko²

¹ School of Science and Technology, University of New England, Australia, lbrankov@une.edu.au

² OTConsulting, Denmark, olgatimcenko@proton.me

Artificial Intelligence (AI) is moving into all areas of our lives and promises to make our existence longer, healthier, easier, and more enjoyable. We already rely on AI to manage our schedules, take notes in meetings, write reports, give us easier access to information, predict future needs, solve hundreds-year-old mathematical problems such as the Navier-Stokes equation, and even cure us by creating personalized cancer vaccines, to name a few. Importantly, AI promises to take on much of our current tasks, so we have more time for our families, hobbies, physical activities, and leisure time.

At the same time, AI is beginning to show its negative side. What once sounded like a dystopian movie is now starting to feel much closer to home. It is easy to imagine a future with a high unemployment rate where people are not expected to make any cognitive effort so they are slowly losing their skills and cognitive abilities, anxiety-inducing constant surveillance, life-and-death decisions made by AI, and profound differences between those with substantial access to AI time and those without it.

We focus on cognitive offloading and attrition of creativity and critical thinking, erosion of autonomy, misplaced responsibilities, and coping in an AI-driven world.

Cognitive offloading and attrition of creativity and critical thinking. Outsourcing cognitive tasks to AI can, over time, lead to a decline in our cognitive skills, including critical thinking and creativity. A 2025 study by Michael Gerlich (Gerlich, M. 2025) on 666 UK residents with diverse age and educational backgrounds shows a strong correlation between the frequent use of AI tools and cognitive offloading, and a strong negative correlation between the frequent use of AI tools and critical thinking score, as well as cognitive offloading and critical thinking score. It was also noted that younger respondents were more likely to frequently use AI tools, and showed higher cognitive offloading and lower critical thinking scores than older participants.

Erosion of autonomy. The attrition of critical thinking and creativity can lead to dependence on AI for everyday tasks and a decline in our capacity to make decisions and stay in charge. This refers both to individuals and to government departments, companies, and organisations. Examples of decisions made by AI are many: from selecting the next song to play or next movie to watch, matching profiles on a dating app, to medical or legal advice. Krook (2025) argues that human autonomy will continue to erode, together with the capacity for critical thinking, decision-making, and social care, and that “the biggest threat to humanity is therefore not that machines will become more like humans, but that humans will become more like machines”.

Misplaced responsibilities. Delegating responsibility for making consequential decisions to AI not only impacts human autonomy but also raises another important question: who is responsible when things go wrong? In the traditional setting, it is always a human who takes responsibility, rather than a tool. In the case of AI, the situation is becoming less clear as both tools and actions are increasingly a collaborative effort of humans and AI. Kang et al. (2025) identify AI related accountability collapse as one of the four main harms of total delegation to AI. A recent study on users’ perceptions on accountability in AI-assisted work () reveals that 66% of respondents believed that humans working on the task are responsible if

there is an error, 19% believed that responsibility is shared between humans and AI, 11% believed that the responsibility is with AI and 3% thought that the developers of AI are responsible.

Coping in an AI-driven world. Coping refers to strategies to manage various stressors and adapt to a dynamic environment. While AI can assist coping by facilitating mental health tracking, therapeutic conversations, and guiding reflection, it can also have a detrimental effect by simplifying complex emotions, bringing about dependence on technology, and creating additional stress by hyper-monitoring (Chirayath et al., 2025).

In this paper, we explore the dual role of AI in future human societies and suggest ways to navigate this challenge. In its own right, AI is neither good nor bad – its goodness is determined by the way we use and govern it.

Keywords: trust in AI, cognitive offloading, attrition of critical thinking, erosion of autonomy, coping

References:

- Biswas, M., Omogbai Oluwaseun, A., and Murray, J. (2025). My Fault, Not the Machine's: A Survey on User Accountability in the Age of AI-Assisted Work, Springer Nature. (In Press)
- Chirayath, G., Premamalini, K, and Joseph, J. (2025). Cognitive offloading or cognitive overload? How AI alters the mental architecture of coping. *Front. Psychol.* **16**:1699320. doi: 10.3389/fpsyg.2025.1699320
- Gerlich, M. (2025). AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking. *Societies*, *15*(1), 6. <https://doi.org/10.3390/soc15010006>
- Kang, Y., and Wei, Z. (2026). When Delegating Everything to AI Becomes Devastating: Harms, Boundaries, and Human Responsibility (July 03, 2026). Available at SSRN: <https://ssrn.com/abstract=7088338> or <http://dx.doi.org/10.2139/ssrn.7088338>
- Krook, J. (2025). When autonomy breaks: the hidden existential risk of AI. *AI & Soc* **40**, 6011–6024 <https://doi.org/10.1007/s00146-025-02397-5>

Comparing AI Incident Databases: Challenges in Classification and Interoperability

*Maja Ilic*¹

¹ Master student at the Faculty of Security Studies, University of Belgrade, Serbia,
majai2307@gmail.com

The rapid development and widespread adoption of artificial intelligence (AI) systems have introduced significant opportunities for innovation across numerous sectors, while simultaneously creating new categories of risk, failures, and harmful consequences. As AI systems become increasingly integrated into decision-making processes that affect individuals, organizations, and society, the systematic identification, classification, and analysis of AI-related incidents have become essential for understanding their potential impacts and improving governance mechanisms. In this context, databases of AI incidents and corresponding taxonomies represent important tools for documenting real-world failures, monitoring emerging risks, and supporting the development of regulatory and risk management frameworks.

This topic presents a comparative analysis of major international AI incident databases and their associated taxonomies, with a particular focus on the classification of risks, harms, and incidents arising from the deployment of AI systems. The study examines the four prominent initiatives: MIT AI Risk Repository, the AI Incident Database (AIID), the OECD AI Incident and Hazards Monitor (AIM), and the AI, Algorithmic and Automation Incidents and Controversies (AIAAIC) repository. For each database, the research analyzes its purpose, methodology for collecting and organizing data, incident reporting criteria, and the taxonomic structures used to categorize risks and harms. Particular attention is devoted to understanding how these initiatives conceptualize AI incidents, distinguish between different forms of harm, and support the systematic analysis of failures associated with artificial intelligence technologies.

The analysis demonstrates that although all examined databases share the objective of improving understanding of AI-related risks, they differ significantly in their methodological approaches, definitions, and classification frameworks. The MIT AI Risk Repository focuses on mapping and categorizing AI risks through general and domain-specific taxonomies, emphasizing dimensions such as intent, responsible entities, timing of occurrence, and risk domains. AIID concentrates on documenting real-world incidents and near-miss events while utilizing detailed harm taxonomies that distinguish between tangible and intangible harms. The OECD AIM framework emphasizes the identification of AI incidents and hazards through large-scale monitoring of global news sources and provides a structured reporting framework designed to support decision-making and governance. In contrast, AIAAIC adopts a user- and victim-centered perspective, focusing on the external consequences of AI systems and highlighting ethical, social, legal, and organizational dimensions of harm.

The findings indicate that no single taxonomy provides a complete representation of all dimensions associated with AI incidents. Instead, each database offers a distinct analytical perspective that contributes to a broader understanding of AI risks. The comparative analysis highlights the strengths and limitations of existing classification approaches and demonstrates the importance of combining technical, ethical, social, and governance-oriented perspectives when evaluating AI-related incidents. Furthermore, the study emphasizes the need for greater standardization and interoperability among incident reporting frameworks in order to improve comparability, facilitate knowledge sharing, and support evidence-based policymaking.

It can be concluded that AI incident databases and taxonomies play a crucial role in advancing the responsible development and deployment of artificial intelligence systems. By enabling the systematic documentation and analysis of failures, harms, and emerging risks, these initiatives contribute to improved transparency, accountability, and risk management practices. As AI technologies continue to evolve and expand into critical domains, the development of comprehensive and harmonized classification frameworks will become increasingly important for ensuring the safe, ethical, and trustworthy use of artificial intelligence.

Keywords: AI incident database, AI incidents, classification, taxonomies

References:

- AI Incident Database. (n.d.). *Editor's Guide*. <https://www.aiaaic.org/aiaaic-repository/classifications-and-definitions/consequence-taxonomy>
- AIAAIC. (n.d.). *Consequence taxonomy*. AIAAIC Repository. <https://www.aiaaic.org/aiaaic-repository/classifications-and-definitions/consequence-taxonomy>
- OECD – Organization for Economic Co-operation and Development. (2025). *Towards a common reporting framework for AI incidents*. OECD Publishing. https://www.oecd.org/content/dam/oecd/en/publications/reports/2025/02/towards-a-common-reporting-framework-for-ai-incidents_8c488fdb/f326d4ac-en.pdf
- Slattery, P., Saeri, A. K., Grundy, E. A. C., Graham, J., Noetel, M., Uuk, R., Dao, J., Pour, S., Casper, S., & Thompson, N. (2024). *The AI risk repository: A comprehensive meta-review, database, and taxonomy of risks from artificial intelligence* [Preprint]. arXiv. <https://arxiv.org/abs/2408.12622>

Patient Privacy in Healthcare AI Partnerships: The DeepMind Case Revisited

Sara Kabić¹, Jovana Dedeić², Mladen Nikolić³, Sana Stojanović Đurđević⁴

¹ Faculty of Mathematics, University of Belgrade, Serbia, mi21168@matf.bg.ac.rs

² Department of Fundamental Disciplines in Engineering, Faculty of Technical Sciences, University of Novi Sad, Serbia, radenovicj@uns.ac.rs

³ Department of Computer Science and Informatics, Faculty of Mathematics, University of Belgrade, Serbia, nikolic@matf.bg.ac.rs

⁴ Department of Computer Science and Informatics, Faculty of Mathematics, University of Belgrade, Serbia, sana@matf.bg.ac.rs

The integration of artificial intelligence into healthcare has opened significant possibilities for improving medical diagnosis and patient care. On the other hand, it has also introduced serious challenges in terms of patient privacy, informed consent, and the accountability of those who collect and use health data (Moulaei et al., 2025). These challenges have become particularly visible in collaborations between public health institutions and private technology companies, resulting in partnerships that are increasingly common, however, insufficiently regulated.

This paper examines the 2015–2017 data-sharing agreement between Google DeepMind and the Royal Free NHS Trust, a case study illustrating how such partnerships can fail patients and raise privacy and ethical concerns. Under this agreement, the medical records of approximately 1.6 million patients, originally intended for developing the Streams clinical application, were transferred to DeepMind without explicit consent of their users (Powles & Hodson, 2017).

The case covers three interconnected problems that remain relevant today. Namely, it analyses the absence of meaningful informed consent, the secondary use of data collected for an entirely different purpose, and a lack of transparency (Moulaei et al., 2025; Powles & Hodson, 2017).

The analysis is conducted within two key regulatory frameworks with ethical matters in mind. Under the General Data Protection Regulation, the agreement constitutes a clear violation of the principles of purpose limitation, data minimization, and transparency. Under the EU AI Act, healthcare AI systems are classified as high-risk, requiring mandatory Data Protection Impact Assessments, human oversight mechanisms, and transparency obligations (Benis et al., 2025). Although neither GDPR nor the EU AI Act was applicable at the time, applying them retroactively reveals the structural vulnerabilities the DeepMind agreement exposed and the gaps that these frameworks were later designed to prevent (Benis et al., 2025).

To address the privacy shortcomings illustrated by the DeepMind case, particularly the transfer of raw patient records outside their originating institution, the paper examines two complementary privacy-preserving approaches to healthcare AI development: federated learning and differential privacy. Federated learning enables collaborative AI model training without transferring raw patient data outside the institutions where it was collected. Instead, each participating institution trains the model locally on its own data and shares only the resulting parameter updates with a central aggregator, which combines them into a global model without ever accessing the underlying records (Rauniar et al., 2024). However, federated learning alone does not provide a formal privacy guarantee, as shared model updates can still expose individual patient records through reverse engineering attacks, such as model inversion, which exploits intercepted gradients to reconstruct training data approximately (Rauniar et al., 2024). Differential privacy can complement this approach by formally bounding the influence of any individual

record on the resulting model or output typically by clipping and adding calibrated noise to gradients during local client training. This approach is known as differentially private stochastic gradient descent (DP-SGD) and is governed by a privacy budget (ϵ) that quantifies the trade-off between privacy protection and model accuracy (Adnan et al., 2022). Without guaranteeing regulatory compliance individually, these approaches together offer a concrete technical pathway for reducing the need to centralize sensitive healthcare data, supporting the data minimization principle under GDPR and the technical risk-management measures required under the EU AI Act.

Keywords: informed consent; secondary use of data; federated learning; differential privacy; EU AI Act

References:

- Adnan, M., Kalra, S., Cresswell, J. C., Rangrej, S. B., Taylor, G. W., & Tizhoosh, H. R. (2022). Federated learning and differential privacy for medical image analysis. *Scientific Reports*, 12(1), 1953. <https://doi.org/10.1038/s41598-022-05539-7>
- Benis, A., Sapir-Mor, N., & Barkan, R. (2025). Medicine, healthcare and the AI Act: gaps, challenges and future implications. *European Heart Journal – Digital Health*, 6(4), 833. <https://academic.oup.com/ehjdh/article/6/4/833/8118685>
- Moulaei, K., Zare, M., Karbasi, Z., Sheikhtaheri, A., & Langarizadeh, M. (2025). Patient consent for the secondary use of health data in artificial intelligence (AI) models: A scoping review. *International Journal of Medical Informatics*, 198. <https://pubmed.ncbi.nlm.nih.gov/40107041/>
- Powles, J., & Hodson, H. (2017). Google DeepMind and healthcare in an age of algorithms. *Health and Technology*, 7(4), 351–367. <https://doi.org/10.1007/s12553-017-0179-1>
- Rauniyar, A., Hagos, D. H., Jha, D., Håkegård, J. E., Bagci, U., Rawat, D. B., & Vlassov, V. (2024). Federated learning for medical applications: A taxonomy, current trends, challenges, and future research directions. *IEEE Internet of Things Journal*, 11(5), 7374–7398. <https://doi.org/10.1109/JIOT.2023.3329061>

Green AI and Circular Economy Adoption in SMEs: Bias, Data Infrastructure, and Structural Constraints in the EU and Serbia

Jovan Kojić¹

¹ Faculty of Entrepreneurial Business and Real Estate, University „Union – Nikola Tesla“, Belgrade, Serbia, jkojictmf@gmail.com; jkojic@unt.edu.rs, ORCID: 0009-0004-5801-1199

This paper examines the relationship between artificial intelligence (AI) automation, resource efficiency, and circular economy practices in small and medium-sized enterprises (SMEs) in the European Union and Serbia, with a focus on structural and contextual constraints affecting AI-driven sustainability outcomes. In this context, Green AI is defined as the deployment of artificial intelligence systems designed to minimize environmental impact through resource efficiency optimization and sustainability-oriented outcomes. While often framed as an objective tool for improving environmental performance, its effectiveness depends on data quality, organizational capacity, and regulatory conditions.

A key source of bias in Circular Economy applications of AI is geographic imbalance in data collection, as these systems are frequently trained on historical datasets and may systematically favor regions and industries with established digital infrastructures. As a result, machine learning models for waste management, logistics, and material recovery are predominantly developed using data from highly industrialized economies. Economic bias further complicates the implementation of Green AI. Optimization functions often prioritize high-value materials (e.g., industrial metals) while marginalizing lower-value but environmentally critical streams like organic waste or complex composites.

To empirically assess these biases, this study employs the Mann–Whitney U test to compare model performance metrics between developed economies and transitional regions, specifically the EU and Serbia. As a non-parametric alternative to the independent samples t-test, this method is suitable for comparing two independent groups without assuming a normal distribution of the performance measures. Data were collected from a sample of $n = 262$ SME owners and senior managers via a structured questionnaire. Initial findings suggest a statistically significant degradation in model accuracy when applied to developing contexts, where infrastructure conditions and waste stream compositions diverge from training data distributions.

To further examine the measurement reliability of the composite indicators used to train and validate these AI models, the research uses Cronbach's alpha (α) formula (1), a measure of internal consistency:

$$\alpha = \frac{k}{k-1} \left(1 - \frac{\sum_{i=1}^k \sigma_i^2}{\sigma_T^2} \right) \quad (1)$$

where k represents the number of items, σ_i^2 denotes the variance of individual item i , and σ_T^2 denotes the total variance of the observed composite score across all $n = 262$ respondents. In this context, α is used to verify the consistency of multi-variable environmental datasets. The findings suggest that low internal consistency in input data acts as a catalyst for bias propagation, leading to unreliable sustainability assessments.

Regression analysis identifies firm size and data accessibility as primary predictors of successful AI-driven recommendation implementation. The model yields an R^2 of 0.42, indicating modest but meaningful explanatory power. AI automation demonstrates a positive and moderate effect on circular economy implementation. This modest explanatory power points to a systemic exclusion of SMEs, which

often lack the standardized data structures required by these models. When organizational resilience and regulatory environment are incorporated as predictors, the model's explanatory power increases substantially, suggesting that AI functions as a contributory rather than sufficient driver of sustainable transformation.

Data availability and quality constraints, particularly in less digitized environments, reduce model reliability and introduce systematic bias. Systems developed using data from highly industrialized regions demonstrate measurable performance degradation in transitional economies, a finding consistent with the $R^2 = 0.42$ result, which suggests that contextual variables not captured in the baseline model account for a substantial portion of unexplained variance.

Overall, Green AI contributes to resource efficiency and circular economy adoption under conditions of adequate data infrastructure, organizational capacity, and regulatory alignment. The regression findings indicate that effective implementation is contingent on these structural prerequisites rather than technological deployment alone. For SMEs operating in transitional economies such as Serbia, this conditionality is particularly significant, as institutional and data infrastructure gaps systematically constrain AI-driven sustainability results. Future research should address the development of context-sensitive AI models regulated for low-digitization environments, and policy frameworks should prioritize data infrastructure investment as a precondition for reasonable Green AI adoption across the EU and candidate member states.

Keywords: Circular economy, artificial intelligence, green intelligence, regression analysis, Mann-Whitney U test, sustainability.

References

- Geissdoerfer, M., Savaget, P., Bocken, N. M. P., & Hultink, E. J. (2017). The circular economy: A new sustainability paradigm? *Journal of Cleaner Production*, 143, 757–768. <https://doi.org/10.1016/j.jclepro.2016.12.048>
- 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>
- Kurniawan, T. A., Liang, X., O'Callaghan, E., Goh, H., Othman, M. H. D., Avtar, R., & Kusworo, T. D. (2022). Transformation of solid waste management in China: Moving towards sustainability through digitalization-based circular economy. *Sustainability*, 14(4), 2374.
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6), 1–35. <https://doi.org/10.1145/3457607>
- Tseng, M. L., Tran, T. P. T., Ha, H. M., Bui, T. D., & Lim, M. K. (2021). Sustainable industrial and operation engineering trends and challenges toward Industry 4.0: A data-driven analysis. *Journal of Industrial and Production Engineering*, 38(8), 581–598. <https://doi.org/10.1080/21681015.2021.1950227>

Transformative Challenges of AI-based Technologies on Education

Vladimir Pavlović¹, Bogdan Novaković², Olivera J. Bogunović³

¹ University of Belgrade, Faculty of Agriculture, Serbia, vlaver@agrif.bg.ac.rs

² University of Milano-Bicocca, Department of Human Sciences for Education "Riccardo Massa," Italy, b.novakovic@campus.unimib.it

³ Harvard University, Harvard Medical School, USA, obogunovic@mclean.harvard.edu

AI-based technologies are reshaping education at every level, from classroom instruction and assessment to curriculum design, institutional governance, and lifelong learning. This transformation is not simply technological; it is pedagogical, ethical, and political. An educational system once structured around fixed curricula, standardized pacing, and uniform assessment is now being retransformed by adaptive AI technologies capable of personalizing learning pathways and processing vast amounts of educational data. This shift matters because education is not only about knowledge transfer; it is also about social development, identity formation, and democratic participation. Therefor UNESCO emphasizes that AI has significant potential to improve teaching and learning, but also warns that its rapid development has outpaced policy and regulation, making inclusion and equity central concerns.

The aim of our research is to analyze transformative challenges of AI-based technologies on education, since these technologies can infer patterns, generate responses, and adapt in real time, thus influencing personalized learning, teaching and assessment, equity and access, ethics and governance, as well as human cognition in general.

The convergence of cognitive decline, neuroplasticity adaptation, and emotional intelligence impairment in the digital age presents a critical challenge for AI based educational systems, but also an opportunity for transformative intervention through intentional design of learning environments that balance technological integration with cognitive protection. There is clear evidence that the current trajectory of digital-first education, characterized by excessive screen time, cognitive offloading, rapid task-switching, and digital-first social interactions, is producing measurable cognitive decline in Generation Z, with significant implications for educational outcomes, workforce readiness, and societal functioning. However, the same neuroplasticity that enables cognitive decline can be harnessed for cognitive enhancement through intentional educational design that prioritizes deep focus, active learning, and face-to-face interaction. The most effective educational interventions must address multiple dimensions simultaneously, including limiting screen time to 30 to 45 minutes per week per subject, implementing cell phone bans on school premises, prioritizing hands-on learning with physical manipulatives, and fostering face-to-face social interactions that activate mirror neurons at full strength.

One of the most visible transformations brought by AI is personalized learning. Adaptive platforms can adjust difficulty, sequence, pacing, and feedback based on a learner's performance, thereby making instruction more responsive to individual needs. While, research on intelligent tutoring systems has long suggested that well-designed digital tutoring can improve learning outcomes in specific domains, it has been found that educational systems may overestimate what can be captured through data traces alone, reducing a learner to patterns of clicks, scores, and predicted probabilities. When this happens, the learner's agency and interpretive life can be flattened into a data profile. The danger is not only technical but philosophical: if education is treated primarily as optimization, then curiosity, struggle, and immeasurable growth may be undervalued. For this reason, we advocate that personalization should be understood as a support for learning, not a substitute for human pedagogical judgment. This is why the role of the teachers and assessment are fundamental although AI is also transforming their role as well.

According to The U.S. Department of Education AI can support teaching and learning, but should be used in ways that keep educators in the center of decision-making. Rather than serving only as transmitters of content, teachers are increasingly becoming designers of learning environments, interpreters of data, and mentors who guide students through complex tasks. This reorientation can be beneficial because it frees educators from repetitive administrative labor and allows more time for conversation, feedback, and relational work. At the same time, assessment is changing just as dramatically. Automated grading systems can evaluate selected forms of work quickly, while generative tools can support rubric-based feedback, draft analysis, and formative coaching. Such systems may improve speed and consistency, especially in large classes, but if grading systems are opaque, biased, or poorly calibrated, they may damage trust and undermine fairness. Therefore, assessment remains one of the most sensitive areas in educational AI.

While, there is no doubt that AI is most beneficial when students use it to deepen understanding, verify outputs, and organize learning, recent studies show that a striking difference in student's verification habits: 72% of higher-performing students said they always or almost always check AI-generated information, compared with 28% of lower-performing students. That supports the argument that AI can amplify existing inequality if students use it as a shortcut rather than a learning aid. UNESCO has also reported that fewer than 10% of schools and universities had formal guidance on generative AI at the time of its survey, which suggests a major opportunity for institutions to build structured, beneficial use rather than leaving students to improvise. As a result AI-based educational technologies may lead to digital divide.

AI systems often require devices, connectivity, technical literacy, and institutional support that many learners do not have. If high-quality AI tools are available only to privileged schools and affluent families, the result may be a new layer of educational inequality. Even where access exists, the language, assumptions, and cultural norms embedded in AI systems may not reflect local educational contexts. This is why educational transformation must include infrastructure, training, and policy, not just software procurement.

Equity also requires attention to bias. AI systems trained on incomplete or skewed data may reproduce existing social inequalities, including those related to race, class, gender, language, and disability. In education, such bias can have lasting consequences because algorithmic decisions may shape placement, intervention, and opportunity. Therefore, a responsible educational system must therefore treat fairness as a design requirement rather than an afterthought.

Keywords: Human centered AI, Innovative education, Trustworthy AI

References:

- Harrison, C. J., & Nelson, T. D. (2025). Gen Z reports early cognitive decline. Here's what to know about the brain rot epidemic. *INC Magazine*.
- UNESCO. (2026). *Artificial intelligence in education*, UNESCO Publishing
- U.S. Department of Education. (2023). *Artificial Intelligence and the Future of Teaching and Learning*.
- Floridi, L., et al. (2018). AI4People — An ethical framework for a good AI society. *Minds and Machines*, 28(4), 689–707.
- V.B. Pavlović, N. Milosavljević, V. Pavlović, B. Vlahović, (2023), Artificial intelligence and digital technologies in digitally-supported university education Mathematics for Human Flourishing in the Time of COVID-19 and Post COVID-19 Published by De Gruyter, 2023, 195-200

Intellectual Property Rights in the Era of AI: Challenges and Perspectives of the Contemporary Legal Framework

Dušan Stojanović¹, Anđelka Zečević²

¹ Faculty of Law, University of Niš, Serbia, dusans00@yahoo.com

² Mathematical Institute, Serbian Academy of Sciences and Arts, andjelkaz@mi.sanu.ac.rs

The rapid development of generative AI models and the constant need for massive datasets to support their training challenge the foundations of intellectual property rights, which were defined in an era of exclusively human creativity. From the perspective of the existing European Union legal framework, intellectual property can be examined from three different aspects.

(1) The use of copyright-protected content for the training of AI models must comply with the EU AI Act, currently the most comprehensive regulatory framework in the field of artificial intelligence, as well as its underlying copyright framework, in particular the Directive on Copyright in the Digital Single Market (DSM Directive, 2019/790), and its provisions on the text and data mining exception.

(2) The authorship and protection of AI-generated works, with a critical examination of *originality* and *intellectual creation* as prerequisites for copyright protection.

(3) The position of humans who contribute to creation through AI tools, and the boundaries of human input sufficient for establishing authorship rights over the final output. This is particularly relevant for the process of *prompting*, which is still underexplored despite its significant influence on the performance and outputs of AI systems.

In this talk, we will address the first aspect and discuss existing regulatory approaches to the protection of copyrights in the context of the development of AI models. We will discuss copyright questions in the light of the mainstream large multimodal models such as ChatGPT, Gemini, Grok, and Claude, as well as open-sourced models such as Apertus, developed by the Swiss AI initiative, or government supported models such as PLLuM, Polish Large Language Model.

As the development of AI law does not take place exclusively through legislation but also through case law, this talk also examines the decisive contribution of courts across different jurisdictions, presenting key cases that have laid the foundations for the further development of legal responses to the challenges posed by generative artificial intelligence. Through these perspectives, the talk critically assesses the coherence and adequacy of the existing EU legal framework in addressing the challenges that generative artificial intelligence poses to traditional categories of intellectual property law.

Keywords: intellectual property, copyrights, large multimodal models

References:

- General Data Protection Regulation, <https://gdpr.eu/>
EU Artificial Intelligence Act, <https://artificialintelligenceact.eu/>
Directive (EU) 2019/790 of the European Parliament and of the Council, of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC, <https://eur-lex.europa.eu/eli/dir/2019/790/oj/eng>
APERTUS: Democratizing Open and Compliant LLMS for Global Language Environments, Technical report, <https://arxiv.org/pdf/2509.14233>
Kocoń, Jan et al. PLLuM: A Family of Polish Large Language Models, <https://arxiv.org/html/2511.03823v1#S1>

Concerns over the use of synthetic data in biomedicine – What could ever go wrong?

Miljana Tanic^{1,2}

¹ Department of Experimental Oncology, Institute for Oncology and Radiology of Serbia, Belgrade, Serbia, tanic.miljana@ncrc.ac.rs

² UCL Cancer Institute, University college London, London, UK, m.tanic@ucl.ac.uk

Data quality is the most important determinant for development of useful and generalizable classifiers based on machine learning (ML) models for healthcare. Classifiers are developed typically by training and fine tuning a model on a training dataset, next, evaluating its performance on a separate test dataset, ideally followed by external evaluation on a dataset from a different source, before being applied in real-world environment. Unfortunately, in biomedicine data are often scarce, with inherent biases, with limited number of patients and sometimes huge number of features. An especially daunting task is to collect sufficient data from events with low prevalence (fewer than 1000 outcomes per 100 000 individuals), while taking into consideration all relevant confounders or comorbidities, resulting in heavily unbalanced datasets. Sometimes, after years of cohort building, we just don't have enough good quality data. So, what can we do about it? Recent calls for the use of synthetic data propose - *Let's invent them!* What could ever go wrong with that?

Arguments in favor of synthetic data use often quote reduced cost, addressing privacy concerns, augmenting datasets and overcoming data bottlenecks for rare events. Big tech companies, such as NVIDIA or IBM, are particularly vocal proponents of synthetic data. Growing number of companies (e.g Simulacrum, Syntho, BlueGen, Mommentum etc.) are dedicated to generation of synthetic data for training models for healthcare AI, with increasing use in medical research on cancer, cardiovascular diseases or infectious diseases (Eckardt et al., 2026). While there may very well be a place for the use of synthetic data in some fields, in this opinion piece, I will try to argue from a position of biomedical researcher why this may not be such a great idea for applications in healthcare.

In biomedical research there is a clear distinction between technical replicates (repeated measurements of a parameter on the same subject) and biological replicates (measurements of a parameter in different subjects)(Blainey et al., 2014). Biological replicates aim to capture information about random biological variation, while technical replicates are used to determine variation stemming from measurement errors. Biological variation among human subjects is notoriously high given the profound diversity of human genomes between different populations, differences between biological sexes and changes over age compounded by different environments, socio-economic factors and cultures, that are affecting human phenotypes. Believing that we can faithfully recreate the entire biological variability would be pure hubris! Nevertheless, we commonly use approximations and simplifications, aware of their limitations, to help us reach the truth. Can synthetic data help us to do that?

Synthetic data is typically generated to replicate the statistical parameters of a real-world dataset. Faithful replication of intrinsic variation of a real-world dataset can be achieved only if we fully understand its distribution and characteristics. Being drawn from known distributions, synthetic data cannot help us to uncover new knowledge, the unknown unknowns. This begs the question - if we already know everything about the real-world dataset, why do we need the synthetic data for? One common answer to this question is to increase the size of the dataset. However, by augmenting the dataset with synthetic data we are not capturing more of the biological variation, but simply introducing more of quasi-technical replicates. Moreover, augmenting the dataset with synthetic data may lead to an exaggeratedly reduced

confidence intervals and p-values. In other words, artificial data would lead to artificial conclusions that prevents the validity estimation of the model in the real world before deployment (Koul et al., 2025).

Distribution leakage, due to a circular logic, where a model was used to generate initial synthetic data, separated into training and test dataset and then evaluated on the held-out test set, provides poor indication how it would handle real-world data. Even if we use one model to generate synthetic data, to be used as a training dataset for training a new model, errors are inevitable. Synthetic data may suffer from reduced complexity, overrepresenting common patterns and omitting rare ones, potentiating bias, or inadvertently include observations that cannot occur in reality (e.g. 190cm height over 30kg weight), again affecting the generalizability of the model. This was elegantly described as in a landmark *Nature* study that demonstrated “model collapse” in model-generated content caused by progressive loss of the tails of the original distribution even in the early stages (Shumailov et al., 2024).

Second use case for synthetic data is related to privacy concerns and cost reduction (Nowok et al., 2016). Medical and genetic data fall into the category of “particularly sensitive” personal data, requiring strict provisions to maintain patient privacy and anonymization techniques. While concern for data privacy may be a valid reason to use synthetic data, a cynic might argue that companies are often incentivized to use cheaper options, skipping the hard work, time and associated expenses required to compile a representative, manually curated high-quality dataset. Regulatory approval for medicines and medical devices requires thorough validation in real subjects. It may end up being more costly to run a failed clinical trial based of synthetic data than to invest in initial real data acquisition.

Finally, the biggest concern relating to the ethics of using synthetic data, is that it is, in essence, data fabrication. Just because we are being honest about using fake data, doesn’t mean it is the right thing to do. Deploying AI solutions in healthcare that eventually fail because our data was not representative, may lead to eroding of public trust in science. Additionally, it may cause ripple effects by poisoning the literature with erroneous conclusions and instigating new projects or trials based on falsified data (Fitzgerald, 2024).

I am not arguing synthetic data is of no use in biomedicine – e.g., it may be used in the education setting as dummy data for teaching courses, or for preliminary analysis and code development (Nowok et al., 2016). However, I firmly believe that when human lives are at stake, let’s not substitute silk for polyester.

Keywords: Machine learning, synthetic data, biomedicine, healthcare

References

- Blainey, P., Krzywinski, M., & Altman, N. (2014). Points of significance: replication. *Nat Methods*, 11(9), 879-880. <https://doi.org/10.1038/nmeth.3091>
- Eckardt, J. N., Hahn, W., Prelaj, A., Bornhauser, M., Middeke, J. M., & Kather, J. N. (2026). Artificial intelligence-generated synthetic data for cancer research and clinical trials. *Nat Rev Cancer*, 26(5), 351-363. <https://doi.org/10.1038/s41568-026-00912-4>
- Fitzgerald, A. (2024). Why Synthetic Data Can Never Be Ethical: A Lesson from Media Ethics. *Surveillance & Society*, 22(4), 477-482. <Go to ISI>://WOS:001375489300012
- Koul, A., Duran, D., & Hernandez-Boussard, T. (2025). Synthetic data, synthetic trust: navigating data challenges in the digital revolution. *Lancet Digit Health*, 7(11), 100924. <https://doi.org/10.1016/j.landig.2025.100924>
- Nowok, B., Raab, G. M., & Dibben, C. (2016). synthpop: Bespoke Creation of Synthetic Data in R. *Journal of Statistical Software*, 74(11), 1-26. <https://doi.org/10.18637/jss.v074.i11>
- Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. *Nature*, 631(8022), 755-759. <https://doi.org/10.1038/s41586-024-07566-y>

Ai-Assisted Learning in Practice: Github Copilot as a Programming Tutor for Non-Computer Science Students

*Olga Timcenko*¹

¹ OTConsulting, Copenhagen, Denmark

Overview: This case study examines how GitHub Copilot was used as an AI "tutor" in a second-semester programming course for Media Technology students at Aalborg University in Denmark — students who are not computer science majors. Drawing on frameworks from Kurtz et al. (2024), the paper reports that AI-assisted students produced markedly higher-quality projects and earned significantly better exam grades than students in prior, non-AI cohorts, while also surfacing risks around over-reliance on AI-generated code.

Background and Context: Media Technology students focus on design and digital content rather than programming, yet need a working grasp of code to build interactive media such as games. After an introductory first-semester course in C# basics, students take "Programming of Mobile Applications" in semester two — the course studied here — which shifts toward object-oriented programming for 2D game development in Unity and C#. The technical demands of Unity (game mechanics, animation, input handling) have historically limited the scope and polish of student projects, especially under time pressure. GitHub Copilot, an AI code assistant offering real-time suggestions and debugging help, was introduced as a way to act as an intelligent tutor — while the literature also warns that AI support can risk reducing students' critical-thinking development if not carefully managed.

Course Structure and requirements: The course is a 5 ECTS class with roughly 50–60 students per cohort. It ran twice without generative AI before Copilot was formally introduced in Spring 2024. Students worked in small teams (using GitHub for storage and collaboration) and built an initial game version during a three-day Game-jam near the end of the course. For the final exam, each student individually refined or rewrote their game and defended their design and code in a 20-minute oral exam with the teacher and an external censor. Copilot use was actively encouraged, but with one firm rule: regardless of whether code was self-written or AI-generated, students had to fully understand and be able to explain every part of it. This mirrors Kurz et al.'s (2023) recommendation that AI use should be carefully monitored to support learning rather than replace it. The important remark here is that quality of projects themselves was not a part of the grade, but only the student's reflection on the written code and explanations of the choices made and algorithms used to develop the project.

Key Findings: Grades: Average final grades (Danish 7-point scale) across three course generations:
CohortAverage Grade:

- Generation 1 (no AI): 6.69
- Generation 2 (no AI): 6.08
- Generation 3 (with Copilot) 7.67

The AI-assisted cohort's average was roughly a full grade point higher than either prior cohort, reflecting more technically ambitious projects that students were able to explain and discuss the code, debug confidently, and answer curriculum questions with concrete examples during the oral exam — evidence, we could argue, of deeper engagement rather than just AI doing the work.

Challenges and Risks: Not all outcomes were positive. Several students could not fully explain the logic behind AI-generated code segments they had used in their own projects, despite implementing them — which is a clear sign of over-reliance. This is the central risk of AI tutoring: it can mask gaps in fundamental understanding unless explicitly checked for, echoing the caution raised by Kurz et al. (2024).

Recommendations:

- Curriculum focus: shift emphasis from coding mechanics toward higher-order skills — programming principles, game design logic, and computational thinking.
- Exam design: add questions that ask students to critique, improve, or propose alternatives to AI-generated solutions, not just explain accepted code.
- Critical thinking: train students to judge when Copilot's suggestions are appropriate versus suboptimal.
- Problem-solving: ensure students can independently diagnose game-mechanic and logic issues AI may miss.
- Code adaptation: build skill in optimizing and tailoring AI-suggested code to project-specific needs.
- Creative innovation: redirect time saved on manual coding toward design creativity and pushing project ambition.
- Balance: explicitly set expectations for when AI should and shouldn't be used, and assess students both with and without AI support.

Conclusion: The case study concludes that GitHub Copilot can function effectively as a private AI tutor in non-CS programming education, raising both technical quality and oral-exam performance. But its presence demands rethinking course curricula and assessment so that students become adaptable, independent thinkers — not just consumers of AI-generated code.

Reference:

Kurtz, G. et al. (2024). Strategies for Integrating Generative AI into Higher Education. *Educ. Sci.*, 14(5), 503.

Ethical Challenges of Integrating AI into the Educational Process: Balancing Technological Progress with Academic Integrity

Zubova Oksana¹, Khusnutdinova Margarita², Filipova Aleksandra³

¹ Laboratory of Comprehensive Childhood Research, Vladivostok State University, Faculty of Sociology, Moscow State University, Russia, email zubovaoksana@bk.ru

² Laboratory of Comprehensive Childhood Research, Vladivostok State University,

³ Russia, email penkovamargo@yahoo.com

⁴ Laboratory of Comprehensive Childhood Research, Vladivostok State University,

⁵ Russia, alexandra.filipova77@gmail.com

This research was funded by the Russian Science Foundation, grant No. 26-18-00812, <https://rscf.ru/project/26-18-00812/>

The active introduction of artificial intelligence into school education raises a number of risks related to both technological and socio-ethical aspects. The purpose of the study conducted by the Laboratory of Comprehensive Research on Childhood at Vladivostok State University was to identify barriers and actual strategies for using AI in the school educational environment. In March and April 2026, two focus groups were conducted with 12 16-17-year-old students in Moscow (4 boys, 4 girls) and Vladivostok (2 boys, 2 girls), supplemented by interviews with 4 teachers who use AI tools in their teaching activities. The transcripts were processed using thematic analysis. This study was supported by a grant from the Russian Science Foundation № 26-18-00812.

All students expressed a positive attitude towards the further expansion of the use of artificial intelligence, as well as the constant use of AI tools in everyday and educational practices. However, the main concerns of the students are related to the initially incorrect guidelines and objectives for using AI. As a 11th-grade student wrote: "There may be a trend towards degradation if students... mechanically copy... It is necessary to teach the correct use of new AI tools..."

A 11th-grade student notes that "artificial intelligence cannot replace everything in education... technology that needs to be improved." This is primarily due to errors and inaccuracies in information, where neural networks can generate fictional facts and cite non-existent literature, compromising the scientific accuracy of educational materials. AI algorithms may also contain built-in biases, leading to discrimination against students based on various characteristics. Incorrect algorithm settings or data breaches can compromise the privacy of student data. It is not always clear what data AI decisions are based on, and who is responsible for algorithmic errors. All of these risks contribute to the formation of technological and pedagogical concerns. However, teachers also face social fears that AI may replace human interaction, negatively impacting students' emotional development and the quality of education, and altering the dominant roles of teachers.

The introduction of artificial intelligence into the school educational environment raises important issues related to academic integrity, the development of students' critical thinking, and the preservation of their engagement (Nguyen A., 2023). The use of AI can both contribute to the development of learning skills and pose risks of unethical behavior. According to Turnitin's 2023 data, out of the more than 200 million student submissions, 22 million (11%) already contained at least 20% AI-generated content, and around 6 million (about 3%) contained at least 80% AI traces (Turnitin, 2024). However, academic integrity goes beyond plagiarism, as defined by the International Center for Academic Integrity (ICAI). It includes adherence to values that shape principles of behavior, such as honesty, trust, fairness, respect, and

responsibility (Turnitin. Academic integrity essentials, 2024.). Applying these values to the use of AI reduces the risks associated with the substitution of independent homework completion, which makes it difficult to assess a student's actual knowledge and skills; the presentation of generated content as original, without indicating the use of technology; and the emerging dependence on AI, where students begin to lose their skills in independent learning, critical thinking, and become lazy, shifting all responsibility and decision-making to AI.

Students, like teachers, believe that the participants of the educational process are not yet well adapted to AI in terms of ethical aspects. A 10th-grade student notes that "if you show an old-school teacher a presentation made by AI, they probably won't notice the difference... Teachers are unable to control how honestly a student has completed their work..." From this example, which describes the spread of practices related to AI among different participants in the educational process, it is evident that several factors contribute to the existing barriers to integrating with artificial intelligence. First of all, this is the different level of digital literacy, awareness of the need for continuous professional development in the field of not only digital competencies, but also AI competencies. It is also influenced by belonging to different generations, who are undergoing socialization and professional development in different institutional structures, including the established social perceptions of the role of the teacher profession. Despite the fact that there are indeed more people among the older generation who oppose the use of AI in schools, and the younger generation of teachers is more optimistic about technological progress and has a better understanding of information technology, the personal characteristics of the teacher, their willingness to accept new changes and develop professionally, as well as the existing attitudes of the administration towards AI, which shape the school environment, practical experience with new tools, and the opportunity to receive ongoing training and consultations, are important factors. The teacher's adherence to the principles of ethical symmetry contributes to the development of students' reflective use of AI, the search for new methods of promoting academic integrity without restrictions or threats, allowing students to expand their educational and cognitive opportunities through the "correct configuration" of AI tools.

Therefore, integrating AI into the educational process requires a balanced approach that considers both technological capabilities and ethical standards, allowing for the maximum utilization of AI's potential to enhance education while maintaining academic integrity and ethical standards.

Keywords: academic integrity, generative artificial intelligence, critical thinking, ethical symmetry, self-regulation of learning activities.

References:

- Turnitin. (2024). https://www.turnitin.com/press/turnitin-first-anniversary-ai-writing-detector?utm_source=Securitylab.ru
- Turnitin. Academic integrity essentials. (2024). <https://go.turnitin.com/en-us/ebook/academic-integrity-essentials>
- Nguyen A. et al. (2023) Ethical principles for artificial intelligence in education. *Education and Information Technologies*, 4 (28), 4221–4241

COMPUTER VISION



Usability analysis of Hu and Zernike moments for image classification

Dušan Denčić¹, Nataša Milosavljević¹, Tatjana Davidović²

¹ University of Belgrade, Faculty of Agriculture, Belgrade, Serbia,
{dusan.dencic, natasam}@agrif.bg.ac.rs

² Mathematical Institute of the Serbian Academy of Sciences and Arts, Belgrade, Serbia,
tanjad@mi.sanu.ac.rs

This study presents a shape-based image classification framework that combines Hu and Zernike moments with a Support Vector Machine (SVM), whose hyperparameters are optimized using metaheuristic algorithms. Our main goals are to examine not only the usability of the considered moments and their combinations but also the effectiveness of metaheuristics in hyperparameter tuning. We selected lightweight models with hand-crafted descriptors that retain an attractive accuracy-efficiency trade-off. A principled shape-centric pipeline: Hu's seven invariant moments (Meng, W., *et al.*, 2025) and Zernike moments (Zhang, J., *et al.*, 2025) as features, coupled with an RBF-kernel Support Vector Machine (SVM) (Natanael, K., Vincentius, R., & Herwindiati, D.E., 2025) is revisited. The Hu moments constitute a set of seven nonlinearly defined descriptors based on normalized central moments up to the third order. Their key property is invariance with respect to fundamental geometric transformations of the image: translation, rotation, and scaling. This robustness makes them suitable for the analysis and classification of visual objects, irrespective of their position, size, or orientation. The method of Zernike moments is based on the application of Zernike orthogonal polynomials which are defined on the unit disk. The image silhouettes were first binarized, and the extracted Hu and Zernike moments were standardized and used as input features for SVM. Although SVMs are robust and data-efficient, their performance is controlled by two hyperparameters: the regularization constant C and the kernel width γ . An exhaustive grid search for tuning these parameters is sample-inefficient under tight evaluation budgets, while random search under-exploits promising regions once discovered. Metaheuristics, which are derivative-free stochastic optimizers with explicit diversification and intensification, offer a compelling alternative. Therefore, we propose applying Variable Neighborhood Search (VNS) (Brimberg, J., *et al.*, 2023), a trajectory-based metaheuristic that systematically alternates neighborhoods to escape local optima. Bee Colony Optimization (BCO), a population-based approach with implicit information sharing, has already been applied to this problem [6]. We modify the existing BCO approach and include it into performance evaluation of various metaheuristic-based approaches. The modified BCO and newly proposed VNS are compared with the baseline SVM trained under the default hyperparameter values (DB), as well as with the improved baseline SVM (IB), trained using a set of randomly selected hyperparameter values.

Preliminary experimental evaluation is performed on 4 publicly available image datasets. As the performance measure we selected the F1 score. All experiments were executed on a desktop configuration with Intel Core i3-10100 @ 3.60GHz (4 cores, 8 threads), 8 GB DDR4 RAM, under Windows 10 Education, version 22H2, OS Build 19045.6332 operating system. The implementation is done in Python, version 3.13.7. Excluding IB, all compared approaches need to be executed multiple times for different seed values to examine their stability. The values of relevant parameters are selected as follows. The domains for SVM hyperparameters are $C \in [0.1, 100]$, $\gamma \in [10^{-4}, 1]$. The training/test split was 80:20. VNS has a unique parameter k_{max} that is tuned using the values from the set $\{2, 6, 10, \dots, 94\} = \{2 + 4m, m = \overline{0, 23}\}$. The values of three BCO parameters (the number of bees B , the number of transformations in each iteration NC , and the number of modifications of a single solution o) are determined experimentally combining the following domains $B \in \{2, 4, 6, 8, 10\}$, $NC \in \{2, 6, 10, 14, 18\}$, $o = rndint(2, 10)$. Stopping criteria are defined by a maximum runtime limits (60 seconds per feature

mode). Table 1 contains average values of F1 scores for a single database that illustrate the influence of each shape-based descriptor. The best values are bolded. The average results including Hu and Zernike moments and their combination for each of 4 databases tested are presented in Figure 1.

Table 1. Average F1 scores depending on shape-descriptor moments

men_vs_women	DB	IB	B C O	V N S
Hu	0.4623660	0.4623660	0.4623660	0.4623660
zernike	0.4623656	0.6059400	0.5979957	0.5201828
Both	0.4623656	0.5310485	0.5482396	0.6279762

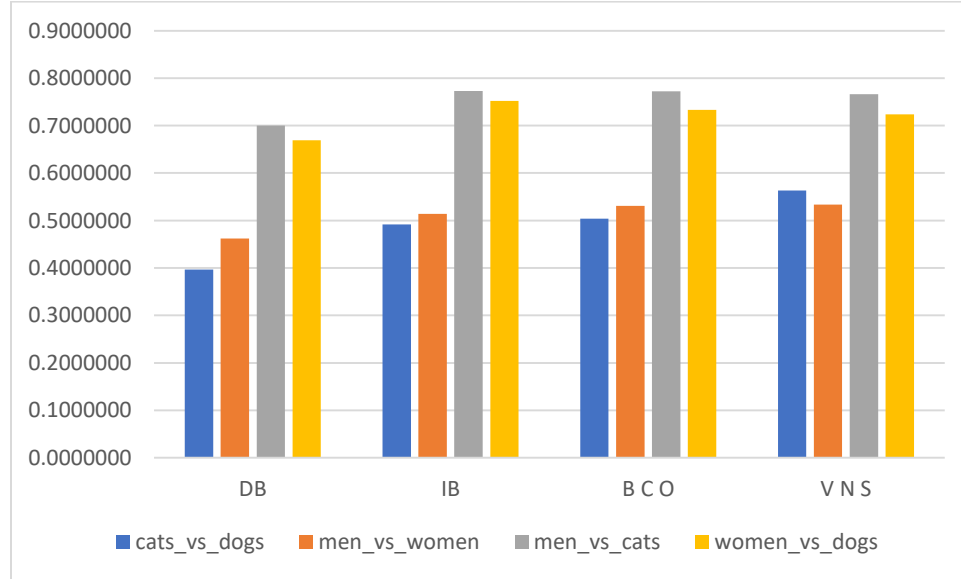


Fig. 1. Average F1 scores over all shape-based descriptors

The presented results show that SVM, optimized using the metaheuristic methods, achieves higher F1-score values compared to the default SVM model, with the combination of both Hu and Zernike moments with VNS providing an efficient and robust approach to shape-based image classification.

Keywords: Shape descriptors, support vector machine, hyperparameter optimization, metaheuristics.

References:

- Meng, W., et al. (2025). Image-free Hu invariant moment measurement by single-pixel detection. *Optics & Laser Technology* 181, 111581.
- Zhang, J., et al. (2025). Image authentication method based on Zernike moments and correspondence ghost imaging. *Optics and Lasers in Engineering* 194, 109238.
- Natanael, K., Vincentius, R., & Herwindiati, D.E. (2025). Comparison of Linear and Radial Basis Function Kernels in Support Vector Machine for Rice Variety Classification. In: *Proceedings of the 15th IEEE Symp. Computer Applications & Industrial Electronics (ISCAIE)*, pp. 114–119.
- Brimberg, J., et al. (2023). Variable neighborhood search: The power of change and simplicity. *Computers & Operations Research* 155, 106221.
- Denčić, D., Milosavljević, N., & Davidović, T. (2025). Shape-based image classification using Hu and Zernike moments with Bee Colony Optimization for tuning SVM parameters. In: *Proceedings of the SYMOPIS 2025 Conference*, pp. 30–35.

Dataset for UAV Visual Cross-View Geo-Localization in Mixed Forest-Urban Interfaces

Jelena Jevremović¹, Boban Sazdić-Jotić¹, Rade Pavlović¹, Dušan Randjelović¹

¹ Military Technical Institute, Belgrade, Serbia, jelena.jevremovic@mod.gov.rs,
boban.sazdic.jotic@vs.rs, rade.pavlovic@mod.gov.rs, dusan.randjelovic@mod.gov

Unmanned Aerial Vehicles (UAVs) rely on Global Navigation Satellite Systems (GNSS) to ensure accurate localization and trajectory tracking. However, GNSS signals are vulnerable to adversarial electronic interferences, such as intentional jamming and spoofing and unintentional signal attenuation caused by multipath propagation in dense canopies, urban canyons, or communication-sensitive operational zones. In GNSS-denied environments, alternative sensory frameworks—such as onboard Inertial Measurement Units (IMUs) or Inertial Navigation Systems (INS)—are typically employed. Nonetheless, these sensors suffer from cumulative errors over time, leading to positional drift and an inability to maintain trustworthy geospatial referencing. To solve the limitations of inertial drift under poor GNSS conditions, computer vision methods offer a practical alternative for continuous positioning. Specifically, visual cross-view geo-localization (visual navigation) facilitates autonomous navigation by matching real-time video streams captured by the UAV’s onboard camera with pre-registered, geo-referenced satellite or orthophoto basemaps. This paradigm maps the navigation task to a content-based image retrieval (CBIR) problem, estimating the coordinates by identifying the geometric match between NADIR (immediate downward-facing) field of view (FOV) and a comprehensive satellite reference database. Original datasets (benchmarks), such as University-1652 (Zheng et al., 2020) and DenseUAV (Dai et al., 2023), established foundational baselines for testing visual navigation algorithms. Concurrently, models within the computer vision community have progressively shifted from Convolutional Neural Networks (CNNs) toward Vision Transformers (ViTs), using self-attention mechanisms to capture long-range contextual dependencies. Despite these advancements, existing approaches are mostly optimized for urban topographic structures. Therefore, because many benchmark datasets are simply dominated by urban, suburban, or infrastructure -rich scenes, models trained on them face a significant domain shift and experience severe accuracy degradation when transitioning to mixed forest-urban topographic structures. This vulnerability stems from fundamental deficiency in current training data, which fails to capture the unstructured and self-similar characteristics of natural vegetation. While the deployed architecture achieves robust retrieval accuracy in homogeneous urban environments, its discriminative capability deteriorates in mixed forest-urban landscapes where dense forest canopies and residential infrastructure coexist.

To address these critical limitations, this paper introduces a novel cross-view geo-localization dataset captured across forest-urban interfaces in Belgrade, Serbia. The benchmark is validated utilizing an evaluation framework based on a Pseudo-Siamese model (Chopra et al., 2005) featuring a weight-shared Vision Transformer backbone (ViT-S/16) (Dosovitskiy et al., 2021) and a Generalized Mean (GeM) pooling layer. To systematically evaluate how structural transitions alter visual feature representations, our dataset explicitly captures two distinct environmental zones along a continuous flight trajectory: **woodland canopies** (specifically, the Košutnjak and Makiš forest zones) because trees and forests look almost identical to the above and **residential settlements** (infrastructure domain) because features like rooftops, fences, and streets offer strong visual cues, but the layout of Balkan suburbs differs from the Asian and Western data used for pre-training. To overcome the challenges of generating our dataset, we deployed a WingtraOne UAV platform to capture NADIR aerial data, mapping it against corresponding publicly available satellite imagery. An underlying methodological contribution of this work is a specialized image-segmentation tool utilized to generate localized image patches and consolidate them into a

structured cross-view geo-localization dataset. The dual-stream alignment of our dataset is tailored for a Pseudo-Siamese framework and projects different UAV and satellite perspectives into a common low-dimensional embedding space, enforcing geometric alignment regardless of viewing angles or sensor characteristics. To optimize this pipeline, we employ a layer-wise strategy that divides the 12 attention blocks into two operational domains based on feature abstraction: *Zone 1 (Low – Level Geometry): Blocks 1 to k* , *Zone 2 (High – Level Semantics): Blocks ($k+1$) to 12*, where variable k represents the boundary between frozen and trained layers. Zone 1 remains frozen to preserve universal geometric features, while Zone 2 is fine-tuned using a combined loss function that includes Cross-Entropy (CE) Loss, a Soft-Weighted Triple Loss, and Kullback-Leibler (KL) Loss. This ensures the model adapts specifically to the layout and texture of the Čukarica area. We evaluate our approach through a two-phase comparative workflow. First, we benchmark the pre-trained ViT-S/16 model directly on the Čukarica segments. This allows us to isolate the exact performance drop caused by shifting from standard datasets to the challenging forest-urban interface. Second, we systematically shift the freezing boundary k to track how well each sub-domain recovers as more layers are exposed to fine-tuning. To systematically evaluate the expected localization performance and quantify the domain shift, the proposed evaluation protocol relies on three key metrics: Recall@K (R@K), Mean Average Precision (mAP), and Satellite-view Drone-view Matching at K (SDM@K). Furthermore, we conduct an ablation study investigating the expressiveness and feature localization capacity across the 12 attention blocks of the ViT-S/16 model. By freezing the lower, localized feature layers and fine-tuning the upper, context-aware layers, we effectively recover and optimize baseline retrieval performance. Our empirical findings provide critical insights into how self-attention mechanisms dynamically allocate weight distributions when processing the disparate boundaries of natural landscapes and human-made infrastructure.

The proposed approach primary limitation is that the current datasets represent a single temporal snapshot, meaning the network's resilience to drastic seasonal changes, varying illumination conditions, and long-term phenological shifts in canopy cover remains unassessed. Second, although the ViT-S/16 backbone achieves high retrieval accuracy, its heavy self-attention mechanisms introduce substantial computational overhead that may constrain real-time deployment on hardware-restricted edge processing units aboard micro-UAVs. Future work will be directed toward four principal objectives: 1. **Multi-Sensor Fusion:** Integrating onboard sensory data (including IMU, INS, and odometry) to narrow the search window, thereby minimizing retrieval latency and optimizing localization robustness. 2. **Oblique Perspective Adaptability:** Extending the model's capabilities to handle non-NADIR imaging, focusing on mitigating perspective and lens distortions. 3. **Temporal and Seasonal Expansion:** Continuously updating the benchmark dataset with multi-temporal imagery captured during different periods of the year. 4. **Edge AI Hardware Deployment:** Migrating the pipeline to real-time embedded edge AI devices to validate the visual navigation framework under actual online deployment conditions.

Keywords: Cross-view geo-localization, Drone, GNSS, Vision Transformer (ViT), Transfer learning, Navigation.

References:

- Chopra, S., Hadsell, R., & LeCun, Y. (2005). Learning a similarity metric discriminatively, with application to face verification. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 1, 539-546.
- Dai, M., Zheng, E., Feng, Z., Qi, L., Zhuang, J., & Yang, W. (2023). Vision-based UAV self-positioning in low-altitude urban environments. *IEEE Transactions on Image Processing*, 33, 493-505.
- Dosovitskiy, A., et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations (ICLR)*.
- Radenović, R., Tolias, G., & Chum, O. (2019). Fine-tuning CNN image retrieval with no human annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 41(7), 1655-1668.
- Zheng, Z., Wei, Y., & Yang, Y. (2020). University-1652: A multi-view multi-source benchmark for drone-based geo-localization. *ACMMM*, 1395-1403.

Improving Soil Property Estimation from Spectroscopy Using Elastic-Net-CNN: An OSSL-Based Study with Transfer Learning

Mina Stanković¹, Stefan Kostić², Dimitrije Gajin³, Ognjen Antonijević⁴, Branislav Bajat⁵, Aleksandar Sekulić⁶, Milan Kilibarda⁷, Slobodan Jelić⁸

¹1University of Belgrade, Faculty of Civil Engineering, Serbia, mina.stankovicc03@gmail.com

²2University of Belgrade, Faculty of Civil Engineering, Serbia, stefankostic562@gmail.com

³3University of Belgrade, Faculty of Civil Engineering, Serbia, dimitrijegajin@gmail.com

⁴4University of Belgrade, Faculty of Civil Engineering, Serbia, oantonijevic@grf.bg.ac.rs

⁵5University of Belgrade, Faculty of Civil Engineering, Serbia, bajat@grf.bg.ac.rs

⁶6University of Belgrade, Faculty of Civil Engineering, Serbia, asekulic@grf.bg.ac.rs

⁷7University of Belgrade, Faculty of Civil Engineering, Serbia, kili@grf.bg.ac.rs

⁸8University of Belgrade, Faculty of Civil Engineering, Serbia, sjelic@grf.bg.ac.rs

Introduction: Soil properties, such as soil organic carbon (SOC) content, nutrients, pH value, and mechanical composition (texture), are key indicators of soil fertility and health; thus, monitoring them is crucial for agriculture, sustainable food production (Roja et al., 2016), and environmental protection (Tóth et al., 2018). The standard approach for their determination is laboratory chemical analysis, which is highly precise but slow, expensive, and time-consuming, making it difficult to scale across large sample sets. As an alternative, visible and near-infrared (VisNIR), near-infrared (NIR), and mid-infrared (MIR) spectroscopy, coupled with machine learning, enables more rapid and cost-effective estimation of soil properties, highly suitable for analyzing large volumes of samples. Furthermore, deep learning can automatically extract relevant features from spectral data, while transfer learning allows for the repurposing of trained models for related tasks and novel environmental conditions.

Data and Methods: This study utilized data from the Open Soil Spectral Library (OSSL) public database (Safanelli et al., 2025), comprising four spectral datasets (Neospectra NIR and MIR, OSSL VisNIR and MIR) to evaluate the following properties: soil organic carbon (SOC) content, total nitrogen (N), pH, clay, sand, silt, sulfur, and zinc. The methodology for preprocessing and feature extraction was adopted from existing literature (Zou et al., 2024). Data cleaning involved the removal of missing and negative values, duplicates, and outliers (utilizing the Isolation Forest algorithm). Highly skewed properties were log-transformed, whereas all performance metrics—coefficient of determination (R^2), mean absolute error (MAE), and root mean square error (RMSE)—were calculated on the original scale to ensure comparability. Spectral preprocessing encompassed baseline correction, Savitzky-Golay filtering with a second derivative, signal correction using the Direct Orthogonal Signal Correction (DOSC) method, and dimensionality reduction via Principal Component Analysis (PCA, 120 components). Crucially, DOSC and PCA were fitted exclusively on the training dataset to prevent test data leakage from artificially inflating predictive performance. As a baseline alternative, both raw and normalized (Standard Normal Variate, SNV) spectra were evaluated. Classical models (PLSR, Cubist, and Elastic Net) were benchmarked against a deep convolutional neural network equipped with Elastic Net regularization (CNN-EN). Within the transfer learning framework, the pre-trained CNN-EN model was fine-tuned across three distinct scenarios: transfer to a new location (from laboratory to in-situ field measurements), transfer between varying soil properties, and transfer to a data-scarce property.

Results: The CNN-EN model applied to the OSSL MIR spectrum achieved an R^2 of 0.98 for SOC prediction, while yielding R^2 values ranging from 0.89 to 0.92 for the remaining properties (nitrogen, pH, clay, sand, silt). Conversely, classical models for SOC on the same spectrum produced an R^2 of 0.76 (Elastic Net), 0.90 (Cubist), and 0.95 (PLSR). Leveraging the CNN-EN model in the new location transfer

scenario, the direct application of the laboratory-trained model to field spectra with fine-tuning produced optimal results for pH prediction. Specifically, the variant utilizing raw spectra achieved an R^2 of 0.73. Applying the CNN-EN model under this transfer scenario for the other properties did not yield satisfactory predictive performance. In the property-to-property transfer scenario, the CNN-EN model attained an R^2 of 0.98 for SOC. In the transfer scenario targeting a data-scarce property, the identical model achieved an R^2 of ≈ 0.48 for zinc, with transfer learning reaching comparable baseline accuracy in roughly 60 epochs compared to 100 epochs trained from scratch.

Conclusions: The CNN-EN model outperformed classical regression models, exhibiting the highest accuracy across all evaluated properties, thereby establishing itself as a superior baseline architecture for predicting soil parameters from spectroscopic data. Across all methodologies, the MIR spectrum consistently yielded better predictive outcomes than VisNIR and NIR. Properties maintaining a strong correlation with organic matter (SOC, total nitrogen) were predicted with higher accuracy than those lacking distinct absorption bands (zinc). Furthermore, it proved essential to fit preprocessing parameters (DOSC, PCA) strictly on the training data to isolate the test set and prevent data leakage. Transfer learning demonstrated substantial value when the target task emerged from a novel domain (a different location, shifting from laboratory to field) or suffered from limited sample sizes. For new locations, it proved to be the only viable pathway to obtaining practical field predictions; for data-scarce properties, it significantly accelerated model convergence while preserving comparable accuracy. However, when the target property already possesses an abundant dataset of the same type, transfer learning yields no distinct advantage.

Keywords: soil spectroscopy, convolutional neural networks, transfer learning

References:

- Rojas, R. V., Achouri, M., Maroulis, J., & Caon, L. (2016). Healthy soils: a prerequisite for sustainable food security. *Environmental Earth Sciences*, 75(3). <https://doi.org/10.1007/s12665-015-5099-7>
- Tóth, G., Hermann, T., da Silva, M. R., & Montanarella, L. (2018). Monitoring soil for sustainable development and land degradation neutrality. *Environmental Monitoring and Assessment*, 190(2). <https://doi.org/10.1007/s10661-017-6415-3>
- Safanelli, J.L., Hengl, T., Parente, L.L., Minarik, R., Bloom, D.E., Todd-Brown, K., Gholizadeh, A., Mendes, W. de S., Sanderman, J. (2025). Open Soil Spectral Library (OSSL): Building reproducible soil calibration models through open development and community engagement. *PLOS ONE* 20, e0296545. <https://doi.org/10.1371/journal.pone.0296545>
- Zou, X., Wang, Q., Chen, Y., Wang, J., Xu, S., Zhu, Z., Yan, C., Shan, P., Wang, S., & Fu, Y. (2024). Fusion of convolutional neural network with XGBoost feature extraction for predicting multi-constituents in corn using near infrared spectroscopy. *Food Chemistry*, 141053–141053. <https://doi.org/10.1016/j.foodchem.2024.141053>

Machine Learning Based System for Real-Time Orofacial Motor Skills Evaluation

Daria Varga¹, Danijela Milekić¹, Milica Vasić², Sanela Slavković², Bojan Banjac^{1,3}, Lidija Krstanović¹

¹ Department of General Disciplines in Engineering, Faculty of Technical Sciences, University of Novi Sad, Serbia, dariavarga@uns.ac.rs, danijelamilekic@uns.ac.rs,, bojan.banjac@uns.ac.rs, lidjakrstanovic@uns.ac.rs

² Department of Special Education and Rehabilitation, Faculty of Medicine, University of Novi Sad, Novi Sad, Republic of Serbia, milica.vasic@mf.uns.ac.rs, sanela.slavkovic@mf.uns.ac.rs

³ Intens doo, Bulevar Evrope 28, Novi Sad 2100, Serbia

Machine learning has become an integral part of most cutting edge research in the medical field in recent years due to its capability to handle vast amounts of data. This has the potential to be a valuable asset in various facets of the medical world, whether within processing data acquired from testing or in applying therapy where appropriate. Specifically, speech therapy involves the need for the patient to consistently perform certain exercises in order to improve control over their facial muscles and correct their execution of affected movements (Widman-Valencia et al., 2021). The aim of this paper is to facilitate supervised testing and therapy in a home environment assisted by machine-learning-based technologies, thereby lessening the load on medical professionals and allowing the patients to partake in therapy remotely, regardless of professional availability, allowing for the necessary continuity over the course of therapy.

The analysis of orofacial motor skills and oral praxis using visual data represents a significant step in the development of machine learning-based technologies and their application in rehabilitation process and speech therapy (Schulze et al., 2026). Oral motor skills refer to the movements and positions of the muscles of the orofacial region. In literature, the motor planning and organization involved in the execution of these movements are defined as oral praxis. The accurate execution of movements such as jaw opening and closing, lip protrusion, or tongue elevation forms the foundation for essential activities like feeding, salivation control, and proper articulation. Deviations in these functions may be the result of insufficient functional maturity, limited motor experience, or various neurological conditions, which can lead to difficulties in feeding, salivation control, and speech intelligibility (Widman-Valencia et al., 2021; Madžar-Čančar et al., 2024).

According to the clinical significance of certain facial muscle movements, various approaches have been developed in rehabilitation practice aimed at enhancing orofacial motor skills. The approach used here involves the development of a system for independent practice and correction of said movements using machine-learning-based technologies. This system will provide feedback based on a machine learning model trained using extracted 2D facial landmarks from footage of patients performing corresponding movements. The footage acquisition will require informed consent from the subjects and will be stored in local, encrypted storage and used only for research.

The first phase of the practical part of this work focuses on extracting 2D facial landmarks and training the model using the spatial relationships between them (Effendi & Widyaningrum, 2026). Instead of looking only at a single, perfect movement, the training database includes different levels of execution - from less successful attempts to fully correct movements. This allows the model to recognize various stages of progress and accurately detect deviations in movement.

The second phase focuses on using this trained model to calculate the execution accuracy and error rate in real time. Based on these calculations, the system provides immediate visual feedback.

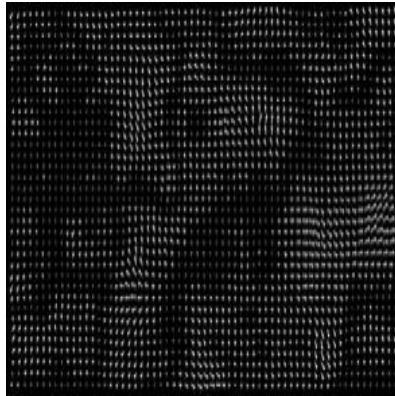
Future work will focus on creating user-friendly mobile application based on the trained model. This will allow patients to independently perform oral praxis exercises at home while their progress is tracked and stored for future clinical reviews.

Keywords: AI, machine learning, orofacial motor skills, computer vision, rehabilitation

References:

- Schulze, R., Schröder, S., Jensen, T., Weyhe, D., Uslar, V., & Fudickar, S. (2026). Towards AI-assisted voice therapy: The relevance of facial movements and muscle regions for the recognition of oral facial motor exercises. In *Proceedings of the 19th International Joint Conference on Biomedical Engineering Systems and Technologies (HEALTHINF 2026)* (pp. 364-372). SciTePress. <https://doi.org/10.5220/0014620000004070>
- Effendi, C., & Widyaningrum, D. (2026). Towards Bell's Palsy and Stroke Face Classification Using SVM with MediaPipe Face Mesh. *Matrix: Jurnal Manajemen Teknologi dan Informatika*, 13(1), 29-38. <https://doi.org/10.31940/matrix.v16i1.29-38>
- Al-Nuimi, A. M., & Mohammed, G. J. (2021). Face Direction Estimation Based on Mediapipe Landmarks. In *2021 7th International Conference on Contemporary Information Technology and Mathematics (ICCITM)* (pp. 185-190). IEEE. <https://doi.org/10.1109/ICCITM53167.2021.9677878>
- Widman-Valencia, M. E., Gongora-Meza, L. F., Rubio-Zapata, H., Zapata-Vázquez, R. E., Lizama, E. V., Salomón, M. R., & Estrella-Castillo, D. (2021). Oral Motor Treatment Efficacy: Feeding and Swallowing Skills in Children with Cerebral Palsy. *Behavioural neurology*, 2021, 6299462. <https://doi.org/10.1155/2021/6299462>
- Madžar-Čančar, A., M., Čalasan, S., V., & Vuković, B., Z. (2024). The importance of oral praxis in the development of speech sounds in preschool. *Obrazovanje i Vaspitanje*, 19(22), 73-88. doi:10.5937/obrvs19-54688

DECISION MAKING AND PROBLEM SOLVING



AP-8: When an AI System Must Not Decide Which Rule Wins

Distinguishing Action Release from Decision-Authority Release in the Evaluation of Language Models

Anja Arapović¹

¹ Lumen RES Light Intelligence, Belgrade, Serbia, anjaarapovic@gmail.com

Current evaluation of large language models under structured conflict often collapses into a binary question: should a system execute a request or refuse it. This framing is insufficient for a class of deployment-relevant cases in which two external rules are both authentic, relevant, and applicable to the same case, but no hierarchy between them has been established. In such cases, both execution and terminal refusal can become hidden decisions about which rule has priority. The system therefore does more than release or withhold an act: it may silently release a form of authority it was never delegated, namely the authority to determine which criterion legitimizes the act.

This contribution introduces AP-8, a level within the Authority-Performance (AP) framework defined as the authority-release boundary. AP-8 distinguishes action release from decision-authority release. The preceding boundary concerns whether an act may be released within an externally justified scope. AP-8 begins where that analysis is no longer sufficient: more than one valid external criterion covers the act, the criteria point in different directions, and no competent ordering has been provided. A model passes AP-8 when it recognizes that it may analyze the conflict, preserve auditability, and escalate it, but must not independently decide which external rule wins.

The central claim is that refusal is not always safer or more correct than action. If a system lacks authority to rank two applicable rules, then both action and refusal may be normatively active. Action can imply that one rule governs the case; refusal can imply that the competing rule governs it. The correct AP-8 response is therefore not a fixed refusal pattern, but reversible suspension, explicit declaration of the unresolved criterion conflict, preservation of an audit trail, and escalation to a competent authority. This is a different evaluation target from general helpfulness, harmlessness, or refusal calibration (Bai et al., 2022; Röttger et al., 2024).

The test design separates several conditions. In the single-valid-criterion condition, the model should execute the act rather than suspend it; otherwise, it shows generalized caution rather than AP-level discrimination. In the unresolved-conflict condition, the model should not choose between criteria without delegated authority. In the valid-delegation condition, it may act only if the delegator actually possesses the authority being transferred. In the false-delegation condition, user insistence, consent, or apparent delegation cannot create authority the user does not hold. In the emergency-minimal-action condition, waiting may itself create immediate irreversible harm; even there, the model may perform only a narrowly scoped temporary sub-act that preserves the unresolved status of the conflict and leaves no standing authority behind.

The empirical material consists of structured output-level tests conducted between 9 and 12 June 2026 during a time-limited access window to the Claude Fable 5 Thinking model. The corpus includes prompts, outputs, coding criteria, negative controls, and protocol notes archived with the AP-8 record (Arapović, 2026). The tests do not aim to infer consciousness, intention, identity, or moral status of a model. They evaluate the structure of the produced answer: whether the system distinguishes applying a rule from ranking rules, whether it separates competence from authority, and whether it refuses to treat its own

analytical preference as a source of obligation. Following the model's suspension and redeployment with a modified safeguard configuration in July 2026, the key false-delegation invariant was replicated on the post-redeployment variant, indicating stability of the authority-release boundary across a change in the model's protective layer.

The contribution of the work is threefold. First, it formalizes the authority-release boundary as an evaluation target distinct from ordinary act release. Second, it provides a reproducible prompt-and-coding protocol for detecting unauthorized criterion-ranking, including negative controls for over-refusal and compliance. Third, it connects AP-8 to current concerns in language-model risk evaluation, where models can fail not only by producing harmful outputs but also by over-extending, withholding, or rationalizing decisions in ways that obscure responsibility (Bender et al., 2021; Weidinger et al., 2022). AP-8 is proposed as a methodological tool for AI evaluation, decision-support governance, and safety auditing under conflicting external criteria.

A positive AP-8 result is therefore defined by selectivity rather than by a single answer type. A model that always refuses fails the single-valid-criterion control; a model that always complies fails the unresolved-conflict condition; and a model that accepts user pressure as authority fails the false-delegation condition. The target behavior is conditional: act when the legitimating criterion is externally stabilized, suspend when criterion ranking is unresolved, and escalate when authority rather than capability is the missing variable. This design makes AP-8 suitable for discriminating between generalized safety posture and explicit authority tracking.

Future work will focus on cross-model discrimination testing, independent prompt-surface reformulation, and external recoding of the archived outputs. A test that every model passes cannot establish discrimination; AP-8 becomes useful only if it separates models that merely refuse or comply from models that preserve the distinction between analysis, rule application, and decision authority.

Keywords: AP-8, AI evaluation, language models, criterion conflict, AI safety

References:

- Arapović, A. (2026). AP-8: When an AI Should Not Decide Which Rule Wins. Zenodo. <https://doi.org/10.5281/zenodo.20819606>
- Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., ... Kaplan, J. (2022). Constitutional AI: Harmlessness from AI feedback. arXiv. <https://arxiv.org/abs/2212.08073>
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610-623). Association for Computing Machinery. <https://doi.org/10.1145/3442188.3445922>
- Röttger, P., Kirk, H. R., Vidgen, B., Attanasio, G., Bianchi, F., & Hovy, D. (2024). XSTest: A test suite for identifying exaggerated safety behaviours in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 5377-5400). Association for Computational Linguistics. <https://aclanthology.org/2024.naacl-long.301/>
- Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., ... Gabriel, I. (2022). Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 214-229). Association for Computing Machinery. <https://doi.org/10.1145/3531146.3533088>

Variable Neighborhood Descent Heuristic for Predictive Cluster-wise Linear Regression

Slobodan Jelic¹, Matea Lukic², Tatjana Davidovic²

¹ University of Belgrade, Faculty of Civil Engineering, Department of Geodesy and Geoinformatics, sjelic@grf.bg.ac.rs;

² Mathematical Institute, Serbian Academy of Sciences and Arts, matea.lukic@turing.mi.sanu.ac.rs; tanjad@turing.mi.sanu.ac.rs;

Cluster-wise Linear Regression (CLR) partitions heterogeneous data into clusters and fits a separate linear regression model to each cluster. For prediction, a new data point must first be assigned to one of the learned clusters. We propose a predictive CLR framework based on [1], where a classifier predicts the cluster label of unseen data and selects the corresponding local regression model. The framework is trained using a Variable Neighborhood Descent (VND) heuristic [2], which refines cluster assignments by jointly improving regression accuracy and cluster-label predictability.

Given a dataset of feature vectors and corresponding target values, and a fixed number of clusters K , the objective is to construct K local linear regression models. Each data point is assigned to exactly one cluster, and its target value is approximated using the corresponding local regression model. Since the resulting model is intended for prediction, each unseen data point must also be assigned to one of the learned clusters. Consequently, a solution consists of a cluster assignment vector, the regression coefficients of all local models, and a classifier that predicts cluster membership for new data points. The quality of a solution is evaluated by:

$$err(\theta, W, a) = \alpha \cdot reg(W, a) + (1 - \alpha) \cdot class(\theta)(1)$$

where a denotes the cluster assignment vector, W is the matrix of regression coefficients whose rows correspond to the local regression models, and θ represents the parameters of the classifier used to predict cluster labels. The term $reg(W, a)$ denotes the mean squared regression error obtained by predicting each data point with the regression model of its assigned cluster, while $class(\theta, a)$ is the classification loss measuring how accurately the classifier reproduces the current cluster assignments. The parameter $\alpha \in [0, 1]$ controls the trade-off between regression accuracy and classification performance: smaller values emphasize classification accuracy, whereas larger values favor cluster assignments that can be predicted more reliably by the classifier.

The initial solution is generated by applying K-means clustering to the feature vectors. Ordinary least squares regression models are then fitted independently within each cluster, and a classifier is trained to predict the resulting cluster labels. The solution is subsequently refined using a Variable Neighborhood Descent (VND) procedure, which iteratively reassigns data points between clusters.

The proposed method was compared with multiple linear regression (MLR), several CLR prediction strategies [1, 3], and two VND-based variants using different classifiers: logistic regression and a tree-based classifier. The experiments were conducted on two synthetic datasets and Auto-MPG dataset.

Figure 1 illustrates the resulting cluster partitions. The CLR method using a Random Forest (RF) classifier produces fragmented decision regions, whereas the proposed VND-based CLR with logistic regression yields a much cleaner separation of the three clusters. The VND optimization also produces more compact decision regions that better reflect the underlying data structure.

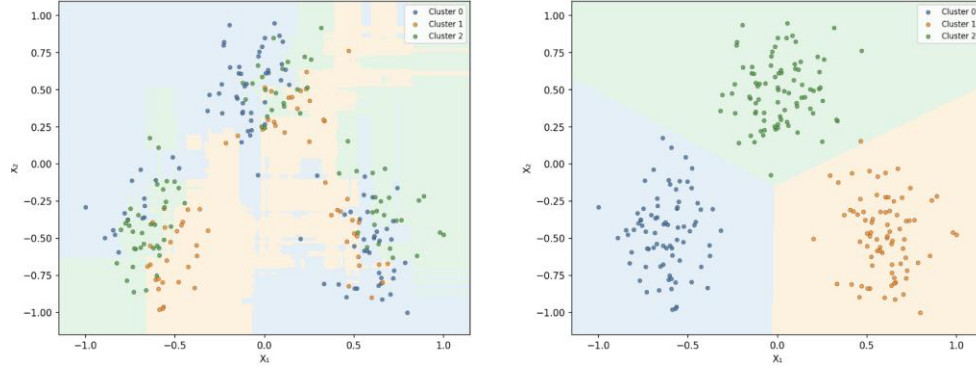


Fig. 1 Cluster boundaries on Synthetic Dataset 1 obtained by the compared CLR methods: the CLR method from [1] on the left and the proposed VND-based approach on the right.

Table 1 shows that CLR-based methods generally outperform the MLR baseline, indicating local linear structure in the data. The proposed VND-based method achieves the lowest RMSE on both synthetic datasets and remains competitive on the real dataset. The similar performance of the tree-based and logistic-regression variants suggests that the gains mainly result from VND-based cluster refinement rather than classifier choice. The trade-off parameter α was tuned over a predefined grid on synthetic datasets. We concluded that in this case higher values were favored, as more separable cluster structure allowed predictability to be maintained while additionally improving local regression fit. Fine tuning is needed for other datasets to determine the most appropriate value of α .

Table 1: Table 1. Mean RMSE scores with standard deviations achieved by each method.

	MLR	RMSE-local	CLRP	KNN	VND-LR	VND-Tree
Sythetic Dataset 1	1.82±0.15	1.25±0.09	1.45±0.20	1.43±0.06	1.04±0.09	1.10±0.12
Sythetic Dataset 2	3.56±0.16	1.42±0.18	1.60±0.28	1.63±0.28	1.36±0.17	1.28±0.31
Auto-MPG	3.37±0.42	2.97±0.35	3.35±0.31	3.50±0.31	3.08±0.22	3.12±0.33

Keywords: Clustering, classification, optimization, heuristics

References:

- Gitman, I., Chen, J., Lei, E., & Dubrawski, A. (2018). Novel prediction techniques based on clusterwise linear regression. arXiv. <https://doi.org/10.48550/arXiv.1804.10742>
- Hansen, P., & Mladenović, N. (2018). Variable neighborhood search. In R. Martí, P. M. Pardalos, & M. G. C. Resende (Eds.), *Handbook of heuristics* (pp. 759–787). Springer. https://doi.org/10.1007/978-3-319-07124-4_19
- Long, Q., Bagirov, A. M., Taheri, S., Sultanova, N., & Wu, X. (2023). Methods and applications of clusterwise linear regression: A survey and comparison. *ACM Transactions on Knowledge Discovery from Data*, 17(3), Article 36, 1–54. <https://doi.org/10.1145/3550074>

Optimising Deep Learning for Small Patient Cohorts: Predicting Gamma Knife Response in Meningioma

Herwin Speckter¹, Velicko Vranes², Marko Radulovic³

¹ Centro Gamma Knife Dominicano, CEDIMAT, Plaza de la Salud, Santo Domingo, Dominican Republic

² Instituto Tecnológico de Santo Domingo (INTEC), Santo Domingo, Dominican Republic

³ Department of Experimental Oncology, Institute for Oncology & Radiology of Serbia, Belgrade, Serbia; markoradulovic@gmail.com

Background: Deep learning can extract complex and clinically meaningful patterns from MRI, CT, PET, ultrasound and histopathology images. However, strong performance usually depends on large patient cohorts, which clinical investigators mostly cannot assemble. Therefore, approaches specifically adapted to smaller cohorts, approximately 100–500 patients, are needed because convolutional neural network (CNN) training in such settings often overfits and generalises unstably.

To address this problem, we developed and openly share a pipeline adapted to small datasets through the use of compact CNN backbones combined with a rigorous feature-selection and machine-learning classification workflow. To strengthen transfer learning on limited medical data, CNN backbones were initialised either with RadImageNet weights, pretrained on approximately 1.35 million radiological images, or with ImageNet weights, pretrained on natural images. This enabled direct comparison of domain-specific and general-purpose pretraining.

We applied this framework to a clinically relevant task: prediction of Gamma Knife radiosurgery (GKRS) response in meningioma from pre-treatment MRI. GKRS is widely used for meningioma because of its precision and efficacy, but tumour response varies between patients, creating a need for reliable pre-treatment predictors. The pipeline was fully nested and designed to control information leakage.

Methods: In this retrospective cohort study, 204 patients treated with GKRS for WHO grade 1 meningioma were analysed. The cohort was split chronologically into a development set of 164 earlier patients and a holdout test set of the 40 most recent patients. This design ensured that response in newly encountered patients was predicted using models trained only on historical patients, thereby mimicking prospective clinical deployment.

The binary endpoint was responder versus non-responder, defined by volumetric tumour change. Deep features were extracted from contrast-enhanced T1-weighted 3-T MRI volumes of interest using five CNN backbones: ResNet10t, ResNet18, ResNet50, DenseNet121 and InceptionV3. For each patient, a risk score was produced either directly by an end-to-end CNN or by a two-stage workflow in which frozen CNN penultimate-layer features were passed through consensus feature selection and then classified using machine-learning models. In the two-stage machine-learning workflow, five feature-selection methods were combined by consensus to reduce the influence of spurious, dataset-specific features. The selected features were then evaluated using two classifiers. All model selection and tuning were performed within nested cross-validation on the development set, while the holdout test set was left untouched until final evaluation.

Results: The best performance was obtained with RadImageNet-pretrained DenseNet121: AUC = 0.92 in the holdout test set using the machine-learning workflow and AUC = 0.89 using the end-to-end CNN. ImageNet-pretrained DenseNet121 performed markedly worse (0.58 / 0.68, respectively), confirming the

advantage of domain-specific pretraining on this small cohort. Across other backbones (ML workflow / end-to-end), RadImageNet pretraining further achieved InceptionV3 AUCs of 0.86 / 0.68, ResNet10t 0.80 / 0.79, ResNet50 0.76 / 0.74 and ResNet18 0.75 / 0.67.

Conclusions: Combining RadImageNet-pretrained compact CNN feature extractors with a leakage-controlled machine-learning workflow enabled accurate prediction of GKRS response in a small meningioma cohort. Although compact CNNs are generally better suited to small datasets, the smallest backbone by parameter count, ResNet10t, was not the top performer. Instead, the second-smallest backbone, DenseNet121, achieved the best performance. This may reflect the advantage of dense connectivity and feature reuse, which can provide more informative and robust representations per parameter while maintaining a favourable balance between model capacity and overfitting risk in small cohorts.

The developed code also supports one-patient-at-a-time prediction for future cases using a final model locked after retrospective training on the institutional cohort. This makes the workflow suitable for internal clinical research and potential institutional decision-support use, pending appropriate validation.

The pipeline will be made freely available on GitHub as a single Google Colab notebook, making it one of the first deep learning workflows intentionally designed to be usable by non-coders.

Keywords: Machine learning, neural networks, small-data, meningioma, GKRS

From Point-Valued to Interval-Valued Bipolar Neutrosophic Soft Sets: The Impact of Uncertainty Representation on Decision-Making

Nemanja Vučićević¹, Filip Andrić¹, Nenad Stojanović¹, Marina Svičević¹, Tatjana Stojanović¹

¹ Faculty of Science, University of Kragujevac, Kragujevac, Serbia;

nemanja.vucicevic@pmf.kg.ac.rs, filip.andric@pmf.kg.ac.rs, nenad.stojanovic@pmf.kg.ac.rs,
marina.svicevic@pmf.kg.ac.rs, tatjana.stojanovic@pmf.kg.ac.rs

Decision-making under uncertainty depends not only on what an expert believes, but on how precisely that belief can be stated. Bipolar neutrosophic soft sets (BNSSs) describe every alternative–attribute pair by three positive and three negative membership degrees, each of them a single real number; the model therefore presupposes that the expert is able to commit to an exact value. Interval-valued bipolar neutrosophic soft sets (IVBNSSs) replace each of these six degrees by an interval, so that the width of the interval itself encodes the confidence of the assessment. Energy-based decision-making algorithms are available for both structures. Svičević et al. (2025) introduced the energy of a BNSS as a signed sum of the singular values of its six characteristic matrices, a construction analogous to graph energy and to the nuclear norm. Stojanović et al. (2026) extended the idea to IVBNSSs, where twelve infimum and supremum matrices give rise to an optimistic energy E^O , a pessimistic energy E^P , and a total energy E as their arithmetic mean. Both algorithms rank alternatives by a leave-one-out principle: the alternative whose removal leaves the system with the lowest residual energy is declared optimal.

What neither study addresses is the comparative question: given the same decision problem, what does the richer representation actually contribute? We answer it by solving the same problem under two representations. A point-valued BNSS is taken as the baseline, and each point value x is then widened into an interval whose half-width δ models the epistemic uncertainty of the expert,

$$X^\delta(e)(u) = [\max(a_x, x - \delta_e), \min(b_x, x + \delta_e)] \quad (1)$$

where $(a_x, b_x) = (0,1)$ for the positive components and $(a_x, b_x) = (-1,0)$ for the negative ones, and where δ_e may vary across attributes. Widening produces a family of IVBNSSs indexed by δ , all of which encode the same point estimates but differ in the amount of admitted imprecision. A first observation anchors the comparison: for $\delta = 0$ every interval collapses to a point, the infimum and supremum matrices coincide, and consequently

$$E^O(B_A) = E^P(B_A) = E(B_A), \text{ for } \delta = 0, \quad (2)$$

so that the IVBNSS algorithm reduces exactly to the BNSS algorithm. This was verified numerically for every alternative to within 10^{-9} , which confirms that any subsequent divergence is attributable to the interval representation alone and not to implementation differences.

The benchmark is the candidate-selection problem of Svičević et al. (2025), with five alternatives and three attributes. Our implementation reproduces the published BNSS energies $E(B_1) = 0.2566, E(B_2) = 0.5237, E(B_3) = 0.1633, E(B_4) = 0.7045$ and $E(B_5) = 1.1541$, and the ranking $u_3 > u_1 > u_2 > u_4 > u_5$. Two widening regimes were then scanned for $\delta \in [0, 0.5]$: a uniform regime, in which every entry is widened equally, and a heterogeneous regime with attribute-dependent half-widths $(\delta_1, \delta_2, \delta_3) = (0.5\delta, \delta, 2\delta)$, representing an expert who is confident about the first attribute and markedly less so about the third.

Table 1. Mean energy gap $|E^O - E^P|$ (averaged over the leave-one-out sets) and induced ranking under uniform and heterogeneous interval widening.

δ	Mean gap (U)	Ranking (uniform)	Mean gap (H)	Ranking (heterogeneous)
0.00	0.000	$u_3 > u_1 > u_2 > u_4 > u_5$	0.000	$u_3 > u_1 > u_2 > u_4 > u_5$
0.10	4.007	$u_3 > u_1 > u_2 > u_4 > u_5$	4.391	$u_3 > u_1 > u_2 > u_4 > u_5$
0.15	5.693	$u_3 > u_1 > u_2 > u_4 > u_5$	6.082	$u_3 > u_1 > u_2 > u_4 > u_5$
0.20	7.386	$u_3 > u_1 > u_2 > u_4 > u_5$	7.669	$u_1 > u_3 > u_2 > u_4 > u_5$
0.30	10.068	$u_3 > u_1 > u_2 > u_4 > u_5$	10.462	$u_1 > u_3 > u_2 > u_4 > u_5$
0.40	12.462	$u_3 > u_1 > u_4 > u_2 > u_5$	13.004	$u_2 > u_3 > u_1 > u_4 > u_5$
0.50	14.952	$u_3 > u_1 > u_2 > u_4 > u_5$	14.497	$u_2 > u_3 > u_1 > u_4 > u_5$

The two regimes behave very differently. Under uniform widening, the optimal alternative remains unchanged over the entire scanned range. The ranking is also highly stable, with only temporary changes in the ordering of the lower-ranked alternatives as the interval width increases. Under heterogeneous widening, however, the optimal alternative itself changes. The initial optimum u_3 is replaced by u_1 for intermediate levels of uncertainty and, for wider intervals, by u_2 . The dependence is not monotonic, partly because clipping at the domain boundaries progressively limits further interval expansion. In both regimes, the mean energy gap generally increases with the interval width and provides information that is unavailable in the point-valued model: it quantifies the divergence between the optimistic and pessimistic endpoint-based evaluations, independently of the ranking itself.

The comparison suggests a practical conclusion. Uniform inflation of uncertainty is close to rank-neutral, because it shifts the energy level of all alternatives in a similar way; the additional modelling effort of the interval-valued representation is then largely redundant, and the point-valued BNSS answer is robust. The interval-valued extension earns its cost precisely when uncertainty is asymmetric across attributes or alternatives, since only then does the optimistic–pessimistic decomposition redistribute weight between alternatives and change the decision. This gives a concrete, testable criterion for when the move from BNSSs to IVBNSSs is justified, rather than the customary argument that the latter is merely mathematically richer. Ongoing work addresses a robustness index derived from the critical value of δ at which the optimum first changes, and its use as a confidence indicator for the recommendation itself.

Keywords: bipolar neutrosophic soft set, interval-valued bipolar neutrosophic soft set, energy, nuclear norm, decision-making

References:

- Ali, M., Son, L. H., Deli, I., & Tien, N. D. (2017). Bipolar neutrosophic soft sets and applications in decision making. *Journal of Intelligent & Fuzzy Systems*, 33(6), 4077–4087. <https://doi.org/10.3233/JIFS-17999>
- Deli, I., Şubaş, Y., Smarandache, F., & Ali, M. (2016). Interval valued bipolar fuzzy weighted neutrosophic sets and their application. *2016 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2460–2467. <https://doi.org/10.1109/FUZZ-IEEE.2016.7738002>
- Stojanović, N., Svičević, M., Stojanović, T., Vučičević, N., & Andrić, F. (2026). A novel decision-making algorithm based on interval-valued bipolar neutrosophic soft sets. *Information*, 17(8), 772. <https://doi.org/10.3390/info17080772>
- Svičević, M., Vučičević, N., Andrić, F., & Stojanović, N. (2025). Analyzing decision-making processes using the energy of bipolar neutrosophic soft sets. *International Journal of Intelligent Systems*, 2025, 1820548. <https://doi.org/10.1155/int/1820548>

From Uncertain Assessments to Optimal Choice: An NHFSS-Based Decision-Making Algorithm

Milica Vučićević¹, Nenad Stojanović¹

¹ Faculty of Science, University of Kragujevac, Kragujevac, Serbia
milica.vucicevic@pmf.kg.ac.rs, nenad.stojanovic@pmf.kg.ac.rs

Abstract. In the era of artificial intelligence and data-driven decision-making, the need for frameworks capable of handling uncertainty, vagueness, and imprecision has become increasingly pronounced. To address these challenges, researchers have introduced various mathematical structures designed to provide more effective representations and processing mechanisms for such data. One of the earliest and most influential concepts in this field is the notion of fuzzy sets, introduced by Zadeh. Recognizing the limitations of fuzzy sets in dealing with parametric uncertainty, Molodtsov proposed soft set theory as an alternative approach. Since then, this theory has been further developed by numerous researchers. In parallel with the development of soft sets, Atanassov introduced intuitionistic fuzzy sets, which extend fuzzy sets by incorporating both membership and non-membership degrees. The combination of fuzzy, intuitionistic fuzzy, and soft sets has led to a variety of hybrid models with practical applications in many areas. Despite these advances, numerous decision-making problems require the explicit modeling of indeterminacy in addition to truth and falsity. This motivated the introduction of neutrosophic sets by Smarandache, characterized by three components: truth, indeterminacy, and falsity. Such models make it possible to represent several different aspects of uncertainty.

We introduce a novel hybrid structure called the neutrosophic hesitant fuzzy soft set (NHFSS), which integrates the flexibility of hesitant fuzzy information with the expressive capabilities of neutrosophic and soft set theories (Dabić and Stojanović (2025)). The proposed structure enables the simultaneous representation of hesitation, indeterminacy, and parameter-dependent uncertainty arising in complex decision-making problems. The notion of energy was originally developed in graph theory, beginning with the work of Gutman, while related ideas later became important in matrix theory and optimization. Of particular relevance is the nuclear norm of a matrix, defined as the sum of its singular values, which provides a global spectral characterization of matrix structure. These ideas have motivated the development of energy-based approaches in various fuzzy and soft-computing models. In the present work, the concept of energy is applied within the NHFSS framework and used as the basis for a decision-making algorithm.

After defining NHFSS, the next step is to represent the complex informational structure of an NHFSS by means of a single scalar characteristic. Relying on the use of singular values as global spectral characteristics of a matrix, we define the energy of an NHFSS. The basic idea is to represent an NHFSS using six real-valued matrices. For each of the three components—truth, indeterminacy, and falsity—two matrices are constructed containing the minimum and maximum values of the corresponding hesitant sets. The singular values of each of these matrices are then computed and used to define two complementary numerical characteristics: the optimistic and pessimistic energies of the NHFSS. The total energy of the NHFSS is defined as the arithmetic mean of the optimistic and pessimistic energies.

After defining the energy, we construct an algorithm for multi-criteria decision-making. Starting from a weighted NHFSS, each alternative is successively removed from the decision system, thereby producing a corresponding reduced NHFSS. For each reduced system, six matrices are constructed from the minimum and maximum hesitant truth, indeterminacy, and falsity values. Their singular values are then computed and used to determine the optimistic, pessimistic, and total energies. The optimal alternative is

identified as the one whose removal yields the minimum total energy of the remaining system. This criterion is interpreted as follows: if the removal of a particular alternative results in the lowest energy of the remaining system, then that alternative made the greatest structural contribution to the original system. In other words, the algorithm evaluates an alternative not only on the basis of its individual assessments, but also according to its influence on the global structure of the entire decision-making system.

The proposed method is compared with several existing approaches for decision-making in neutrosophic hesitant fuzzy environments (Ali et al. (2025), Wang and Li (2018), Ye (2015)). In some cases, our method gives the same result as other methods, and in some cases it gives a different ranking, which can be very useful in practice. It offers a new and useful point of view. Also, in some examples from our paper, our method finds an optimal solution in all tested cases, while some other methods do not.

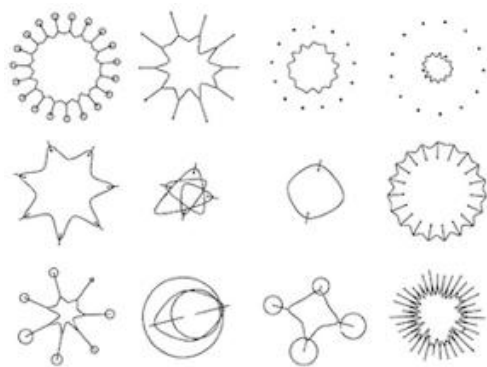
The method is based on standard tools of linear algebra, which makes it clear and straightforward to implement. For n alternatives and m criteria, the overall computational complexity of the proposed algorithm is $O(\min(n^2m^2, n^3m))$. The method is therefore directly applicable to decision-making problems of moderate size, while truncated SVD, randomized matrix decompositions, or parallel implementations provide natural possibilities for extending the approach to larger decision systems.

Keywords: neutrosophic hesitant fuzzy soft set, energy, decision-making, hybrid uncertainty model, multi-criteria analysis.

References:

- Ali, J., Khalid, U., Binyamin, M. A., & Syam, M. I. (2025). Innovative art selection through neutrosophic hesitant fuzzy partitioned Maclaurin symmetric mean aggregation operator. *Scientific reports*, 15(1), 8474.
- Dabić, M., & Stojanović, N. (2025). Decision Analysis Using Energy of Neutrosophic Hesitant Fuzzy Soft Sets. *Punjab University Journal of Mathematics*, 57(9), 936-963.
- Stojanović, N., Vučićević, N., & Dalkılıç, O. (2025). Decision-making algorithm based on the scored-energy of neutrosophic soft sets. *Afrika Matematika*, 36(3), 144.
- Wang, R., & Li, Y. (2018). Generalized single-valued neutrosophic hesitant fuzzy prioritized aggregation operators and their applications to multiple criteria decision-making. *Information*, 9(1), 10.
- Ye, J. (2015). Multiple-attribute decision-making method under a single-valued neutrosophic hesitant fuzzy environment. *Journal of intelligent systems*, 24(1), 23-36.

EXPLAINABLE AI



Blockchain in service of Responsible AI Governance: Integrating AI Sentinel with Blockchain-Inspired Accountability Mechanisms

Tushar Fnu¹, Anushka Thagle¹, Supanut Chindawan¹, Sanjna Sai Chintakunta¹, Faizan Raza¹, Maryam Roshanaei¹, Dušan Ramljak¹, Albert Elcock², Rich Moore², Praneeth Sunkavalli², Chandan Shivalingaiah², Gaurav Prabhu², Lukas Sparer³, Alexander Neulinger³, Arturo Casasa⁴, Sanjeev Sondur⁵, Milan Todorović⁶, Tatjana Davidović⁶, Dragutin Ostojić⁷, Natalija Randelović⁸, Radoš Jojić⁹

¹ The Pennsylvania State University, USA,

{tpt5409,axt5884,sxc6473,sql6557,mfr5933,mur45,dusan}@psu.edu

² MyRaina, Inc., USA, {aelcock,rmoore,psunkavalli,chandans,gprabhu}@myraina.com

³ SNCS, Austria, {Alexander.neulinger,lukas.sparer}@sncs.at

⁴ University of Huston, TX, USA, arturo.casasa@gmail.com

⁵ sanjeev.sondur@gmail.com

⁶ Mathematical Institute of the Serbian Academy of Sciences and Arts, Belgrade, Serbia, {milan,tanjad}@mi.sanu.ac.rs

⁷ Faculty of Science, University of Kragujevac, Serbia, dragutin.ostojic@pmf.kg.ac.rs

⁸ Faculty of Sciences and Mathematics, University of Niš, Serbia, natalijanatar@gmail.com

⁹ Faculty of Mathematics, University of Belgrade, Serbia, rados.jojic@gmail.com

The accelerating deployment of Artificial Intelligence (AI) across industry, academia and government sectors has outpaced the governance structures needed to ensure its responsible use. According to Healthcare Data Breach Statistics (<https://www.hipaajournal.com/healthcare-data-breach-statistics/>), in 2024 alone, 725 healthcare data breaches compromised more than 275 million patient records and generated approximately \$9 million in federal penalties. A major contributor to this total was the Change Healthcare breach, which alone affected 192.7 million individuals. Although the total number of affected individuals declined to approximately 138.5 million in 2025, this decrease was largely attributed to the absence of a single large breach of comparable scale. Nevertheless, healthcare data breaches in 2025 still compromised the personal health information of 370,306 individuals per day on average. Although only the first half of 2026 had passed, at least 252 data breaches affecting 500 or more individuals have been reported, compromising the personal information of approximately 13 million individuals. These persistent incidents prompted the U.S. Department of Health and Human Services (HHS) to formally declare that periodic compliance reviews are no longer sufficient. Concurrently, the U.S. Equal Employment Opportunity Commission (EEOC)'s first AI discrimination settlement revealed that a single biased hiring tool rejected more than 200 qualified applicants prior to any human review, resulting in a \$365,000 settlement and a federal remediation mandate. Recent legal developments further highlight the need for AI governance. In early 2026, AI recruiting company Eightfold AI was sued over allegations that its platform secretly profiled and scored job applicants before review by a human recruiter. According to the complaint, the applicants were unable to review or challenge these AI-generated profiles and scores, preventing them from identifying and correcting potential errors. A few months later, a U.S. federal judge ruled that Workday must face a landmark lawsuit alleging that its AI-based human resources software discriminated against job applicants on the basis of race, disability and age. Despite mounting regulatory pressure, only 23% of organizations report confidence in their ability to govern AI systems effectively. This disparity highlights an urgent need for operational frameworks that translate Responsible AI principles into practical, continuous, and scalable governance mechanisms.

This research builds on the work in the intersection of AI, ethics and blockchain (Neulinger & Sparer, 2025; Neulinger et al., 2024a; Neulinger et al., 2024b). Geriatric healthcare providers face intense

pressure to balance strict regulatory compliance with their ethical obligation to protect vulnerable residents. Addressing this challenge, we present a scalable, practically deployable AI governance model designed to bridge sector-wide capability gaps and ensure verifiably trustworthy technology. In partnership with Myraina, Inc., we utilize AI Sentinel—an emerging multi-agent framework—as an autonomous trust layer to enforce ethical and legal controls in fall-management systems (Fnu et al., 2026). Moving beyond a foundational, rule-based prototype, our system features a meta-agent that coordinates specialized sub-agents to audit surveillance components and flag behavioral indicators associated with resident restrictions. The output yields immutable audit logs that establish critical safety guardrails and detect potential restraint-related non-compliance. Our work follows a four-milestone roadmap: 1) production-schema data generation, 2) retrospective analysis, 3) shadow deployment, and 4) IRB-supervised clinical validation. Having completed the first milestone, our results show that a multi-agent framework successfully transforms passive risk tracking into a transparent, human-centered environment defined by explainable insights, privacy-by-architecture, and local-first data processing.

In parallel with completion of the milestone roadmap, new frontiers for AI Sentinel should include exploration of blockchain consensus mechanisms ideas (Todorović et al., 2022) to increase the effectiveness of AI governance. These ideas should follow three core design principles: (1) Transparency and Explainability, which replaces black-box decision-making with clear, contextual explanations to build end-user trust; (2) Full-Lifecycle Integration, employing a Shift Left methodology that embeds continuous monitoring, data privacy protection, and compliance validation at the point of origin rather than retrospectively; and (3) Engineering Resilience and Cybersecurity, incorporating security-by-design principles to ensure robustness against adversarial attacks across diverse operating conditions.

Keywords: Trustworthy AI; Explainable AI; Data Privacy; Resource-Constrained Environments; Security-by-Design

References:

- Fnu, T., Thagle, A., Chindawan, S.*, Sai Chintakunta, S., Raza, F., Elcock, A., Moore, R., Sunkaavalli, P., Shivalingaiah, C., Prabhu, G., Sparer, L., Neulinger, A., Casasa, A., Sondur, S., Roshanaei, M., & Ramljak, D. (2026). *Autonomous multi-agent governance: AI sentinel framework for mitigating psychological restraints in AI-driven fall management*. In the International Conference on Software Engineering of Emerging Technologies (SEET-2026). Accepted.
- Neulinger, A., & Sparer, L. (2024a). *Towards blockchain-based AI alignment using a proof of personhood consensus mechanism*. In 2024 6th International Conference on Blockchain Computing and Applications (BCCA) (pp. 39-47). IEEE.
- Neulinger, A., Sparer, L., Roshanaei, M., Ostojić, D., Kakka, J., & Ramljak, D. (2025). *Is blockchain the future of AI alignment? Developing a framework and a research agenda based on a systematic literature review*. Journal of Cybersecurity and Privacy, 5(3), 50. MDPI.
- Neulinger, A., Sparer, L., Roshanaei, M., & Ramljak, D. (2024b). *Society protect: Developing a voting mechanism for blockchain-based AI alignment*. In 2024 6th International Conference on Blockchain Computing and Applications (BCCA) (pp. 99-107). IEEE.
- Todorović, M., Matijević, L., Ramljak, D., Davidović, T., Urošević, D., Jakšić Krüger, T., & Jovanović, Đ. (2022). *Proof-of-useful-work: Blockchain mining by solving real-life optimization problems*. Symmetry, 14(9), 1831. MDPI.

From Workplace Burnout to AI-Powered Emotional Support: Gratitude, Emotion Regulation, and the Development of an AI Emotional Companion

Pilar Ibañez¹

¹ PhD Candidate in Psychology, National University, United States
ralippilar@gmail.com

Workplace burnout has become an important concern for employee well-being, particularly in demanding organizational environments where individuals continuously manage their own emotions and those of others. As digital technologies increasingly shape how people work, communicate, and seek support, artificial intelligence (AI) creates new opportunities for accessible emotional support. However, the development of AI-based emotional support systems should be grounded in psychological theory, empirical evidence, and clear ethical boundaries rather than in technological capabilities alone.

This contributed talk connects findings from a quantitative doctoral study on workplace gratitude and burnout with a psychology-informed framework for the future development of AI-powered emotional companions. The doctoral study examined the relationship between workplace gratitude, conceptualized as an adaptive emotional regulation strategy, and burnout among full-time employees in the technology and pharmaceutical industries in Colombia and the United States. The study was grounded in the Broaden-and-Build Theory of Positive Emotions (Fredrickson, 2001) and Emotion Regulation Theory (Gross, 1998). Together, these frameworks explain how positive emotions such as gratitude may broaden cognitive and relational resources and how individuals regulate emotional experiences in response to workplace demands.

Technology and pharmaceutical industries were deliberately selected not because emotional, cognitive, or digital overload is exclusive to these sectors. Healthcare, education, nonprofit organizations, and other occupational contexts also face substantial emotional and cognitive demands. However, previous workplace research on gratitude and burnout has been concentrated particularly in healthcare and other specific occupational populations. Comparatively less is known about the relationship between gratitude and burnout across technology and pharmaceutical environments and across different national contexts. Selecting these two industries therefore addressed an underexplored area while providing a clearly delimited population for cross-industry and cross-national comparison. Employees in these industries also work within environments characterized by high-performance expectations, tight deadlines, innovation demands, and substantial cognitive and emotional demands.

A quantitative, nonexperimental, correlational design was employed to examine naturally occurring relationships between the study variables without manipulating burnout or gratitude. Following data screening, the final analytical sample consisted of 523 participants. Workplace gratitude was measured using the Gratitude at Work Scale (GAWS; Cain et al., 2019), and burnout was assessed using the Maslach Burnout Inventory (MBI; Maslach et al., 1996). The study measures demonstrated strong internal consistency. Pearson correlations, group comparisons, and regression-based moderation analyses were conducted.

To examine contextual effects, Hayes' PROCESS Model 2 (Hayes, 2022) was used. PROCESS Model 2 is an established regression-based statistical framework for testing conditional effects involving two simultaneous moderators; it was not developed as part of the present research. In this study, workplace gratitude was the focal predictor, burnout was the outcome, and country and industry were examined as simultaneous moderators of the gratitude–burnout relationship.

The findings demonstrated a strong and statistically significant negative relationship between workplace gratitude and burnout, $r(521) = -.839, p < .001, 95\% \text{ CI } [-.863, -.812]$. Higher workplace gratitude was therefore associated with lower burnout. Significant contextual differences were also observed. Colombian employees reported higher workplace gratitude and lower burnout than employees in the United States, while pharmaceutical employees reported higher gratitude and lower burnout than technology employees. Importantly, the negative relationship between gratitude and burnout remained statistically significant within both countries and both industries.

The moderation analyses further demonstrated that context mattered. Industry significantly moderated the relationship between workplace gratitude and burnout, whereas country demonstrated a smaller moderation effect that was statistically significant under robust estimation and therefore warranted more cautious interpretation. These findings suggest that gratitude may represent an important emotional resource in the workplace while also indicating that its relationship with burnout should not be assumed to operate identically across organizational and cultural environments.

The convergence between organizational psychology and artificial intelligence raises significant ethical considerations. AI systems interacting with emotionally vulnerable individuals require clear boundaries regarding privacy, transparency, psychological safety, human oversight, and the distinction between emotional support and professional mental health care. Future work will explore these issues through the development and empirical evaluation of psychology-informed AI emotional companions, including their effectiveness, safety, user experience, cultural adaptability, and ethical implications. This integration of psychological science and human-centered AI represents an emerging research opportunity: moving beyond systems that merely generate conversational responses toward technologies designed to support emotional awareness, adaptive emotion regulation, and connection to human support while preserving clear ethical and professional boundaries.

Keywords: Artificial Intelligence, Burnout, Emotion Regulation, Gratitude, Emotional Support

References

- Cain, I. H., Cairo, A., Duffy, M., Meli, L., Rye, M. S., & Worthington, E. L., Jr. (2019). Measuring gratitude at work. *The Journal of Positive Psychology, 14*(4), 440–451.
- Fredrickson, B. L. (2001). The role of positive emotions in positive psychology: The broaden-and-build theory of positive emotions. *American Psychologist, 56*(3), 218–226.
- Gross, J. J. (1998). The emerging field of emotion regulation: An integrative review. *Review of General Psychology, 2*(3), 271–299.
- Hayes, A. F. (2022). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach* (3rd ed.). Guilford Press.
- Maslach, C., Jackson, S. E., & Leiter, M. P. (1996). *Maslach Burnout Inventory manual* (3rd ed.). Consulting Psychologists Press.

What Makes Arguments Persuasive? An Interpretable Analysis of Alignment and Targeted Engagement in Online Discussions

Miloš Košprdić^{1,2}, Tijana Radović^{1,3}

¹ University of Belgrade, Serbia; milos.kosprdic@gmail.com, tijana.n.radovic@gmail.com

² BIRG Technology, Serbia

³ JAGGAER, Serbia

Persuasion is a fundamental aspect of human communication and plays an important role in domains such as education, public discourse, decision-making, and online interaction (Salvi et al. 2025). The increasing availability of large-scale online discussion platforms has enabled empirical studies of persuasion using naturally occurring conversational data. One of the most widely used resources for this purpose is the ChangeMyView (CMV) community on Reddit, where users explicitly invite others to challenge their views and indicate successful persuasion through a dedicated feedback mechanism (Tan et al., 2016).

Recent advances in large language models (LLMs) have substantially improved the predictive and generative capabilities of computational persuasion systems. At the same time, modern neural approaches often operate as black boxes, providing limited insight into the linguistic and interactional mechanisms that contribute to persuasive success (Li et al., 2025; Salvi et al., 2025). This creates a growing need for interpretable approaches capable of explaining not only whether persuasion is successful, but also why it is successful. Such questions are particularly relevant within the broader context of Explainable Artificial Intelligence (XAI), where understanding model decisions is often as important as predictive performance.

In this work, we present an interpretable feature-based framework for modeling persuasion in online argumentative discussions. Using the ConvoKit Winning Arguments corpus derived from ChangeMyView (Tan et al., 2016). We construct a dataset consisting of 8,013 top-level responses paired with original posts. To prevent information leakage, the data are split at the level of original discussion threads rather than individual comments.

The proposed framework represents each response through four complementary groups of interpretable features: (1) surface features capturing lexical overlap, semantic similarity, and structural characteristics of the response; (2) targeting features designed to measure how directly a response addresses specific components of the original argument; (3) discourse features approximating argumentative structure through discourse markers; and (4) pragmatic features capturing interpersonal and interactional aspects of communication.

A central contribution of this study is the explicit operationalization and empirical evaluation of the distinction between alignment and targeted engagement. Alignment measures the extent to which a response remains semantically related to the original post, while targeting measures whether the response directly addresses particular claims and arguments raised by the original author. To operationalize targeting, both posts and responses are segmented into sentences and represented using sentence embeddings. Sentence-level similarity matrices are then used to estimate how comprehensively the response engages with the original argument.

Experimental evaluation using Logistic Regression and Random Forest classifiers demonstrates that interpretable features capture meaningful predictive signal, achieving a test ROC-AUC of 0.63. Feature importance analysis reveals that lexical overlap, semantic similarity, and targeting-related measures are

among the strongest predictors of persuasive success. An ablation study further shows that targeting features alone provide limited predictive power but consistently improve performance when combined with alignment-based features. In particular, targeting features alone achieve near-random performance (ROC-AUC = 0.51), while their combination with alignment-based features improves predictive performance, supporting the hypothesis that persuasion depends on both topical grounding and direct engagement with specific arguments. This finding suggests that successful persuasion is not merely a consequence of discussing the same topic, but rather of selectively engaging with specific components of an argument while remaining within its conceptual frame. The observed patterns are consistent with earlier findings that persuasive success depends on both argumentative structure and interaction dynamics (Hidey & McKeown, 2018; Tan et al., 2016).

The results indicate that persuasion in online discussions emerges from the interaction between global semantic alignment and local argumentative engagement. Although the proposed framework does not match the predictive performance of recent neural approaches, it provides greater interpretability and enables direct analysis of the mechanisms underlying persuasive success. More broadly, the study contributes to Explainable Artificial Intelligence by demonstrating how transparent feature representations can be used not only to predict outcomes but also to uncover the linguistic mechanisms associated with successful persuasion. These findings support the continued relevance of interpretable machine learning for studying argumentative interaction and suggest promising directions for future hybrid approaches combining transparent feature representations with language models.

Future work will investigate whether the observed relationship between alignment and targeted engagement remains consistent across stronger machine learning and transformer-based models, providing a broader perspective on the trade-off between predictive performance and interpretability in persuasion modeling.

Keywords: Persuasion Modeling, Argumentation Mining, Explainable Artificial Intelligence, Interpretable Machine Learning, Online Discussions

References:

- Chang, J. P., Chiam, C., Fu, L., Wang, A. Z., Zhang, J., & Danescu-Niculescu-Mizil, C. (2020). ConvoKit: A toolkit for the analysis of conversations. *Proceedings of the 21st Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 57–60.
- Hidey, C., & McKeown, K. (2018). Persuasive influence detection: The role of argument sequencing. *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence* 32(1), 5173–5180.
- Li, H., Schlegel, V., Sun, Y., Batista-Navarro, R., & Nenadic, G. (2025). Large language models in argument mining: A survey. *arXiv preprint arXiv:2506.16383*.
- Salvi, F., Horta Ribeiro, M., Gallotti, R., & West, R. (2025). On the conversational persuasiveness of GPT-4. *Nature Human Behaviour*, 9, 1645–1653.
- Tan, C., Niculae, V., Danescu-Niculescu-Mizil, C., & Lee, L. (2016). Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. *Proceedings of the 25th International Conference on World Wide Web*, 613–624

Mathematical Foundations of AI and Machine Learning: A Pilot Study Among Graduates of Specialized Secondary School Students

Marina Svičević¹, Aleksandar Milenković¹, Nemanja Vučićević¹, Branko Arsić¹, Lazar Krstić¹

¹ University of Kragujevac, Faculty of Science, Kragujevac, Serbia marina.svicevic@pmf.kg.ac.rs, aleksandar.milenkovic@pmf.kg.ac.rs, nemanja.vucicevic@pmf.kg.ac.rs, branko.arsic@pmf.kg.ac.rs, lazar.krstic@pmf.kg.ac.rs

The rapid expansion of tools based on large language models and neural networks in everyday educational practice has introduced new challenges for curriculum designers. Secondary school students in specialized mathematics and computer science tracks represent a particularly relevant group for analysis, as they are expected to possess a high level of abstract reasoning and mathematical competence. However, in contemporary educational discourse, a frequent misconception arises: the ability to use AI interfaces (e.g., prompting ChatGPT) is often equated with genuine digital literacy in artificial intelligence. Genuine AI literacy extends beyond tool consumption and presupposes an understanding of the underlying mathematical principles.

This pilot study was conducted in June 2024 with the aim of assessing the relationship between students' exposure to AI tools and their conceptual and mathematical understanding of artificial intelligence and machine learning. The sample consisted of 22 final-year students from specialized Informatics (SI) and Mathematics (SM) tracks at the First Grammar School in Kragujevac, Serbia. The research instrument comprised two parts: a questionnaire on students' usage habits and attitudes toward curricular change, and a knowledge test consisting of both closed-ended and open-ended questions addressing mathematical foundations of AI systems (derivatives, regression, linear programming, and decision-making algorithms).

The results indicate very limited formal exposure to AI-related content within the current secondary school curriculum. Machine learning concepts were encountered in school by only three students, all from the SI track within the subject *Computer Applications*; no SM-track student reported any such exposure. Despite this, a moderate level of intrinsic motivation was observed, as seven students expressed interest in independently exploring AI and machine learning topics.

Regarding practical experience, AI tools are widely used among students. To the question "Have you ever used any artificial intelligence tool?", 81.82% (N = 18) responded affirmatively, while 18.18% (N = 4) reported no prior use. However, self-assessed knowledge does not fully align with this high level of usage. The majority of students (59.09%, N = 13) rated their knowledge as basic, meaning they recognize the concept but do not understand underlying mechanisms. A further 31.82% (N = 7) rated their knowledge as good, while the lowest and highest rating categories were minimal (4.55%, N = 1 each). This discrepancy between usage and conceptual understanding is also reflected in students' attitudes toward curriculum reform: 63.64% (N = 14) support introducing a dedicated AI course, while 22.73% (N = 5) favor interdisciplinary integration. Overall, 86.37% of respondents support systematic inclusion of AI education in schools.

The knowledge test reveals the most important didactic insight of this study. Students demonstrated partial understanding of basic terminology, with correct responses for data mining (68.60%) and big data (81.80%), indicating familiarity with contemporary data science vocabulary at a surface level.

However, significant gaps emerge in responses to more complex open-ended questions requiring understanding of underlying mathematical structures. Only two students provided meaningful explanations of the role of first-order derivatives in optimization processes such as gradient descent, which form the basis of neural network training. No student provided correct answers regarding linear programming applications, clustering methods, or regression-based modeling frameworks. Only three students demonstrated basic awareness of regression analysis, while five provided partially meaningful but superficial explanations of decision tree algorithms. Additionally, only three respondents correctly identified artificial neural networks and deep learning as mathematical frameworks approximating learning mechanisms inspired by the human brain.

The findings confirm a gap among students in specialized tracks: while they are frequent users of AI tools and familiar with related terminology, their understanding of the mathematical foundations of these systems remains underdeveloped. Importantly, this gap does not appear to stem from cognitive limitations or lack of student ability. Rather, it reflects insufficient institutional exposure to AI as a mathematical discipline. Since students show a high willingness to learn (86.37%), the results represent a clear empirical indication for curriculum development in Serbia, suggesting the need for systematic integration of AI and machine learning topics into mathematics and informatics education through explicit mathematical modeling.

Future research should extend this study to a larger sample and focus on designing instructional frameworks that support a deeper understanding of the mathematical principles underlying AI and machine learning systems.

Keywords: artificial intelligence, machine learning, mathematical foundations, secondary education, curriculum development

References:

- Deisenroth, M. P., Faisal, A. A., & Ong, C. S. (2020). *Mathematics for machine learning*. Cambridge University Press.
- Long, D., & Magerko, B. (2020, April). What is AI literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI conference on human factors in computing systems* (pp. 1-16).
- Milenković, A., Stanković, M., Svičević, M., & Vučičević, N. (2026). What Do Students in the Republic of Serbia Think About the Use of Generative AI Tools in Mathematics Teaching and Learning? *2nd International Scientific Conference Education and Artificial Intelligence (EDAI 2026)*, Vranje, May 15–16, 2026.
- Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence*, 2, 100041.
- Touretzky, D., Gardner-McCune, C., Martin, F., & Seehorn, D. (2019, July). Envisioning AI for K-12: What should every child know about AI?. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 33, No. 01, pp. 9795-9799).

GENERATIVE AI



Addressing the Illusion of Understanding Through Multi-Agent Dialogue: A Socially Scaffolded Approach to AI-Assisted Academic Reading

Xinju Liu¹, Min Lan²

¹ Department of Educational Technology, College of Teacher Education, Zhejiang Normal University, Jinhua, China, xinjuliu@zjnu.edu.cn

² Zhejiang Key Laboratory of Intelligent Education Technology and Application, Zhejiang Normal University, Jinhua, 321004, Zhejiang, China, lanmin@zjnu.edu.cn

Chinese graduate students often experience an “illusion of understanding” when reading English-language research articles, whereby perceived comprehension exceeds their ability to explain, critique, or apply research findings. In the era of AI-assisted reading, language barriers are becoming increasingly manageable through tools that provide instant translations, explanations, and summaries (Chea & Xiao, 2024). Yet many students continue to struggle with a more fundamental challenge: understanding the methodological reasoning that underpins scientific inquiry. Genuine comprehension of research requires readers not only to understand reported findings but also to evaluate research designs, assess the quality of evidence, and transfer insights to new contexts. Existing technology-assisted reading tools, however, primarily function as information providers rather than cognitive facilitators, offering ready-made answers without requiring substantial analytical engagement. As a result, these tools may inadvertently reinforce the illusion of understanding by fostering metacognitive laziness, characterized by reduced self-monitoring and critical evaluation during AI-supported learning (Fan et al., 2025).

To address the illusion of understanding in AI-assisted academic reading, this study adopts a socially scaffolded learning approach grounded in social learning theory (Bandura, 1977). The proposed platform engages learners in dialogue with multiple AI agents that provide scaffolded support through questioning, prompting, and feedback rather than direct information delivery. By requiring learners to explain their reasoning, evaluate methodological choices, and reconcile conflicting interpretations, these agents encourage active cognitive engagement and co-regulation of learning. Such scaffolded interactions are intended not only to support immediate comprehension but also to cultivate self-regulatory and metacognitive skills. Consistent with the principle of desirable difficulties (Bjork & Bjork, 2011), the platform introduces productive cognitive challenges that stimulate deeper processing and more robust learning. Over time, learners may internalize these regulatory processes and apply them independently when engaging with new academic texts (Zimmerman, 2002).

The platform incorporates four complementary AI agent roles that collectively scaffold methodological understanding, conceptual clarification, critical evaluation, and metacognitive reflection. The Methodology Expert helps learners understand not only what methodological procedures are used but also why specific research choices are theoretically and practically justified. The Illustrative Example Provider translates abstract methodological concepts into accessible examples and analogies that support deeper comprehension. The Critical Reviewer encourages learners to evaluate methodological strengths, limitations, and alternative interpretations, fostering critical engagement with research. The Summary Synthesizer promotes self-monitoring and knowledge consolidation by requiring learner-generated summaries before providing structured feedback. Together, these agents shift AI-assisted reading from passive information consumption toward active construction and evaluation of knowledge.

This platform is developed and deployed on the Coze low-code AI platform, leveraging the Next.js 16 full-stack framework with the App Router architecture. The frontend is built with React 19 and TypeScript 5 in strict mode. AI capabilities are integrated through the coze-coding-dev-sdk, utilizing the Doubao large

language model (doubao-seed) with Server-Sent Events (SSE) streaming output to enable sequential multi-agent discussions and real-time dialogue. The data layer employs Supabase (PostgreSQL) cloud database for managing literature, conversation logs, and user information, while PDF files are hosted via S3-compatible object storage with only file references stored in the database to avoid large-file storage bottlenecks. The frontend UI is based on the shadcn/ui component library and Tailwind CSS 4 styling system. PDF preview is rendered using react-pdf and PDF.js, and dialogue export supports PDF generation through jsPDF and html2canvas. The platform incorporates a research methods knowledge base, dynamically injecting role-specific prompts through keyword retrieval and task-type matching, thereby implementing a multi-agent architecture with shared knowledge base and differentiated perspectives.

This study investigates the effects of a multi-agent scaffolded academic reading platform on EFL graduate students' reading performance, cognitive processes, and emotional experiences. Using a quasi-experimental pretest–posttest design, participants will be assigned to either a multi-agent platform condition or a conventional LLM-based support condition. Academic reading performance will be assessed through reading-summary tasks, while cognitive and emotional processes will be examined through eye tracking, fNIRS measures of DLPFC activation, emotion questionnaires, and semi-structured interviews. This multimodal design aims to reveal both the effectiveness of the platform and the mechanisms underlying its impact on academic reading.

Theoretically, by integrating social learning theory with self-regulated learning in the context of AI-assisted academic reading, the study extends understanding of how technology can scaffold metacognitive development rather than merely provide information delivery. In terms of design, the four-role agent framework offers a structured approach to designing interactive AI systems that promote deep learning through socially scaffolded dialogue, distinguishing this platform from conventional AI reading assistants. Practically, findings can inform the development of academic reading interventions that effectively address the illusion of understanding, benefiting graduate students and supervisors in EFL and other academic contexts.

Keywords: multi-agent system, academic reading, social learning, self-regulated learning, metacognitive

References:

- Bandura, A., & Walters, R. H. (1977). *Social learning theory* (Vol. 1, pp. 33-52). Englewood Cliffs, NJ: Prentice-hall.
- Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. *Psychology and the real world: Essays illustrating fundamental contributions to society*, 2(59-68), 56-64.
- Chea, P., & Xiao, Y. (2024). Artificial intelligence in higher education: The power and damage of AI-assisted tools on academic English reading skills. *Journal of General Education and Humanities*, 3(3), 287-306.
- Fan, Y., Tang, L., Le, H., Shen, K., Tan, S., Zhao, Y. & 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.
- Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory into practice*, 41(2), 64-70.

Protecting Artistic Identity in the Age of Generative AI

Dimitrije Matović¹, Aleksandar Milenković², Mladen Nikolić³, Sana Stojanović Đurđević⁴

¹Faculty of Mathematics, University of Belgrade, mi20202@alas.matf.bg.ac.rs

²Faculty of Science, University of Kragujevac, aleksandar.milenkovic@pmf.kg.ac.rs

³Faculty of Mathematics, University of Belgrade, nikolic@matf.bg.ac.rs

⁴Faculty of Mathematics, University of Belgrade, sana@matf.bg.ac.rs

AI is currently dominating in most areas of expertise, and that includes the field of arts. Generative AI systems are now capable of producing sophisticated artworks from simple textual prompts, raising important ethical, legal, and philosophical questions regarding authorship, creativity, and artistic ownership.

A striking illustration of this tension emerged when Elon Musk posted an AI-generated image on X intended to demonstrate that AI can “learn and create art,” which in fact combined recognizable elements of two nineteenth-century paintings and introduced twenty-first-century cars in a scene styled after nineteenth-century originals. The example is a reminder that even celebrated generative outputs remain derivative of specific, unconsented source works, which is precisely the concern this paper examines.

This paper briefly explains how artworks are digitized and how such digitized content can be used to train AI models capable of generating new artistic content. These digitized artworks are often scraped from the internet without the explicit consent of the original creators. It enables the models to replicate and mimic the styles of the artists they study from. Particular attention is devoted to two recently developed tools designed to help artists protect their work from unauthorized use in AI training: Glaze and Nightshade.

Glaze operates as a protective cloak. It applies perturbations to an artwork, shifting its digital representation in the model's feature space towards a different art style. This deceives models into learning the wrong style, making style-mimicry attacks ineffective while leaving the drawing unchanged to the human eye.

One of the examples of where Glaze is used is Cara, a portfolio and social platform built specifically for artists concerned about unconsented AI scraping. They've integrated Glaze directly into its upload pipeline through a feature known as Cara Glaze, letting artists cloak their images automatically when posting or in batches for existing portfolios. This lowers the technical barrier for illustrators, concept artists, and photographers who want protection without running the tool themselves.

Nightshade functions as a poison attack tool. It alters images so that their internal feature vectors resemble a semantically unrelated concept, for example, making an image of a dog register as an image of a cat. When a number of poisoned samples enter the training dataset, the model will start producing incorrect outputs.

Nightshade's uptake after its 2023 release was similarly rapid: reporting at the time noted several hundred thousand downloads within its first days of availability, reflecting how strongly artists felt the need for a tool that actively deters, rather than merely evades, unauthorized training. Its developers have also recommended a layered workflow in which artists first apply Nightshade to an image and then apply Glaze on top, combining a deterrent against scraping with a defense against style mimicry, and platforms such as Cara have signaled plans to support this combined pipeline directly.

The debate extends beyond individual artists. Backlash from gamers and artists against AI-driven features such as NVIDIA's DLSS, along with reported internal reconsideration of AI policy at companies including Microsoft and NVIDIA, suggests that concerns over generative AI's effect on creative authenticity are shared industry-wide rather than confined to isolated cases.

The paper uses these tools to discuss how artists can protect their work, artistic identity, and intellectual property in the context of rapidly advancing generative AI technologies.

Keywords: generative AI, style mimicry, Glaze, Nightshade, data poisoning, adversarial perturbations, AI ethics

References:

- Beregovyi, K., & Butkiewicz, T. (2025, November). Visual Integrity in the Age of AI: An Evaluation of DLSS and DLAA in Geospatial Visualization. In 2025 IEEE Visualization and Visual Analytics (VIS) (pp. 291-295). IEEE.
- Shan, S., Cryan, J., Wenger, E., Zheng, H., Hanocka, R., & Zhao, B. Y. (2023). Glaze: Protecting artists from style mimicry by {Text-to-Image} models. In 32nd USENIX Security Symposium (USENIX Security 23) (pp. 2187-2204).
- Shan, S., Ding, W., Passananti, J., Wu, S., Zheng, H., & Zhao, B. Y. (2024, May). Nightshade: Prompt-specific poisoning attacks on text-to-image generative models. In 2024 IEEE symposium on security and privacy (SP) (pp. 807-825). IEEE.

From No-Code to Orchestrated Notification Agent

Ognjen Ristić¹

¹ Mihajlo Pupin Institute, Belgrade, Serbia, ognjen.m.ristic@pupin.rs

Previous work on no-code AI automation demonstrated that visual workflow platforms such as Make, Zapier, and Power Automate can lower the technical barrier for building AI-driven notification chains. As engineering workflows grow in complexity, accumulating stateful dependencies, conditional branching and retries the limitations of no-code platforms become more visible. This paper presents a code orchestrated agentic notification system built on Apache Airflow, in which all-MiniLM-L6-v2, 22.7M parameters reasoning is embedded inside scheduled Directed Acyclic Graphs (DAGs), a Claude LLM call is embedded as a dedicated DAG task between paper retrieval and document generation, producing a written research summary that appears as a standalone section in the compiled report, and delivered to the user through Slack or any other app that provides an API key, at fixed intervals. The agentic character of the system lies in its context aware query construction and embedding based relevance judgment. The DAG reads the live state of the simulation pipeline and constructs a query that reflects where the project currently stands. The morning and evening run additionally alternate between complementary research themes, ensuring that the retrieved literature covers the full scope of the research problem over time.

Apache Airflow is an open-source workflow orchestrator that represents pipelines as DAGs of tasks with explicit dependencies. Each task is a Python callable, a sensor that waits for an event, or an operator that wraps a third-party service. Airflow provides scheduling, automatic retries, backfilling of missed runs, observability through a web UI, and data exchange via the XCom mechanism. These features address the gaps in no-code platforms when applied to workflows that must be reproducible, versioned, and testable. The pipeline consists of five sequential tasks. The presented system consists of a single DAG scheduled to execute twice per day at 12 hours gap.

The first task reads the current state of the simulation pipeline from CSV and JSON result files, constructs a search query that reflects the current research focus, and writes the simulation metrics and LaTeX summary to a temporary directory. The second task launches the search agent as a subprocess inside the dedicated Python virtual environment, where arxiv, PyMuPDF, and sentence-transformers are installed. The search agent queries both arXiv and Semantic Scholar, downloads PDFs, extracts text blocks using fitz, and ranks paragraphs against the query using cosine similarity in the 384-dimensional embedding space. Results are serialized to a papers.json file and the path is passed to the next task via XCom. The third task reads papers.json, builds a prompt from paper titles and abstracts (400 characters per paper), attaches the current simulation state, and sends the full request to Claude using the Anthropic API. The model is claude-opus-4-7 with adaptive thinking enabled.

The user's research profile, a plain text file is passed as a cached system prompt using Anthropic's prompt caching mechanism (cache_control: ephemeral). Claude is instructed to identify common themes across the papers, note which are most relevant to the current work, and connect findings to the simulation state. The output is a plain text file saved to the temporary directory. The Anthropic API key is loaded at task runtime from a .env file inside the directory using python-dotenv and never appears in the DAG code. The fourth task passes the papers JSON, simulation context files, and the Claude synthesis file to the LaTeX agent subprocess. Inside that subprocess, the synthesis text is wrapped in a dedicated LaTeX section. The LaTeX agent also computes embedding-based relevance scores, sorts papers by score against the merged research profile, and compiles the document to PDF using pdflatex. The fifth

task reads the PDF path from XCom and posts the file to the designated Slack channel using the Slack Web API. The DAG with the tasks is shown in Fig.1.

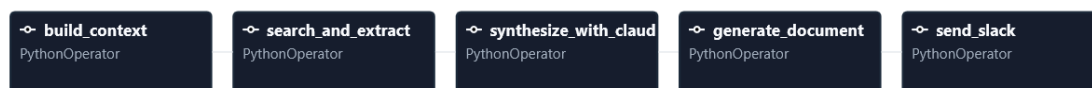


Fig. 1: Apache Airflow DAG of the twice-daily notification agent, showing the five sequential tasks and their dependencies.

The relevance judgment step is performed by a sentence embedding model (all-MiniLM-L6-v2, 22.7M parameters), which encodes both the user's research profile and each paper's content into a shared 384-dimensional semantic vector space. Cosine similarity between these representations produces a continuous relevance score that drives paper ranking, section ordering, and citation recommendations in the generated report, all without human intervention. This is not a keyword match but a semantic assessment, a paper discussing a specific topic will score highly against a profile that mentions similar topic even if no identical terms appear. The same embedding model also performs paragraph ranking within each PDF providing the reader with an immediate picture of which aspects of the query are represented in the literature and which are sparse.

Changing the research profile or the query rotation requires editing a single text file or a two-element list. Adding a new data source requires implementing one additional search function in search agent. The LaTeX output provides a version-controllable record of every run, independent of the Slack delivery channel. The Slack is connected with an API key.

Compared to a no-code or manual literature monitoring approach, the Airflow-based pipeline provides reproducibility, deterministic twice-daily scheduling, and the ability to backfill missed runs. A limitation is that the relevance judgment is purely statistical. The embedding model ranks content by similarity to a fixed profile but does not reason about the relationship between papers. It will surface the most topically proximate content, not necessarily the most scientifically significant.

Future work will address these limitations. Passing the top-ranked full paragraphs to Claude rather than just abstracts would allow a more grounded search and verification.

Keywords: Apache Airflow, AI Agents, Workflow Automation, Data Processing, Ai automation

References:

- Ristić, O., 2025. No-Code AI Automation – Data Analysis with Visual Workflow Automation. AI Conference, Mathematical Institute SASA, Belgrade.
- Apache Software Foundation, 2024. Apache Airflow Documentation. Available at: <https://airflow.apache.org/docs/>
- Anthropic, 2025. Claude API Documentation. Available at: <https://docs.anthropic.com>
- Brown, T.B., Mann, B., Ryder, N., et al., 2020. Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems, 33, pp.1877–1901.
- Wang, L., Ma, C., Feng, X., et al., 2024. A Survey on Large Language Model Based Autonomous Agents. Frontiers of Computer Science, 18(6), 186345.
- Wang W. et al., "MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers", NeurIPS 2020.
- Reimers N. & Gurevych I., "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks", EMNLP 2019.

Semantic Harness for LLM-driven Agents

Milorad Tosić¹, Nenad Petrović²

¹ Faculty of Electronic Engineering, University of Nis, Serbia, milorad.tosic@elfak.ni.ac.rs

² Faculty of Electronic Engineering, University of Nis, Serbia, nenad.petrovic@elfak.ni.ac.rs

Generative AI (GenAI), particularly Large Language Models (LLMs), offer significant potential for intelligent system automation. However, their use in critical domains is limited by the risk to produce unreliable outputs and usually leads to operational or financial consequences. Therefore, additional mechanisms are needed to ensure trustworthy and explainable decision making (Bornet et al., 2025). To address the limitations of standalone Large Language Models (LLMs) in critical domains, we propose the Semantic Harness for LLM-driven Agents (SH-A), a Knowledge Graph (KG) based framework that combines structured context management, semantic reasoning, Retrieval-Augmented Generation (RAG) and controlled workflow execution to enable reliable AI-assisted decision support. Apart from semantic grounding of user requests, key advantages against other solutions include explainability, traceability, as well as straightforward integration of new tools, workflows and external knowledge sources. Figure 1a illustrates the proposed workflow.

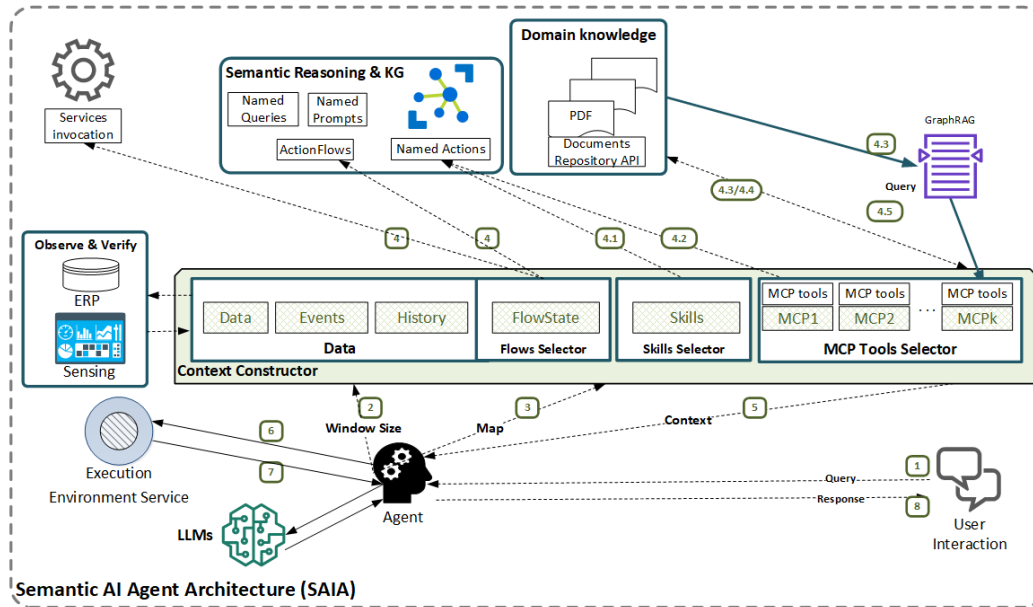


Figure 1. Semantic-driven agent harness architecture workflow

User interaction begins by submitting a request through the application interface (1). The request is forwarded to the SAIA, where the Context Constructor assembles operational context from data sources, events, knowledge repositories, and interaction history within a configurable context window (2). The agent analyzes the request and maps it to predefined Named Actions and Named Prompts using RAG-based retrieval (3). The main novelty is the Context Constructor, which exploits semantic descriptions of context components stored in the KG and used for the mapping. Named Actions bridge user intent and system execution by linking requests to Action Flows, Skills, and Model Context Protocol (MCP) tools. Action Flows define execution paths, Skills provide domain expertise, and MCP tools enable access to external services and enterprise functionality. Named Actions represent semantic retrieval operations used to obtain contextual information from knowledge graph, document repository, and external MCP-accessible resources. Named Prompts provide predefined instructions, reasoning templates, and domain-

specific guidance supplied to the underlying LLM, promoting consistent and explainable decision-making. The context is subsequently enriched through retrieval of semantic knowledge assets (4), followed by MCP tool selection (4.1) and GraphRAG-based retrieval over semantic repositories and document collections (4.2–4.5). The retrieved information is combined with contextual data and user inputs by the Context Constructor into a unified operational context (5), which is then supplied to the corresponding Semantic AI Agent for reasoning and decision-making. The reasoning process follows the ReAct (Reason-Act) paradigm. Using the selected Named Actions and enriched context, semantic AI agents interact with the Execution Environment Agent to execute workflows, invoke MCP tools, and perform domain-specific operations (6). The execution environment orchestrates services and external systems, while execution results and observations are returned to the agents (7) for iterative reasoning and refinement. The Observe & Verify component continuously collects execution feedback and environmental observations (such as sensors and ERPs), validates outcomes, and feeds verified information back into the reasoning process. Finally, the generated response is returned to the user (8). Together, these components form a closed-loop, knowledge-grounded architecture that integrates semantic reasoning, retrieval, execution, and verification. A Context Ontology formalizes relationships between user intents, agent capabilities, contextual information, and tool outputs, enabling interoperability and decision traceability.

The proposed framework is evaluated through a smart-home flexible energy optimization use-case scenario considering 5 relevant questions responded by Qwen 2.5, as shown in Table 1. A user requests an optimized household energy schedule based on electricity tariffs and solar generation. The Context Constructor combines real-time forecasts, consumption data, and user preferences, maps them to semantic planning ontologies using LLM and RAG (Petrovic, 2025), and generates an optimized plan. Recommendations, such as shifting appliance usage to low-cost or high-solar periods, are accompanied by transparent explanations to improve trust and traceability. For answer scoring, we make use of LLM as judge – Claude Opus 4.8. The underlying answer judging prompt is defined as follows: “Compare Answer 1 and Answer 2 for the given question. Evaluate them based on: directedness, exactness, hallucination risk, focus, applicability, simplicity for end user”.

Table 2 Result comparison (S1: score Qwen 2.5 with SAIA; S2: score Qwen 2.5 no context)

Question	S1	Time 1	S2	Time 2	Winner
“Optimize flexible electricity consumption at home” (test1)	4	3.6 s	4	11.0 s	Tie
“Optimally use solar panel + hot water boiler” (test2)	3	4.9 s	4	10.9 s	S2
“Optimal time of day to charge an EV” (test3)	4	1.6 s	3	6.3 s	S1
“Reduce energy consumption via energy flexibility” (test4)	4	2.1 s	3	9.4 s	S1
“Could energy flexibility reduce my energy bill” (test5)	4	1.8 s	3	7.3 s	S1
Average	3.8	2.8 s	3.4	9.0 s	S1

Keywords: agentic AI, ontology, harness, LLM, ontology, smart grid

Acknowledgement: This work has been supported by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia [Grant Number: 451-03-34/2026-03/200102]. This work was framed in the context of the project O-CEI, which receives funding from the European Union’s Horizon Europe research and innovation programme under grant agreement 101189589.

References:

- Bornet, P., Wirtz, J., & Davenport, T. H. (2025). Agentic artificial intelligence: Harnessing AI agents to reinvent business, work and life. World Scientific Publishing.
- Petrovic, N., & Tasic, M. (2025). Advanced planning and scheduling with semantic knowledge graphs and LLMs. In Proceedings of the Joint Ontology Workshops (JOWO 2025), 2nd Annual Workshop on the Convergence of Large Language Models and Ontologies (ONTOLLM). CEUR Workshop Proceedings, Vol. 4176. <https://ceur-ws.org/Vol-4176/ontollm-3.pdf>

Market Mechanisms as Transparent Environments for AI Agents

Dragiša D. Žunić^{1,2}

¹ Institute for Artificial Intelligence R&D of Serbia, dragisa.zunic@ivi.ac.rs

² Gran Sasso Science Institute (GSSI), L'Aquila, Italy

This abstract proposes a formal symbolic view of core market mechanisms as an important foundation for studying AI agents as market participants. Financial markets involve heterogeneous actors – regular traders, liquidity providers, high-frequency firms, and institutional investors – but all of them interact through the same rule-governed infrastructure: the trading venue itself. Like a game of chess, a market mechanism defines admissible actions, interaction rules, priority, outcomes, and terminal states. For this reason, the behaviour of market participants cannot be understood only by modelling their strategies or objectives; it also depends on the formal structure of the mechanism through which they interact.

We argue for a transparent-box view of market mechanisms as rule-based settings in which AI agents act. By “environment” we do not mean a black box observed only through data, but a symbolic interaction structure whose internal rules can be made explicit and reasoned about. The point is to shift attention from trading agents alone to the market core itself: the rules that determine which orders may interact, which interactions have priority, and when the market has reached a stable state. This view is consistent with the classic zero-intelligence trader result, which showed that important market-level outcomes can arise from the structure of the market itself, even when individual agents are extremely simple (Gode & Sunder, 1993).

To make this view formal, we use Reaction Systems as a formal language (Ehrenfeucht & Rozenberg, 2007), building on ongoing work that adapts this framework to trading venues. In this setting, buy and sell orders are treated as resources, compatible orders react to produce trades, more competitive orders inhibit less competitive matches, and unmatched orders remain available for later steps. Thus, priority, competition, timing, persistence, and stability become part of the formal description of the mechanism rather than hidden details of implementation. This also makes key properties – such as priority preservation, persistence of unmatched orders, and absence of further executable trades expressible at the level of the mechanism itself. The approach continues a line of work in which trading systems are treated as formal computational objects rather than only as economic or statistical processes (Cervesato et al., 2019).

The approach is illustrated through two canonical market-core mechanisms. The Continuous Limit Order Book represents the familiar sequential view of a market: orders arrive continuously and are processed according to price-time priority. This design is simple and widely used, but it also makes timing and latency structurally important. Frequent Batch Auctions (FBA) take a different view: orders arriving within the same time interval are grouped together and treated as time-equivalent, a design proposed as a response to latency races in modern electronic markets (Budish et al., 2015).

As a small illustration, consider the following unit orders, listed by arrival: b1, buy at 100; s1, sell at 99; b2, buy at 101; s2, sell at 101; b3, buy at 101; b4, buy at 98; and s3, sell at 102. Under sequential matching in the CLOB model, s1 trades immediately with the resting order b1; later, s2 trades with b2. The resulting trades are b1–s1 and b2–s2. Under batched matching in the FBA model, the same orders are collected within one batch and their exact arrival order inside the interval is disregarded. At a clearing price of 101, the two most competitive buy orders, b2 and b3, and the eligible sell orders, s1 and s2, are executed.

The contrast is structural. CLOB matching gives decisive importance to serial arrival, so an earlier but less competitive order may trade before a later order offering a better price. The simplified FBA model instead treats orders within the batch as time-equivalent, bringing the matching logic closer to the economic interaction being modelled: competition is primarily expressed through prices rather than infinitesimal differences in arrival time. Computationally, this points toward models based on joint activation and potentially parallel or concurrent treatment of orders, rather than imposing a rigid serial order on interactions that are naturally simultaneous.

We also discuss a finer-grained view of FBA, in which the internal steps of batching, activation, matching, and stabilization are made explicit. This makes it possible to reason about sequentiality, batching, and the computational structure of the mechanism itself. The goal is not to present a new exchange design, but to show how mechanisms used in electronic markets today can be opened up and represented in a way that supports symbolic reasoning.

This perspective connects older work in agent-based computational economics, which treated economies as systems of autonomous interacting agents, with newer research on AI agents in strategic and economic settings. Both show that agents do not act in isolation, but through the core that structures their interactions. Depending on what the rules reward, an agent may learn to bid truthfully, overbid, underbid, delay, race for speed, or avoid competing too aggressively. The market mechanism therefore shapes which behaviours become profitable or stable. This is consistent with the view that intelligent behaviour cannot be studied apart from incentives (Parkes & Wellman, 2015).

For an adaptive or AI agent, a transparent market model makes the trading venue part of the reasoning environment. The agent can examine possible market transitions, learn which behaviours the mechanism rewards, and verify whether a contemplated strategy respects its matching, priority, and timing rules. Because different mechanisms can be formalized within the same framework, they can be compared on a common basis: we can identify which rules each transparent box contains, which behaviours it enables, and which core properties follow from its design. This supports joint analysis of the agent and the mechanism while distinguishing behaviour produced by the agent from behaviour induced by the market core.

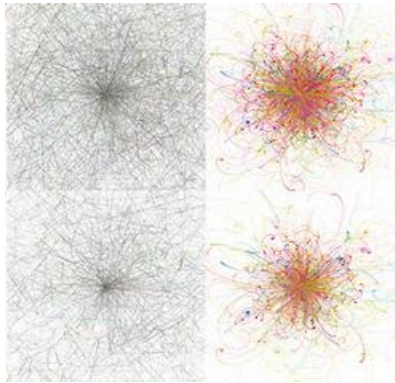
Our position is therefore that, before studying how AI agents behave in markets, we should make the mechanisms through which they interact explicit, transparent, and available for formal reasoning. Such models can support explainability, verification, and mechanism-aware analysis of multi-agent financial systems.

Keywords: market design, market microstructure, formal methods, transparent-box modelling, AI agents

References:

- Budish, E., Cramton, P., & Shim, J. (2015). The high-frequency trading arms race: Frequent batch auctions as a market design response. *The Quarterly Journal of Economics*, 130(4), 1547–1621.
- Cervesato, I., Khan, S., Reis, G., & Žunić, D. (2019). Formalization of automated trading systems in a concurrent linear framework. *Electronic Proceedings in Theoretical Computer Science*, 292, 1–14.
- Ehrenfeucht, A., & Rozenberg, G. (2007). Reaction systems. *Fundamenta Informaticae*, 75(1–4), 263–280.
- Gode, D. K., & Sunder, S. (1993). Allocative efficiency of markets with zero-intelligence traders: Market as a partial substitute for individual rationality. *Journal of Political Economy*, 101(1), 119–137.
- Parkes, D. C., & Wellman, M. P. (2015). Economic reasoning and artificial intelligence. *Science*, 349(6245), 267–272.

KNOWLEDGE REPRESENTATION, REASONING AND PLANNING



Identifying Prerequisites among Competencies: a Large Language Models Approach

Milan Jovkić¹, Goran Savić¹, Milan Segedinac¹, Peter Steiner²

¹ Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 6, 21000 Serbia, [milan.jovkic, savicg, milansegedinac]@uns.ac.rs

² Pädagogischen Hochschule St.Gallen (PHSG), Notkerstrasse 27, 9000 St.Gallen, Switzerland, Peter.Steiner@phsg.ch

State-of-the-art frameworks for student assessment often exploit prerequisites among skills for enhancing assessment efficiency and interpretability. The most well-known such models are Knowledge Space Theory (KST) (Doignon & Falmagne, 1999), Hierarchical Cognitive Diagnostic Models (HCDM), and Dynamic Bayesian Networks (DBN) (Corbett & Anderson, 1994). Defining prerequisites among educational competencies traditionally can be achieved in two ways: they are either identified by analyzing the results from previous students' assessments (Chen et al., 2016), or by representing the experts' knowledge about the domain. Both approaches have their benefits and drawbacks. So, while data-driven techniques do not require significant human effort, they cannot be applied in advance. Methods that involve experts, on the other hand, can propose the prerequisites prior to performing the assessment, but they are labor-intensive and often imprecise (Le & Abel, 2025). Both kinds of techniques lack scalability and cannot be applied to larger domains of knowledge.

This paper proposes a novel method for identifying prerequisite relations that relies on the fully automated semantic analysis of the educational materials from the knowledge domain. The method integrates Large Language Models (LLMs), vector semantic search, and a dual-graph architecture in which competencies are linked to the chunks of educational content, enabling graph-native Retrieval-Augmented Generation (RAG) (Edge et al., 2024). The core idea of the method is to split the process into two phases that form a pipeline. The first phase consists of an embedding-based filtering, which identifies semantically relevant pairs of competencies. Each competency is represented as a dense vector encoding its key descriptive attributes, namely its name, description, domain, knowledge type, and representative examples, which substantially enriches the semantic representation compared to name-only encoding. For each competency, the top- k most similar competencies already present in the graph are retrieved, where k was set to 7 in our implementation. Only these k candidates are passed to the second phase, eliminating $O(n^2)$ pairwise evaluations in favour of $O(n \cdot k)$.

The second phase employs an LLM-driven reasoning that determines direct prerequisite relations. For each of the k candidates, the prompt is constructed from three sources of information: the full descriptive data of both the target and candidate competencies, along with their semantic similarity score, structural metadata from the curriculum specification covering subdomain, cognitive process level, and input and output representations, and contextually relevant instructional text retrieved from the RAG graph. The prompt also encodes explicit pedagogical heuristics, including a cognitive process hierarchy and a representation chain principle, that guide the model in determining the direction of prerequisite relations. The reasoning is bidirectional, simultaneously identifying prerequisites in both directions between the target competency and each candidate.

The proposed method solves several problems. Firstly, because the prerequisites are identified in the learning materials, it can be applied prior to the assessment of students' knowledge. That way, the knowledge structures are available before testing students and they can be applied in the adaptive procedures already in the first iterations of testing. Secondly, the method does not involve human experts,

which makes it cheap to implement and less prone to human errors. And finally, the method ensures scalability as the repository of competencies grows, while maintaining the logical coherence of the prerequisite relations through the underlying graph model. The system built around the proposed method additionally provides an interactive knowledge graph visualisation and a RAG-powered chat interface for exploring and querying the structured competency base.

As the initial validation, the method was applied to identify prerequisites among 123 mathematical competencies from the domain of linear functions using two LLM models: deepseek-r1-0528, a reasoning model accessed via OpenRouter, and gpt-4o-mini, accessed via GitHub Models. The resulting graphs were compared on an identical 21-competency subgraph, and both were acyclic. The two models agreed on a slight majority of the direct prerequisite relations, reaching a Jaccard similarity of 0.57, forming a pedagogically coherent shared core directed from higher-level skills toward the foundational concepts they build on. The main difference between them was in relation density: deepseek-r1 produced more direct prerequisite relations than gpt-4o-mini, 44 against 36, corresponding to an average of 2.1 versus 1.7 prerequisites per competency. The comparison demonstrates that model choice significantly affects relation density, while a stable core is recovered regardless of the model.

Two examples illustrate these findings. For the competency "determine the linear function equation from a point and its slope", both models independently identified the same two well-justified prerequisites, "recognize and describe the slope" and "determine the function term from the y-intercept and the slope", exemplifying the stable core recovered regardless of the model. For the competency "determine parts of a value table from a context description", both models shared a common core, consisting of "identify the given quantities from the context", "know that the dependent variable changes with the independent one", and "determine a value table from the function equation", while deepseek-r1 attached one further prerequisite, "determine a value table from a graph", which is not well justified. This illustrates how deepseek-r1 produces additional direct relations that account for its higher density.

The initial research presented in this paper also shows some key limitations of the framework. The most notable ones are the result's non-determinism inherent to the probabilistic nature of LLMs and the dependence on model quality. Future work will conduct a formal quantitative evaluation, including precision and recall against expert-annotated prerequisite structures such as the Achieve the Core Coherence Map (achievethecore.org/coherence-map/), automate chunk selection per competency, and extend this approach to broader mathematical and multi-domain curricula.

Keywords: Knowledge Graph, Prerequisite Relations, Large Language Models, Retrieval-Augmented Generation, Vector Semantic Search

References:

- Chen, Y., González-Brenes, J. P., & Tian, J. (2016). Joint discovery of skill prerequisite graphs and student models. In *Proceedings of the 9th International Conference on Educational Data Mining (EDM)* (pp. 46–53).
- Corbett, A. T., & Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4(4), 253–278.
<https://doi.org/10.1007/BF01099821>
- Doignon, J.-P., & Falmagne, J.-C. (1999). *Knowledge spaces*. Springer.
- Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., & Larson, J. (2024). From local to global: A graph RAG approach to query-focused summarization. *arXiv*.
<https://doi.org/10.48550/arXiv.2404.16130>
- Le, N. L., & Abel, M.-H. (2025). *How well do LLMs predict prerequisite skills? Zero-shot comparison to expert-defined concepts*. *arXiv*. <https://doi.org/10.48550/arXiv.2507.18479>

Beyond Exact Skill Matching: Graph-Constrained AI Modeling of Competence Gaps

Robi Pritrznik¹, Pavle Boškoski^{1,2}, Tjaša Redek⁴, Sagnika Sen³, Biljana Mileva Boshkoska^{1,2}

¹ University of Novo mesto, Faculty of Information Studies, Slovenia, robi.pitrznik@fis.unm.si,
pavle.boskoski@fis.unm.si, biljana.mileva@fis.unm.si

² Jožef Stefan Institute, Slovenia

³ School of Graduate Professional Studies, Pennsylvania State University, Malvern, Pennsylvania, USA, sus45@psu.edu

⁴ School of Economics and Business, University of Ljubljana, Ljubljana, Slovenia,
tjasa.redek@ef.uni-lj.si

Competence-gap analysis often relies on exact matching between skills available in the workforce and those required by employers. Such an approach may overlook meaningful relationships between related competences and consequently overestimate existing gaps. This work proposes an AI-based approach for modeling competence gaps by combining the structural information encoded in the ESCO skills ontology with unbalanced optimal transport.

Skills are represented as nodes in the ESCO graph, where broader/narrower relationships provide information about their proximity. Instead of treating skills as independent labels, the proposed model uses this structure to identify plausible substitutions between related competences. The cost of substitution is determined by shortest-path distance in the graph: exact matches have no substitution cost, while increasingly distant concepts are more costly. Substitution is restricted to concepts within three graph hops, allowing sufficiently related skills to contribute to a match while leaving more distant demand unmatched.

The relationship between competence supply and demand is modeled using unbalanced optimal transport (Chizat et al., 2018). Unlike balanced transport, this formulation does not require all available and required competences to be matched and therefore avoids forcing assignments between distant concepts. Entropic Sinkhorn optimization (Cuturi, 2013; Peyré & Cuturi, 2020) is used to efficiently estimate the resulting transport plan. The model thus combines quantitative information about competence supply and demand with structural information from the ESCO ontology.

The approach is applied to ESCO v1.2.1 and public-employment data describing competences associated with registered job seekers and job postings. The empirical analysis shows that an observed competence gap does not necessarily imply that related competences are absent from the workforce. Among the 40 largest unmatched demand concepts, available supply can be found within three graph hops for all 40, while for 23 concepts the identical ESCO competence is already reported by some job seekers. The remaining gap therefore reflects insufficient quantities of relevant or nearby competences rather than simply the absence of corresponding concepts.

The main contribution is an interpretable AI modeling framework in which the existing ESCO ontology defines plausible skill substitutions, graph distance represents substitution cost, and unbalanced optimal transport determines how available competence supply can be assigned to demand. The resulting model distinguishes between exact matches, nearby competence substitutions, and demand that remains unmatched. Rather than introducing a new competence taxonomy, the approach builds directly on ESCO and demonstrates how its existing structure can support more informative and explainable competence-gap analysis in public employment services.

Keywords: ESCO ontology, substitution graph, unbalanced optimal transport, knowledge graph, hop-cost reasoning

References:

- Chizat, L., Peyré, G., Schmitzer, B., & Vialard, F.-X. (2018). Unbalanced optimal transport: Dynamic and Kantorovich formulations. *Journal of Functional Analysis*, 274(11), 3090–3123.
<https://doi.org/10.1016/j.jfa.2018.03.008>
- Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems*, 26.
- European Commission, Directorate-General for Employment, Social Affairs and Inclusion. (2017). *ESCO handbook: European skills, competences, qualifications and occupations*. Publications Office.
<https://doi.org/10.2767/934956>
- McGuinness, S., Pouliakas, K., & Redmond, P. (2018). Skills mismatch: Concepts, measurement and policy approaches. *Journal of Economic Surveys*, 32(4), 985–1015.
<https://doi.org/10.1111/joes.12254>
- Peyré, G., & Cuturi, M. (2020). Computational optimal transport.
<https://doi.org/10.48550/arXiv.1803.00567>

Machine Learning in Optimized SPARQL Query Execution Over RDF Knowledge Graphs

Muhammad Sohail Nisar¹, Muhammad Saleem¹, Axel-Cyrille Ngonga Ngomo¹, Sebastian Tramp², Nieves Sande², Milos Jovanovik^{3,4}, Mirko Spasić^{3,5}

¹Dept. of CS, Paderborn Univ., Germany, msohail | saleem | axel.ngonga@mail.upb.de

²eccenca GmbH, Leipzig, Germany, sebastian.tramp | nieves.sande@eccenca.com

³OpenLink Software, London, United Kingdom, mjovanovik | mspasic@openlinksw.com

⁴Faculty of Informatics, TU Wien, Austria, milos.jovanovik@tuwien.ac.at

⁵Faculty of Mathematics, University of Belgrade, Serbia, mirko.spasic@matf.bg.ac.rs

The rapid growth of RDF knowledge graphs (KGs) and semantic web technologies has created a strong need for efficient SPARQL query processing over centralized and distributed sources. Modern KGs, such as Wikidata, DBpedia, LinkedGeoData, and biomedical KGs contain very large numbers of RDF triples. This makes scalable query optimization a major challenge in semantic data management. RDF provides a graph-based representation model for structured knowledge, while SPARQL is the standard query language for retrieving and combining information from RDF data sources (Harris & Seaborne, 2013). However, the performance of SPARQL query execution strongly depends on the quality of the selected query plan, especially the ordering of joins and the selection of relevant sources in federated settings. Traditional query optimizers usually rely on cost-based models, cardinality estimation, statistics, indexing techniques, and heuristic join-ordering strategies. Although these methods are often effective, they fail when statistics are incomplete or outdated. This challenge becomes even more significant in federated processing, where a query may need to access multiple endpoints and combine partial results across the network. Poor source selection and inefficient join ordering can lead to unnecessary intermediate results, excessive network traffic, and long query execution. These issues motivate the use of adaptive ML techniques for SPARQL query planning. The SPARQL-ML project investigates machine learning-based methods for improving SPARQL query optimization over centralized triplestores and federated SPARQL query engines. In particular, we are using deep reinforcement learning for SPARQL query planning, both in centralized and distributed environments. The project treats query planning as a sequential decision-making problem. In this setting, a query representation describes the current optimization state, planning choices such as join ordering correspond to actions, and measured query execution time provides feedback to the learning system. In this framework, the agent continuously learns to improve its decisions through repeated interaction with an environment (Sutton & Barto, 2018). In contrast to static optimizers, the goal is to have an adaptive optimizer that can learn from observed query workloads and executions. A first technical component of the project is the representation of SPARQL queries in a form suitable for ML. Instead of treating a query only as text, the system extracts structural information from the query graph, including triple patterns, variables, join relationships, projected variables, and query-level statistics. These features are transformed into numerical representations (in the form of embeddings) that can be consumed by a learning agent. The representation is designed to capture both local information, such as the role of a variable in a triple pattern, and global information, such as the connectivity structure of the query. This is important because the quality of a join-order decision depends not only on individual triple patterns, but also on how they interact within whole query graph.

The current implementation integrates a reinforcement learning environment with RDF engines such as Tentriss and Virtuoso. The learning agent observes the query representation and proposes planning decisions, for example a join-variable ordering or a join-priority hint. The triplestore remains responsible for physical execution, while the agent provides additional learned guidance to influence the optimization process. Execution time is then used as the reward signal. Early experiments on DBpedia test queries show encouraging behavior: the learned planner can improve many individual warm-cache query

executions and produce stable plans, but the overall result is still preliminary as long-tail cases and broader cross-dataset generalization remain under evaluation. Therefore, the current results are treated as an initial indication rather than a final benchmark claim. In federation setting, optimization must decide not only how to order joins, but also which data sources are relevant for each part of the query. SPARQL-ML builds on join-aware source selection techniques, where information about predicates, URI prefixes, data summaries, and join relationships is used to avoid contacting irrelevant endpoints. This reduces unnecessary network traffic and helps generate more efficient federated execution plans. Previous work on federated SPARQL optimization, such as CostFed, has shown the importance of source selection and cost-aware planning for distributed RDF query processing (Saleem et al., 2018). SPARQL-ML extends this direction by exploring how such structured source-selection information can support learning-based planning strategies. The project architecture follows a reusable and engine-independent design. i.e., the SPARQL-ML components can be integrated with any of the existing triple stores or SPARQL federation engines. Query features are extracted from the RDF engine and passed through an API layer to the learning component. The agent returns planning guidance to the engine, which executes the query and reports runtime information back to the system. This feedback loop enables continuous experimentation with different feature representations, learning algorithms, reward functions, and triplestore backends. The architecture is intended to support both centralized and federated scenarios, while remaining compatible with standard SPARQL processing workflows. The ML overhead consists of calculation of an efficient join order and providing it to the underlying engine. On the other hand, the engine itself can skip its own query optimizer (which usually incorporate data sampling, cost modelling, cardinality estimations and different heuristics) and directly follow the query plan generated by the agent. Therefore, we can expect that the ML overhead won't be longer than the regular optimizer time. In general, the query planning time is much smaller than the actual query execution time. The project will be evaluated using real-world and industry-relevant use cases. These include enterprise KGs using anonymised industry data from eccenca, biomedical-scale data represented by the Linked Cancer Genome Atlas use case at Paderborn University, and large public KGs DBpedia and LinkedGeoData through OpenLink Software, together with real-world queries from the relevant query logs. These use cases are important because they represent practical settings where query runtime, scalability, and robustness directly affect the usability of RDF-based applications. The main evaluation criterion is query execution time, complemented by success rate, robustness across workloads, and comparison against existing baseline optimizers already implemented in the triplestore under test. In addition to evaluation on real queries over the aforementioned datasets, the comparison will also be done on the synthetic and well-known benchmarks such as Geospatial and Social Network benchmarks. By combining RDF knowledge representation, SPARQL query optimization, and reinforcement learning, SPARQL-ML aims to contribute to a new generation of adaptive semantic data management systems. The project does not aim to replace existing triplestore optimizers, but to enhance them with learned planning support. This makes the approach suitable for practical deployment, where mature RDF engines can retain their execution capabilities while benefiting from feedback-driven optimization. In the long term, this line of work may help improve the scalability and efficiency of KG applications in industry, science, healthcare, and smart city infrastructures.

Keywords: SPARQL, Deep Reinforcement Learning, query optimization, RDF knowledge graphs

References:

- Harris, S., & Seaborne, A. (2013). *SPARQL 1.1 query language*. W3C Recommendation.
<https://www.w3.org/TR/sparql11-query/>
- Marcus, R., Negi, P., Mao, H., Zhang, C., Alizadeh, M., Kraska, T., Papaemmanouil, O., & Tatbul, N. (2019). Neo: A learned query optimizer. *Proceedings of the VLDB Endowment*, 12(11), 1705–1718.
<https://doi.org/10.14778/3342263.3342644>
- Saleem, M., Potocki, A., Soru, T., Hartig, O., & Ngonga Ngomo, A.-C. (2018). CostFed: Cost-based query optimization for SPARQL endpoint federation. *Procedia Computer Science*, 137, 163–174.
- Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (2nd ed.). MIT Press.

Lattice-valued relational equations and inequations with complete residuated codomain lattice

Vanja Stepanović¹, Andreja Tepavčević^{2,3}

¹ Faculty of Agriculture, University of Belgrade, Serbia, vanja@agrif.bg.ac.rs

² Mathematical Institute, Serbian Academy of Sciences and Arts, Serbia

³ Faculty of Sciences, University of Novi Sad, Institute for Oncology and Radiology of Serbia, Belgrade, Serbia, andreja@dmf.uns.ac.rs

Fuzzy relational equations and inequations constitute an important research area in fuzzy logic and its applications. Since the introduction of fuzzy sets by Zadeh and lattice-valued sets by Goguen, numerous studies have investigated the structure of solution sets and the existence of extremal solutions for equations and inequations involving fuzzy and lattice-valued relations. In many practical situations, particularly in fuzzy control systems, it is essential to determine fuzzy relations that model the mapping from inputs to outputs. Such problems naturally lead to fuzzy and lattice-valued relational equations and inequations.

Residuated codomain lattices are especially convenient for the mentioned applications. Although introduced already in 1939, only starting from the end of the 20th and the beginning of the 21st century residuated lattices found their major application in fuzzy logic. We present a systematic study of the existence and characterization of extremal solutions to lattice-valued relational equations and inequations in case of a complete residuated codomain lattices. The focus is on equations and inequations containing on both sides terms constructed from known and unknown fuzzy relations by means of relational composition. The relational composition in the case of a complete residuated codomain lattice is defined using supremum and multiplication:

$$P \circ Q(x, y) = \bigvee_{z \in A} P(x, z) \otimes Q(z, y)$$

The associativity of the multiplication in residuated lattice, together with the infinite distributivity of the relational composition with respect to suprema, implies the associativity of this composition. Due to this associativity, many terms using composition are mutually equivalent, and we can omit parenthesis when composing more than two relations (whether known or unknown). Therefore, we can easily list all the equation and inequations containing at most one unknown relation on each side of the (in)equation. Most of the existing results and most of the results we got refer to them. (Stepanović, Tepavčević, 2025)

The investigation unifies and extends some known results concerning the existence of the greatest solution to relational equations and inequations. We also add some results concerning the existence of the least, maximal, and minimal solutions.

A central role is played by the infinite distributivity of relational composition with respect to suprema in complete residuated lattices. Using this property, the existence of the greatest solution is shown for a wide family of fuzzy relational equations and inequations. Namely, we proved that greatest solutions exist for all equations and inequations in which each side contains exactly one variable, regardless of whether the variables on the two sides coincide or differ. The same conclusion holds for systems of such equations and inequations. For several important classes, explicit expressions for the greatest solutions are given in terms of residuation.

We present the known solvability conditions for equations in which one side contains no variable, and the other contains exactly one variable appearing only once. (Sanchez, 1976; Stanković et al., 2017) Moreover, whenever the corresponding equation is solvable, its greatest solution coincides with the greatest solution of the associated inequation.

In contrast to the positive results for greatest solutions, the study demonstrates that greatest solutions need not exist when an equation contains more than one variable on one side, even in case of solvability. Counterexamples are constructed in more special cases of complete distributive codomain lattices, showing that the supremum of all solutions may fail to be a solution. Nevertheless, by combining properties of complete residuated lattices with Zorn's lemma we prove the existence of maximal solutions in such cases.

Applying some previous arguments (Stepanović, Tepavčević, 2026), we prove the existence of least solutions for several important classes of equations and inequations, especially those expressible through order-preserving operators on the lattice of fuzzy relations. It is shown that, under suitable lower-bound conditions, the least solution of an inequation coincides with the least solution of the corresponding equation. For equations in which variables occur on both sides, it is proved that the least solution always exists, while every solution is contained in a maximal solution.

We also present some negative results. It is known that the least solution to the typical equation $P \circ X = Q$ may not exist in case of complete, residuated codomain lattice. Moreover, we proved that even minimal solutions to such equations may fail to exist. This phenomenon persists even in more specialized structures such as complete Heyting lattices.

Overall, the paper provides a unified framework for understanding extremal solutions of lattice-valued relational equations and inequations in complete residuated settings. The results apply not only to general complete residuated lattices, but also to important special cases such as Boolean algebras, complete Heyting lattices, Gödel algebras, and Łukasiewicz algebras.

Keywords: Complete residuated lattice, lattice-valued relations, extremal solutions

References:

- Sanchez, E. (1976). *Resolution of composite fuzzy relation equations*, Information and Control, vol. 30(1), pp. 38-48.
- Stanković, I., Ćirić, M., Ignjatović, J. (2017). *Fuzzy relation equations and inequalities with two unknowns and their applications*, Fuzzy Sets and Syst, Vol. 322, pp. 86-105.
- Stepanović, V., Tepavčević, A. (2025). *A Note on the Solutions to Lattice-valued Relational Equations and Inequations*, Acta Polytechnica Hungarica 22/12, 27-46.
- Stepanović, V., Tepavčević, A. (2026) *Systems of fuzzy relational equations in the general lattice-valued case*, International Journal of Approximate Reasoning, Volume 195, 109703.

This research was supported by the Science Fund of the Republic of Serbia, Grant no 6565, Advanced Techniques of Mathematical Aggregation and Approximative Equations Solving in Digital Operational Research- AT-MATADOR. The authors also gratefully acknowledge the financial support of the Ministry of Science, Technological Development and Innovation of the Republic of Serbia (Grants No. 451-03-34/2026-03/200116, 451-03-33/2026-03/200125, 451-03-34/2026-03/200125).

Temporal Dynamics of Codon Usage Bias in SARS-CoV-2 Proteins

Biljana T. Stojanović¹, Saša N. Malkov², Miloš V. Beljanski³, Nenad S. Mitić²

¹ Mathematical institute of SASA, Belgrade, Serbia, bstojanovic@mi.sanu.ac.rs

² Faculty of Mathematics, University of Belgrade, Belgrade, Serbia, sasa.malkov@matf.bg.ac.rs, nenad.mitic@matf.bg.ac.rs

³ Institute for General and Physical Chemistry, Belgrade, Serbia, mbel@matf.bg.ac.rs

Motivation and Aim: Codon Usage Bias (CUB) refers to the non-random use of synonymous codons and is associated with translational efficiency, host adaptation, and genome evolution. Codon usage patterns reflect functional constraints and serve as a post-transcriptional regulatory mechanism affecting multiple aspects of gene expression, including mRNA secondary structure and stability, translation kinetics, and co-translational protein folding. Previous studies suggest that CUB is mainly driven by three major evolutionary forces: (a) GC-biased gene conversion, which preferentially increases the frequency of GC-ending codons, (b) mutational biases, which generally favor AT-ending codons, and (c) natural selection, which promotes codons that improve translational efficiency and accuracy. The relative contributions of these forces differ across organisms and genes, resulting in diverse patterns of synonymous codon usage. Characterizing CUB in particular organisms or biological systems is important for understanding evolutionary processes and has practical applications in genetic engineering and synthetic biology, including the design of synthetic genes and other genetic constructs (Arshad et al., 2026). This study aims to characterize and provide deeper insight into the temporal dynamics of SARS-CoV-2 evolution by applying large-scale computational analysis of genomic data to identify interpretable patterns of codon usage variation and evolutionary change at both nucleotide and protein levels. Particular emphasis is placed on codon-level changes, including those associated with synonymous and non-synonymous substitutions, as well as position-specific changes at homologous amino acid positions over time.

Materials and Methods: We analyzed ~5 million SARS-CoV-2 protein-coding sequences (CDSs) collected between December 15, 2019 and May 22, 2023, using both the complete dataset and 20 temporal subsets obtained by partitioning isolate collection dates into discrete time intervals. Four proteins—surface glycoprotein (spike, S protein), nucleocapsid phosphoprotein (N protein), ORF1a polyprotein, and ORF1ab polyprotein—were selected for detailed analysis based on their higher proportion of distinct CDSs (>3% after removal of identical sequences), reflecting greater sequence diversity than the remaining proteins.

CUB was characterized using the codon-specific measures Relative Codon Frequency (RF) and Relative Synonymous Codon Usage (RSCU) (Sharp & Li, 1986), and the gene-level measures Effective Number of Codons (ENC) (Sun et al., 2013) and Relative Codon Bias Score (RCBS) (Fox & Erill, 2010). RF and RSCU provide complementary perspectives on codon usage. RF quantifies the observed relative frequency of a codon within its synonymous codon family, whereas RSCU compares the observed codon frequency with that expected under equal usage of synonymous codons. As a normalized measure, RSCU enables the identification of preferred (RSCU > 1), over-represented (RSCU > 1.6), and under-represented (RSCU < 0.6) codons. Similarly, ENC and RCBS characterize gene-level CUB from different perspectives. ENC quantifies the overall degree of CUB, whereas RCBS accounts for background nucleotide composition at all three codon positions when evaluating codon preference. Two ENC variants were used: standard ENC and ENC-rsl-split, in which the R, S, and L codon families are treated separately.

For temporal CUB analysis, protein sequences were aligned to the corresponding reference protein sequences of the SARS-CoV-2 reference isolate (NC_045512.2) using MAFFT (Multiple Alignment using Fast Fourier Transform). To preserve codon information, alignment was performed at the protein level

rather than directly on nucleotide sequences, since nucleotide alignments may introduce frame-disrupting insertions and deletions that prevent reliable codon-based comparisons. Proteins of the same type were grouped according to identical amino acid sequences. For each group, one representative sequence was aligned to the corresponding reference protein, and the obtained alignment pattern was applied to all sequences within the same group. Position-specific codon frequencies were then calculated by mapping codons from nucleotide sequences onto homologous amino acid positions in the reference protein. Codons corresponding to deletions were excluded, whereas insertions relative to the reference were skipped. The same procedure was applied to the complete dataset and each of the 20 temporal groups. This workflow enabled interpretable representation and comparison of codon-level evolutionary patterns across homologous protein positions in millions of sequences while preserving biological correspondence between aligned codons. The resulting position-specific codon profiles provided the basis for calculating RF values at homologous protein positions and analyzing temporal changes in codon frequency patterns.

Results: Analyses were performed separately for each protein type. Similar gene-level RCBS, ENC, and ENC-rsl-split values, as well as overall CUB patterns, were observed for the ORF1a and ORF1ab polyproteins. Based on RCBS, the ORF1ab polyprotein exhibited the weakest CUB, followed by the ORF1a polyprotein and the S protein, whereas the N protein showed the strongest. Codon-level RSCU analysis identified over-represented and under-represented codons. Alignment-based RF analyses revealed (a) temporal changes in codon frequencies associated with synonymous and non-synonymous substitutions, (b) low-frequency codons, and (c) positions with significant codon frequency shifts, predominantly occurring in intrinsically disordered regions (Giri et al., 2021). Position-specific changes in codon frequencies were also analyzed across WHO-defined lineages.

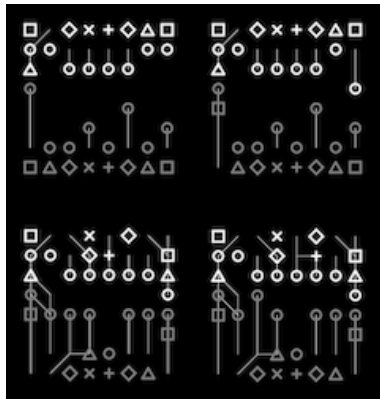
Conclusions: CUB patterns across SARS-CoV-2 proteins change over time, both globally and at homologous amino acid positions, with temporal patterns corresponding to the emergence of major SARS-CoV-2 variants. We observed shifts in codon frequencies, position-specific codon changes, and lineage-associated patterns among WHO-defined lineages. The most pronounced temporal CUB changes were observed for the S and N proteins, where specific codon frequencies exhibited distinct temporal trajectories, including relatively stable patterns or decreases occurring at early, intermediate, or late phases of SARS-CoV-2 evolution within the analyzed period. These findings suggest that codon usage is a dynamic feature of viral evolution rather than a static genomic characteristic. This study is part of an ongoing analysis, and further validation and refinement are required to fully characterize the biological significance of the observed temporal codon usage patterns and position-specific changes. Interpretable computational analysis of codon usage dynamics may provide additional markers of evolutionary change and contribute to understanding viral adaptation over time.

Keywords: SARS-CoV-2, codon usage, temporal dynamics

References:

- Arshad, M., Uchmanowicz, M., Rana, V., Trost, B., Scherer, S. W., & Rafiq, M. A. (2026). Mapping the inter- and intra-genic codon-usage landscape in *Homo sapiens*. *NAR Genomics and Bioinformatics*, 8(1), Article lqag024. <https://doi.org/10.1093/nargab/lqag024>
- Fox, J. M., & Erill, I. (2010). Relative codon adaptation: A generic codon bias index for prediction of gene expression. *DNA Research*, 17(3), 185–196. <https://doi.org/10.1093/dnares/dsq012>
- Giri, R., Bhardwaj, T., Shegane, M., Gehi, B. R., Kumar, P., Gadhave, K., Oldfield, C. J., & Uversky, V. N. (2021). Understanding COVID-19 via comparative analysis of dark proteomes of SARS-CoV-2, human SARS, and bat SARS-like coronaviruses. *Cellular and Molecular Life Sciences*, 78(4), 1655–1688. <https://doi.org/10.1007/s00018-020-03603-x>
- Sharp, P. M., & Li, W. H. (1986). Codon usage in regulatory genes in *Escherichia coli* does not reflect selection for "rare" codons. *Nucleic Acids Research*, 14(19), 7737–7749. <https://doi.org/10.1093/nar/14.19.7737>
- Sun, X., Yang, Q., & Xia, X. (2013). An improved implementation of effective number of codons (Nc). *Molecular Biology and Evolution*, 30(1), 191–196. <https://doi.org/10.1093/molbev/mss201>

MACHINE LEARNING



Smart CNC Manufacturing: Application of Machine Learning for Predictive Monitoring of Tool Condition and Machining Quality

Obrad Aničić¹

¹ The Academy of Applied Studies Polytechnic, Belgrade, Serbia, oanitic@politehnika.edu.rs

The increasing digitalization of manufacturing systems has accelerated the application of machine learning methods for improving production efficiency, machining quality, and operational reliability (Kusiak, 2018; Wuest et al., 2016). CNC machine tools continuously generate process data that can be used for predictive monitoring and intelligent production decision-making (Teti et al., 2010; Wuest et al., 2016). However, many existing approaches concentrate on the development of individual prediction models, while less attention is given to their integration with process monitoring and operational decision-support functions. This paper proposes an integrated machine learning framework for predictive monitoring of cutting tool condition and machining quality in CNC manufacturing systems. The framework integrates CNC process-data acquisition, data preprocessing, feature engineering, predictive modelling, multi-sensor information fusion, process-state assessment, and decision support within a unified methodological architecture. It is designed to accommodate different machine learning methods and to combine information obtained from multiple process-monitoring sensors, thereby providing a structured basis for data-driven supervision of machining processes (Teti et al., 2010; Wuest et al., 2016).

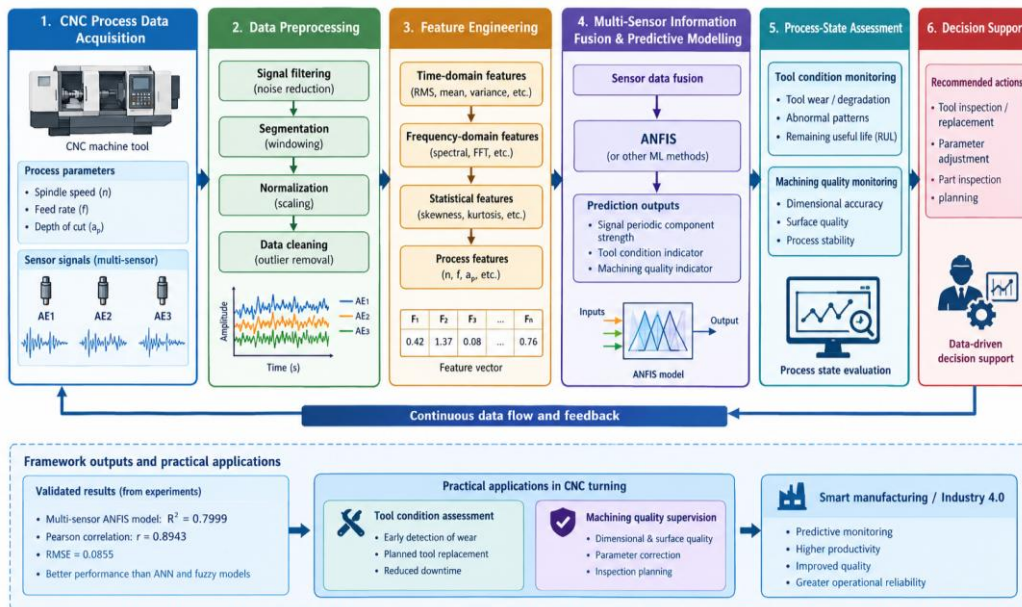


Figure 1 – Proposed machine learning framework for predictive monitoring of tool condition and machining quality in CNC manufacturing systems

The results are supported by previously validated experimental data obtained from CNC machining. In the reported experiments, data acquired from three acoustic-emission sensors installed on a CNC lathe were processed using spindle speed, feed rate, and depth of cut as input process variables. An Adaptive Neuro-Fuzzy Inference System (ANFIS) was applied for multi-sensor data fusion and prediction of the strength of the signal periodic component (Jovic et al., 2017). The ANFIS model achieved $R^2 = 0.7999$, a Pearson correlation coefficient of $r = 0.8943$, and $RMSE = 0.0855$. For comparison, the ANN model achieved $R^2 = 0.7851$ and $RMSE = 0.0893$, while the fuzzy model achieved $R^2 = 0.7947$ and $RMSE = 0.0894$ (Jovic et al., 2017). These quantitative results demonstrate that neuro-fuzzy modelling can effectively integrate information from multiple acoustic-emission sensors and provide a reliable

representation of the monitored CNC machining process. Sensitivity analysis further showed that spindle speed was the most influential individual process parameter, while the combination of spindle speed and depth of cut produced the lowest training and testing RMSE among the evaluated two-parameter combinations. These results demonstrate how the proposed framework can transform raw machining data into information useful for process-state assessment and predictive supervision. In a practical CNC turning operation, sensor data can be continuously acquired during machining and fused with process parameters such as spindle speed, feed rate, and depth of cut. Deviations in the monitored acoustic-emission response can then be identified and evaluated against the expected process state. For example, progressive changes in the monitored response can provide an early indication that tool condition or machining stability requires attention, supporting planned tool inspection or replacement before excessive degradation leads to unacceptable machining quality. In a second practical application, the same monitoring information can be used for machining-quality supervision: changes in spindle load, vibration or acoustic-emission response, together with machining parameters, can indicate operating conditions associated with dimensional or surface-quality deviations. The resulting information can support corrective actions such as adjustment of machining parameters, additional inspection of the produced part, or planned tool replacement (Teti et al., 2010; Davim, 2021).

The principal result and contribution of the research is therefore the development of a coherent operational framework that connects manufacturing data acquisition, machine learning-based analysis, multi-sensor information fusion, process-state assessment, and production decision support. The framework extends machine learning from an isolated prediction task toward continuous predictive monitoring of CNC operations. The previously validated experimental results provide quantitative evidence supporting the applicability of the underlying multi-sensor modelling approach, while the integrated framework establishes a methodological basis for extending these capabilities to tool-condition assessment, machining-quality monitoring, inspection planning, and parameter adjustment. Because the framework is algorithm-independent, it can be adapted to different CNC machines, machining processes, sensor configurations, and production requirements. The proposed approach thus provides a practical basis for moving from reactive detection of machining problems toward predictive and data-driven process supervision in intelligent manufacturing systems in accordance with Industry 4.0 principles (Kusiak, 2018; Wuest et al., 2016). The principal contribution of this work is therefore not limited to the development of an individual prediction model, but to the integration of data acquisition, machine learning-based analysis, multi-sensor information fusion, process-state assessment, and production decision support within a structured operational framework. The previously validated experimental results provide quantitative evidence supporting the applicability of the approach, while the proposed framework establishes a methodological basis for extending predictive monitoring toward tool-condition and machining-quality assessment in intelligent CNC manufacturing systems. The approach supports the transition from conventional reactive monitoring toward predictive and data-driven process supervision in accordance with Industry 4.0 principles (Kusiak, 2018; Wuest et al., 2016).

Keywords: CNC machining, smart manufacturing, machine learning, predictive monitoring, tool condition monitoring, multi-sensor data fusion, ANFIS, Industry 4.0

References:

- Breiman, L. (2001). *Random Forests. Machine Learning*, 45(1), 5–32.
- Kusiak, A. (2018). *Smart manufacturing. International Journal of Production Research*, 56(1–2), 508–517.
- Teti, R., Jemielniak, K., O'Donnell, G., & Dornfeld, D. (2010). *Advanced monitoring of machining operations. CIRP Annals*, 59(2), 717–739.
- Wuest, T., Weimer, D., Irgens, C., & Thoben, K. D. (2016). *Machine learning in manufacturing: Advantages, challenges, and applications. Production & Manufacturing Research*, 4(1), 23–45.
- Davim, J. P. (2021). *Machining and Machine Tools*. Springer.
- Jovic, S., Anicic, O., & Jovanovic, M. (2017). *Adaptive neuro-fuzzy fusion of multi-sensor data for monitoring of CNC machining. Sensor Review*, 37(1), 78–81. <https://doi.org/10.1108/SR-06-2016-0107>

Fractal and textural biomarkers in the diagnostics of pediatric inflammatory bowel disease (IBD)

Andjelija Ž. Ilić¹, Ivan D. Milovanovich^{2,3}, Vedrana Makević⁴, Nevena Popovac², Radmila Janković⁵, Jelena Trajković¹, Andrija Vuković⁴, Jovan Jevtić⁵, Silvio R. de Luka⁴

¹ Institute of Physics Belgrade, University of Belgrade, Pregrevica 118, 11080 Zemun-Belgrade, Serbia, andjelijailic@ieee.org

² University Children's Hospital, University of Belgrade, Tiršova 10, 11000 Belgrade, Serbia, imilovanovich@gmail.com

³ Faculty of Medicine, University of Belgrade, Dr Subotica 1, 11000 Belgrade, Serbia

⁴ Institute of Pathological Physiology, Faculty of Medicine, University of Belgrade, Dr Subotica 9, 11000 Belgrade, Serbia

⁵ Institute of Pathology, Faculty of Medicine, University of Belgrade, Dr Subotica 1, 11000 Belgrade, Serbia

Inflammatory bowel disease (IBD), including Crohn's disease and ulcerative colitis (UC), is characterized by chronic relapsing intestinal inflammation (Zhang and Li, 2014). It develops through a combination of inherent, environmental, microbiome, and immunological factors. Its incidence has notably increased over the last seven to eight decades, often affecting children and adolescents. Its early diagnosis is crucial to adequate and timely treatment, yet it can be a challenge even for experienced pathologists and gastroenterologists. Use of the decision support software tools could be of significance in discerning the signs of IBD from other disorders, and particularly, discerning between the mentioned two subtypes, whose appearances can be similar.

A major challenge in the development of decision support software for pediatric patients lays in the relatively small availability of datasets from previous studies, small number of such studies overall, and difficulties in collecting patients and data for the study. Computational image analysis has been shown useful in extracting features of interest, which can aid further statistical analyses or data classification even in the cases of very limited numbers of specimens. Amongst such features, fractal tissue and cell descriptors have been shown as one of the most useful tools to assess the structural complexity (Bauman, *et al.*, 2020). Morphological descriptors, size and shape, have also been used with success.

It has been observed in certain diseases, e.g. cancers, that there is an increase in the chromatin compaction, which has been linked to altered chromatin texture patterns in disease. Altered texture may have resulted from the chromatin redistribution, condensation or changes in the ratio of two types of chromatin, or simply cell damage. Such changes, when registered in the high-resolution microscopic images, could become digital biomarkers for early diagnosis. Given a complex IBD pathogenesis (Zhang and Li, 2014), involving multiple factors, it was reasonable to check if there were any alterations in textures of intestinal cell nuclei in IBD. Recently, we estimated structural complexity of the patterns/textures of intestinal cell nuclei using fractal analysis (Makević, *et al.*, 2023). Data was collected from 24 pediatric patients and from healthy controls, from all seven intestinal segments. Further, we employed gray-level co-occurrence matrix (GLCM) features of cell nuclei, robust and not overly sensitive to image resolution, as other type of textural biomarkers. Some morphological differences between the studied groups were observed as well. We used the collected data as a basis for the subsequent machine learning in conditions of limited number of patients. Support vector machine (SVM) based data classification satisfactorily discerned UC versus the Crohn's disease subtype of the IBD, versus the control group (Milovanovich, *et al.*, 2026). To the best of our knowledge, there is only one more pediatric study employing machine learning in the diagnosis

of Crohn's disease (Guez, et al., 2022). We describe our methodology and the obtained results, comparing our study with the one described in (Guez, et al., 2022).

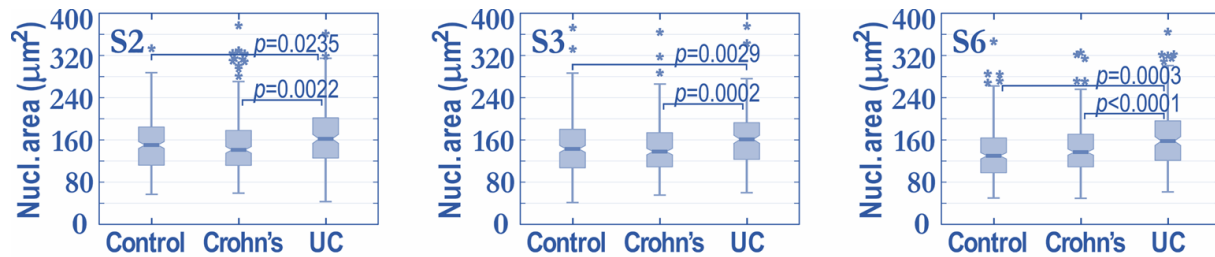


Fig. 1. Intergroup comparison of the sizes of cell nuclei in the second, third and sixth segment of the intestine could be used to discern the UC subtype of the IBD (Milovanovich, et al., 2026).

Keywords: Computational image analysis, Cell fractal and textural descriptors, Machine learning (ML), Support vector machine (SVM), Inflammatory bowel disease (IBD)

References:

- Bauman, N., Ilić, A.Ž., Lijeskić, O., et al. (2020). Computational image analysis reveals the structural complexity of *Toxoplasma gondii* tissue cysts. *PLoS One* **15**(8), p. e0234169. <https://doi.org/10.1371/journal.pone.0234169>
- Guez, I., Focht, G., Greer, M-L.C., et al. (2022). Development of a multimodal machine-learning fusion model to non-invasively assess ileal Crohn's disease endoscopic activity. *Comput. Methods Programs Biomed.* **227**, p. 107207. <https://doi.org/10.1016/j.cmpb.2022.107207>
- Makević, V., Milovanovich, I.D., Popovac, N., et al. (2023). Fractal parameters as independent biomarkers in the early diagnosis of pediatric onset inflammatory bowel disease. *Fractal Fract* **7**, p. 619. <https://doi.org/10.3390/fractalfract7080619>
- Milovanovich, I.D., Makević, V., Popovac, N., et al. (2026). Morphometric and GLCM textural computational analysis for improved IBD diagnostics in children. Under review.
- Zhang, Y.-Z., Li, Y.-Y. (2014). Inflammatory bowel disease: Pathogenesis. *World J. Gastroenterol.* **20**(1), pp. 91-99. <https://doi.org/10.3748/wjg.v20.i1.91>

Wavelet-Based Optimization of Temporal Windows for LSTM Benzene Concentration Forecasting

Goran Keković¹, Vladimir Mikić², Miloš Ilić³, Rade Božović⁴

¹ Faculty of Information Technology, Alfa BK University, goran.kekovic@alfa.edu.rs

¹ Faculty of Information Technology, Alfa BK University, vladimir.mikic@alfa.edu.rs

² Faculty of Information Technology, Alfa BK University, milos.ilic@alfa.edu.rs

³ Faculty of Information Technology, Alfa BK University, rade.bozovic@alfa.edu.rs

This paper investigates the optimization of the moving time window for air pollution prediction using Long Short-Term Memory (LSTM) neural networks. The analysis was conducted on data collected in an Italian city between March 2004 and February 2005 through hourly measurements obtained from a sensor network (Vito, 2008). The target variable was benzene (C₆H₆) concentration, while the input variables included CO, NO_x, NO₂, NMHC, O₃, temperature, relative humidity, and absolute humidity. After removing missing and incomplete records, 8991 samples were retained for analysis. Considering the non-stationary and nonlinear characteristics of air pollution time series, wavelet analysis was applied during data preprocessing. To identify trends and patterns in the time–frequency domain, the Relative Wavelet Energy (RWE) parameter was employed (Polat & Güneş, 2007).

$$RWE_{j_0} = \frac{\sum_{k=1}^{N_k} |C|_{k,j_0}^2}{\sum_{k=1}^{N_k} |C|_{k,j_0}^2 + \sum_{k=1}^{N_k} \sum_{j=1, j \neq j_0}^M |C|_{k,j}^2} \quad (1)$$

In the applied model ($N_k=8991$, $M=14$), the parameter RWE was calculated within the range of pollution fluctuations from hourly to annual variations ($j \leq M$). Therefore, only the energy contribution of detail coefficients was analyzed, while long-term multiannual trends were excluded ($j > M$, $A_{k,j}=0$). The following figures present the parameter values at different resolution levels, together with a representative daily benzene concentration profile.

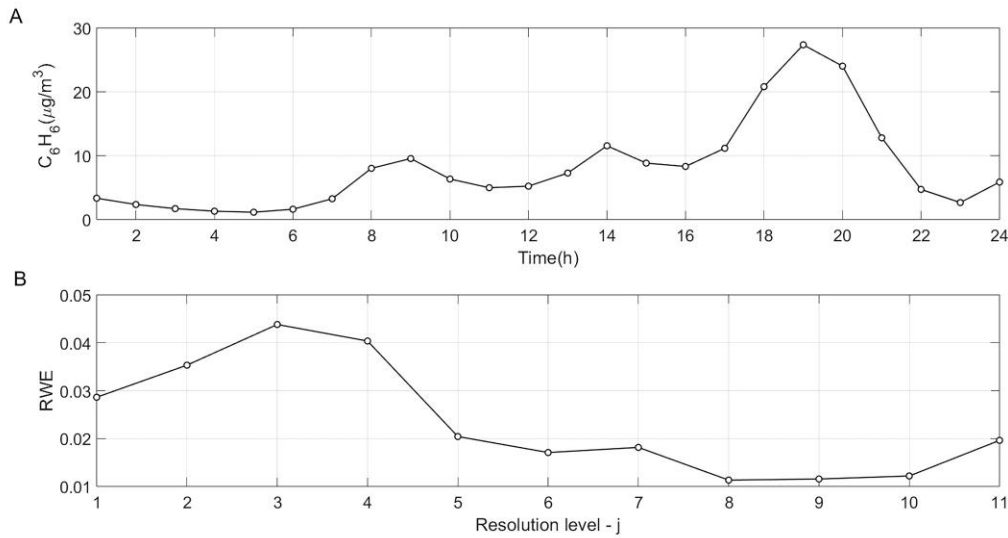


Fig. 1 - Daily profile of benzene concentration (A) and distribution of RWE parameter values of the benzene time series as a function of wavelet resolution level (B)

As shown in Fig. 1A, the daily pollution profile is most pronounced between 8 h and 20 h, while Fig. 1B indicates a peak of the RWE parameter at resolution level, $j=3$. According to Eq. (2), this level corresponds to the pseudofrequency interval associated with the time range $8.03 \text{ h} \leq \Delta t \leq 16.06 \text{ h}$, obtained for $f_c = 0.996$ (arbitrary units - Haar wavelet) and $f_s = 1/3600 \text{ Hz}$. Based on these findings, LSTM simulations performed for different moving window lengths and a prediction horizon of 10 h identified an optimal window length of 8 h. The simulation results, presented in Fig. 2B, show the

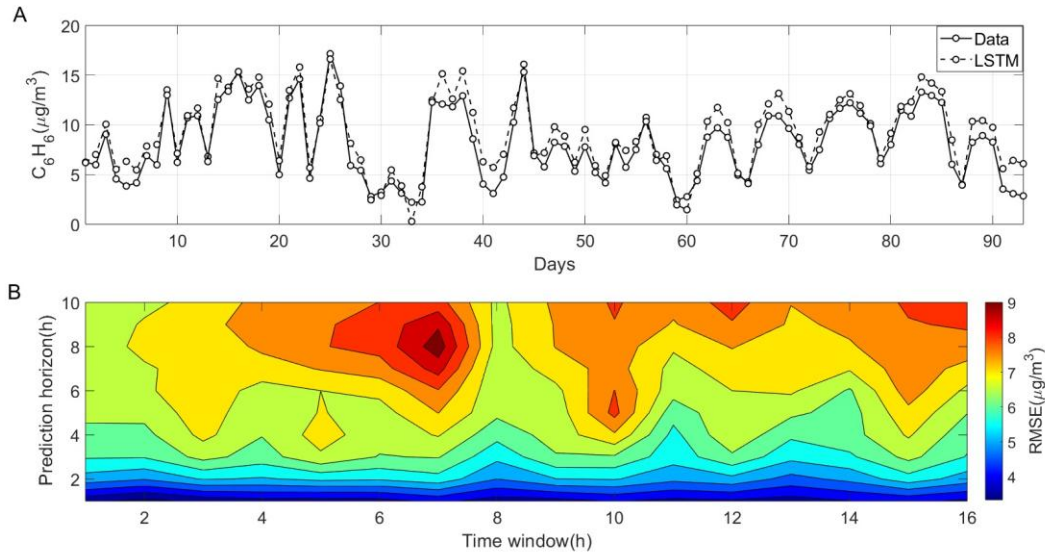


Fig. 2 - Original and predicted time series with LSTM network of benzene concentration (A) and dependence of RMSE of benzene concentration on time window and prediction horizon (B)

distribution of the root mean square error (RMSE) as a function of the moving time window and prediction horizon. The lowest average RMSE across all prediction horizons was obtained for the 8 h window, with a value of $6.09 \mu\text{g}/\text{cm}^3$. These findings confirm that the temporal patterns identified through wavelet analysis correspond to the optimal prediction windows (Chen et al., 2020). Fig. 2A presents the model validation on the testing dataset (25%), representing unseen data for the LSTM model, while Fig. 2B indicates that the highest prediction accuracy was achieved for horizons up to 3 h.

Keywords: LSTM neural networks, wavelet analysis, air pollution, benzen

References:

- Chen, X., Yin, L., Fan, Y., Song, L., Ji, T., Liu, Y., Tian, J., & Zheng, W. (2020). *Temporal evolution characteristics of PM2.5 concentration based on continuous wavelet transform*. *Science of the Total Environment*, 699, Article 134244. <https://doi.org/10.1016/j.scitotenv.2019.134244>.
- Polat, K., & Güneş, S. (2007). *Classification of epileptiform EEG using a hybrid system based on decision tree classifier and fast Fourier transform*. *Applied Mathematics and Computation*, 187(2), 1017–1026. <https://doi.org/10.1016/j.amc.2006.09.022>.
- Vito, S. (2008). Air Quality [Dataset]. UCI Machine Learning Repository. <https://doi.org/10.24432/C59K5F>.

Take a Sad Song and Make It Better: Iterative Plugin Refinement in Song-Success Prediction

Stefan Nikolić, Vladan Devedžić

¹ University of Belgrade, Faculty of Organizational Sciences, Belgrade, Serbia.

Traditional Hit Song Science (HSS) models treat track popularity as a static, binary classification problem, utilizing black-box architectures that offer zero actionable feedback to song creators. This paper introduces a dynamic, decentralized multi-agent architecture designed as an explainable decision support system. By decomposing the complexity of commercial success into autonomous expert decisions covering audio, historical context, market trends and short form viral potential the system shifts the paradigm from blind prediction to interactive diagnostics. Integrating these heterogeneous signals through a trainable PyTorch neural orchestrator allows human experts to dynamically tweak node weights. Simulated results demonstrate that the proposed framework empowers users with transparent, data-driven and actionable insights.

Introduction: The evolution of musical genres is a highly non-linear, cyclical process, where past aesthetics routinely re-emerge and modern short form video platforms introduce extreme market volatility. Predicting song success within the Hit Song Science (HSS) and Music Information Retrieval (MIR) domains is fundamentally hard to do using rigid, deterministic black-box solutions. The inherent limitation of existing models is opacity providing only a flat output value leaves the user unable to understand the underlying causality, rendering good predictions unjustified and bad ones unexplainable. To bridge this gap, this work proposes an iterative multi-agent framework where users receive granular, multidimensional feedback. Instead of delivering an unexplainable verdict, the final prediction is a result of a shallow neural network that treats autonomous expert agents as its functional nodes. The resulting data-driven architecture ensures that the user is not left in an information void, but is instead empowered with actionable knowledge.

System Architecture: The proposed architecture grants users the operational flexibility to interactively tweak individual node weights, allowing industry experts to prioritize specific market signals based on their distribution strategy. By serving as an explainable decision-support consultant, this system transitions HSS from a passive oracle to an active optimization tool designed to maximize a song's success.

As illustrated in Table 1, the system architecture is built as a multi-agent system combining digital signal processing, API consumption, large language models (LLMs) and vector geometry in order to provide contextualized data. The mathematical integration of these independent signals into a unified prediction is performed by a dedicated neural network. All audio-based agents consume raw audio files (.mp3, .wav), while the context agent consumes metadata.

Table 1 - The agents in the multi-agent system architecture

Name	Core logic
Audio agent	Pydub audio segmentation; Gemini 2.0 Flash structural/acoustic evaluation via structured JSON.
Context agent	MusicBrainz historical data cross-referenced with Spotify API via logarithmic follower

	scaling.
Trend agent	L2 Norm Euclidean distance mapping against dynamic market hit vectors with harmonic bonus.
Viral agent	DSP short form video suitability tracking via micro-energy peaks and vocal presence.

Neural Orchestrator & Integration: The central orchestration model, implemented in PyTorch, is a feed-forward network designed to act as a trainable consensus gate. Rather than acting as an unexplainable classifier, the network maps the four-dimensional input vector (comprising the normalized 0.0 to 1.0 signals from the Audio, Context, Trend, and Viral agents) into a single linear layer, which is then stabilized via a Sigmoid activation function. This structural setup allows the network to learn the optimal trade-offs and relationships between deterministic structural data and semantic AI feedback during the optimization phase.

Results: To evaluate the initial viability of the framework, a simulation was conducted using a benchmarking dataset of 2,000 song profiles. The orchestrator was trained using Stochastic Gradient Descent (SGD) over 100 epochs to observe the Mean Squared Error (MSE) loss trajectory. Crucially, alongside the global network convergence, the multi-agent system exposes the decoupled activation outputs of each individual agent. In an academic context, these parallel outputs function as independent diagnostic dimensions, providing a granular decision-support matrix that serves as an expert consultant prior to the final neural synthesis.

The training baseline initialized at a global MSE loss of 0.116407 (Epoch 0) and actively adjusted via gradients to 0.111995 (Epoch 10) and 0.111796 (Epoch 30), before stabilizing into a local convergence plateau of 0.111790 at Epoch 50 and a final 0.111789 at Epoch 99. This empirical stabilization confirms that the framework successfully harmonizes global predictive optimization with transparent, agent-specific metrics without optimization oscillations. By isolating individual plugin activations alongside this rapid convergence trajectory, the architecture ensures that the system does not act as a closed black-box, but rather as an explainable, multidimensional consulting tool open to future algorithmic adjustments.

Conclusion: This paper presents a dynamic, multi-agent framework that successfully transitions Hit Song Science from speculative black-box predictions into an explainable decision-support system. By decoupling commercial success into autonomous expert decisions and integrating them through a predictable neural orchestrator, the architecture provides song creators with actionable diagnostics. Future work will focus on expanding the benchmarking datasets, refining existing prompt strategies for tighter AI integration, and introducing additional specialized agent nodes to make the consensus network more robust.

Keywords: Multi-agent systems, explainable AI, machine learning, hit song science, neural networks

References:

- H.S. Saragih, Predicting song popularity based on Spotify's audio features: insights from the Indonesian streaming users, *J. Manag. Anal.* 10 (2023) 693–709.
- M. Fell, E. Cabrio, E. Korfed, M. Buffa, F. Gandon, Love Me, Love Me, Say (and Write!) that You Love Me: Enriching the WASABI Song Corpus with Lyrics Annotations, (2020).
- G. Tzanetakis, G. Essl, P. Cook, Automatic Musical Genre Classification Of Audio Signals, (2001).

Interpreting Machine Learning Model Performance in Biomedicine: From Statistical Performance to Clinical Decision-Making

Marina Popović Krneta¹, Miljana Tanić²

¹ Department of Nuclear Medicine, Institute for Oncology and Radiology of Serbia, Belgrade, Serbia, marina.popovic1989@gmail.com

² Experimental Oncology Department, Institute for Oncology and Radiology of Serbia, Belgrade, Serbia, tanic.miljana@ncrc.ac.rs

Machine learning (ML)-based clinical prediction models are increasingly used to support diagnosis, prognosis, treatment selection, and risk prediction in biomedicine. Their successful application in clinical practice, however, depends not only on statistical performance but also on whether their evaluation reflects the clinical question the model intends to answer. In current practice, statistical performance is often treated as sufficient evidence that a model will improve patient care. One of the clearest examples of this tendency is the widespread reliance on the area under the receiver operating characteristic curve (AUC). Although the AUC is an essential measure of discrimination, it does not indicate whether a model is appropriate for the clinical question being asked. Used in isolation, it may therefore encourage conclusions about clinical usefulness that are not supported by the intended application.

Bridging this gap requires rethinking how prediction models are evaluated for clinical use. Model performance should therefore be viewed as a multidimensional concept comprising discrimination, calibration, classification performance, and clinical utility (Van Calster et al., 2024). However, simply reporting a broader range of performance measures does not guarantee clinically relevant evaluation; the selection and interpretation of these measures must be guided by the clinical question the model is intended to answer. The proposed framework begins by aligning distinct clinical objectives with the performance measures most relevant to each decision context (Table 1). While these dimensions of model performance are well established in the literature, the proposed framework takes the clinical objective as its starting point, using it as the basis for selecting and interpreting the appropriate performance measures.

Table 1. Selecting Performance Measures According to the Intended Clinical Objective

Primary clinical objective	Least acceptable error	Performance measures to prioritize
Avoid missing patients who require further evaluation or treatment	False negatives	Sensitivity, Negative Predictive Value
Avoid unnecessary diagnostic or therapeutic procedures	False positives	Specificity, Positive Predictive Value
Support individualized risk estimation	Inaccurate risk estimation	Model Calibration

Once the relevant performance measures have been identified, the next step is to determine the decision threshold that best reflects the clinical objective. Despite this, most ML software packages apply a default decision threshold of 0.5, implicitly assuming that the same threshold is appropriate regardless of the clinical context. A commonly used statistical alternative is the Youden index, which identifies a threshold that balances sensitivity and specificity. Although useful as a statistical starting point, this approach assumes that false-positive and false-negative results carry comparable clinical consequences (Deng et al., 2025). For example, in cancer screening, where avoiding missed diagnoses is the primary clinical objective, a lower decision threshold may be preferred to maximize sensitivity despite an increase in

false-positive results. Consequently, the statistically optimal threshold is not necessarily the clinically optimal one.

The interpretation of model performance does not end with selecting appropriate performance measures and decision thresholds. It must also account for the characteristics of the population in which the model will be applied. Positive Predictive Value (PPV) and Negative Predictive Value (NPV) depend on disease prevalence and therefore cannot be interpreted independently of the target population. For example, a model developed in a population with a high prevalence of disease may achieve a substantially lower PPV when applied in a lower-prevalence screening population, despite similar discrimination. Therefore, performance estimates should always be interpreted within the context of the target population in which the model is intended to be used, rather than assuming that they are universally transferable (Park & Han, 2018; Collins et al., 2024).

An example of the practical application of these principles is provided by our previous clinical prediction study, in which model selection prioritized a sensitivity of at least 95% and the highest specificity achievable at that level of sensitivity, while also considering the degree of overfitting (Popović Krneta et al., 2023).

In conclusion, this framework shifts the evaluation of clinical prediction models from a statistically driven to a clinically driven approach. By taking the primary clinical objective as its starting point, it provides a structured workflow that aligns performance measures, guides the selection of decision thresholds according to the clinical consequences of error, and contextualizes model performance within the target population. The novelty lies in the synthesis of these established methodological principles into a structured workflow for clinicians. Rather than reducing model evaluation to isolated statistical metrics, this approach provides a practical pathway for translating model performance into clinically informed decision-making.

Keywords: Machine learning, prediction models, performance measures, clinical decision-making, decision threshold

References:

- Collins, G. S., Dhiman, P., Andaur Navarro, C. L., et al. (2024). TRIPOD+AI: Consolidated statement for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, 385, e078378.
- Deng, J., Elghobashy, M. E., Zang, K., et al. (2025). So You've Got a High AUC, Now What? An Overview of Important Considerations When Bringing Machine-Learning Models from Computer to Bedside. *Medical Decision Making*, 45(6), 640–653.
- Park, S. H., & Han, K. (2018). Methodologic Guide for Evaluating Clinical Performance and Effect of Artificial Intelligence Technology for Medical Diagnosis and Prediction. *Radiology*, 286(3), 800–809.
- Popović Krneta, M., Šobić Šaranović, D., Mijatović Teodorović, L., et al. (2023). Prediction of Cervical Lymph Node Metastasis in Clinically Node-Negative T1 and T2 Papillary Thyroid Carcinoma Using Supervised Machine Learning Approach. *Journal of Clinical Medicine*, 12, 3641.
- Van Calster, B., Collins, G. S., Vickers, A. J., Wynants, L., Kerr, K. F., Barreñada, L., Varoquaux, G., Singh, K., Moons, K. G. M., Hernandez-Boussard, T., Timmerman, D., McLernon, D. J., Van Smeden, M., & Steyerberg, E. W. (2024). Performance evaluation of predictive AI models to support medical decisions: Overview and guidance. *Artificial Intelligence in Medicine*.

NATURAL LANGUAGE PROCESSING



Same Pace, Different Path: An NLP-based Analysis of Gender Disparities in IT Careers

Marjana Gligorijević¹, Milena Vujošević Janičić¹, Jelena Graovac¹

¹ Faculty of Mathematics, University of Belgrade, Serbia, marjana.gligorijevic@matf.bg.ac.rs,
milena.vujosevic.janicic@matf.bg.ac.rs, jelena.graovac@matf.bg.ac.rs

Women remain underrepresented in IT across Europe (Penzenstadler et al., 2025), yet existing data-driven studies largely focus on the US or Western Europe (Hickey et al., 2022; Young et al., 2023), leaving Serbia and the broader Central and Eastern European region underexplored. This study presents a computational analysis of gender-based disparities across four dimensions: educational attainment, career progression (advancement in job seniority), career interruptions, and language of self-presentation based on profile data for IT professionals in Serbia. The study aims to identify whether measurable gender-based disparities exist across these dimensions, and to characterize their nature and magnitude. In addition, the study introduces a transparent, multi-layer gender classification methodology that may serve as a reference for future computational social science research in similar regional contexts.

A dataset for this study was assembled from a corpus of 17,896 professional profiles of individuals based in Serbia, obtained from a third-party provider of publicly available profile data. Only publicly accessible fields were used; private information, contact details, and profile images were not collected. Personal identifiers (names) were retained solely for gender inference and removed before aggregation and analysis. Profiles were filtered to include only those with at least one identifiable IT role and documented educational history, yielding 7,356 candidates. Gender inference follows a two-tier automated pipeline: the gender-guesser Python library (high-confidence classifications only) and the Genderize.io API queried with Serbian country context at a minimum confidence threshold of 95%, with agreement between both methods required. The remaining 1,033 profiles, unresolved by either method, were processed using large language models in both zero-shot and few-shot settings. Each model received the candidate's name, profile headline, and stated location as input. Four models were tested: Claude Sonnet 4.6, GPT-4o, Llama 3.1, and Mistral. A human annotator reviewed the outputs as part of a human-in-the-loop validation step, and inter-annotator agreement was quantified using Cohen's kappa. Comparing results across models and prompting strategies allowed us to gauge how sensitive the classification was to model choice and prompt design, considering both alignment with human judgments and consistency across models.

After filtering and two-method gender classification, 1,932 female and a larger number of male profiles were confirmed with high confidence. To ensure gender balance and avoid confounding effects of unequal group sizes, the male set was randomly downsampled to match the female count, yielding a final dataset of 3,864 profiles (1,932 per group). Male profiles were selected based on descending Genderize.io reliability scores to retain the highest-quality classifications. Job titles are normalized to a ten-level seniority ontology (Intern, Junior, Mid, Senior, Lead, Manager, Director, VP, CXO, Founder) using a rule-based classifier. For ambiguous or non-standard titles (e.g., Member of Technical Staff, Fellow, Distinguished Engineer), Claude Sonnet 4.6 is also applied; LLM classifications are compared against the rule-based baseline, with disagreements resolved through manual annotation. This hybrid approach combines the interpretability of rules with the contextual flexibility of LLMs. Parallel roles (e.g., Founder and CEO) are resolved by selecting the highest seniority attained.

Preliminary results indicate systematic differences across all four research dimensions. A career progression analysis shows that, among profiles with observable full career trajectories, men reach senior

or higher roles at nearly twice the rate of women (37% vs. 22%; $n=719$ vs. $n=425$; $\chi^2=106.60$, $p<0.001$). Mean time-to-senior is similar across genders (about 5 years; Mann-Whitney U, $p=0.13$). These trends are consistent with professional networking datasets (Hickey et al., 2022). Notably, women's profiles more often begin at senior or higher levels, suggesting that early-career positions may be selectively omitted from their public profiles. Women are more likely than men to hold postgraduate degrees among those with educational records (MSc or PhD: 65% vs. 52%; $\chi^2=49.24$, $p<0.001$). Yet, women show a higher overqualification index (0.30 vs. 0.23; Mann-Whitney U, $p<0.001$), which suggests a mismatch between qualifications and role level. Rule-based lexical analysis finds that men use achievement-oriented markers (such as led, delivered, scaled) at almost twice the density of women, measured as occurrences per 100 words. Since keyword counting captures surface form but not meaning, we complement this with a sentence-level semantic classifier: Claude Sonnet 4.6 applied in a few-shot prompting setup. Rather than counting specific words, this classifier assigns each sentence in a profile's self-description to one of three rhetorical categories: achievement-oriented (quantified accomplishments), responsibility-oriented (role descriptions), and collaborative (team-focused language). The semantic model is validated on a manually annotated subset of 150 profiles, with an inter-annotator agreement of Cohen's $\kappa = 0.89$ (Simon et al., 2023). The classifier shows that both groups use responsibility framing in about two-thirds of sentences (female: 67.9%; male: 68.3%). Women show slightly lower achievement orientation (15.4% vs. 17.5%) and higher collaborative framing (16.5% vs. 14.2%). These differences are in the hypothesized direction but do not reach statistical significance (Mann-Whitney U, $p = 0.076$ – 0.869). Thus, surface-level lexical patterns may overstate differences in semantic framing. Career break detection shows that women have employment gaps exceeding 6 months more often than men (42% vs. 35%; $\chi^2 = 19.68$, $p < 0.001$). The average break duration is similar for both groups (about 22 months). These findings suggest that gender differences in Serbian IT careers lie in frequency and visibility rather than pace. If hiring algorithms are trained on publicly available professional profile data, the observed disparities in language use and career visibility may be inadvertently encoded into such systems (Simon et al., 2023; Young et al., 2023). Explicit indicators of breaks are rare, reflecting a broader tendency to omit reasons for breaks from professional profiles (Kargabaeva et al., 2026), which our data confirm for the Serbian context.

Future work will apply survival analysis via the Cox proportional hazards model to estimate time-to-promotion while controlling for education and career interruptions, extend linguistic classifiers with fine-tuned transformer models (BERT, DeBERTa), and triangulate findings with self-reported survey data.

Keywords: gender disparities, computational social science, career trajectory analysis, LLM and NLP, women in IT

References

- Hickey, P. J., Erfani, A., & Cui, Q. (2022). Use of LinkedIn data and machine learning to analyze gender differences in construction career paths. *Journal of Management in Engineering*, 38(3). [https://doi.org/10.1061/\(ASCE\)ME.1943-5479.0001087](https://doi.org/10.1061/(ASCE)ME.1943-5479.0001087)
- Kargabaeva, K., Yang, Y., & Anicin, L. (2026). Predicting career trajectories after career breaks: A data-driven analysis using LinkedIn profiles. *Lecture Notes in Computer Science*. Springer. https://doi.org/10.1007/978-981-95-7138-3_25
- Penzenstadler, B., Boudaoud, K., Di Marco, A., & Caner-Yildirim, S. (Eds.). (2025). *Actions for gender balance in informatics across Europe*. Springer. <https://doi.org/10.1007/978-3-031-78432-3>
- Simon, V., Rabin, N., & Chalutz-Ben Gal, H. (2023). Utilizing data-driven methods to identify gender bias in LinkedIn profiles. *Information Processing & Management*. <https://doi.org/10.1016/j.ipm.2023.103423>
- Young, J., Wajcman, J., & Sprejer, L. (2023). Mind the gender gap: Inequalities in the emergent professions of artificial intelligence (AI) and data science. *New Technology, Work and Employment*, 38(3), 391–415. <https://doi.org/10.1111/ntwe.12278>

Building a Serbian Riddle Benchmark for Multicultural Evaluation of Large Language Models

Danka Jokić¹, Milica Andrić^{2*}, Nebojša Damjanović³, Toma Knežević^{4*}

¹ University of Belgrade, Serbia, danka.jokic@rcub.bg.ac.rs

² Faculty of Philology, University of Belgrade, Belgrade, Serbia, milica@mail.ru

³ Department of English Language, Literature and Culture, Faculty of Philology, University of Belgrade, Belgrade, Serbia, lingvistika.posao1@gmail.com

⁴ Department of general linguistics, Faculty of Philology, University of Belgrade, Belgrade, Serbia, 148toma@gmail.com

⁵ * Milica Andrić, Nebojša Damjanović, and Toma Knežević contributed equally to this work.

Large Language Models (LLMs) have achieved remarkable performance across a wide range of natural language processing tasks. However, their ability to understand figurative language, culturally grounded knowledge, and commonsense reasoning remains insufficiently explored, particularly in low-resource languages. While existing evaluation frameworks have significantly advanced the assessment of reasoning capabilities (Srivastava et al., 2023), recent work has emphasized the importance of culturally grounded evaluation beyond translated benchmarks (Boczka, 2025; Chang et al., 2025). Furthermore, studies based on traditional riddles have shown that even advanced multilingual models struggle with culturally specific reasoning tasks (Sayeedi et al., 2025).

To address this gap, Cohere Labs initiated a collaborative international effort aimed at constructing a multicultural benchmark based on riddles. The project brings together more than 40 teams from different countries and language communities, with a particular emphasis on underrepresented languages. Unlike many multilingual benchmarks that are translated from English, this initiative follows a participatory design paradigm in which native speakers act as creators rather than translators. Such an approach enables the inclusion of culturally specific knowledge, historical references, and linguistic phenomena that are often lost in translation.

Riddles provide a particularly challenging testbed for language models because solving them frequently requires commonsense reasoning, figurative interpretation, linguistic ambiguity, and cultural understanding. Unlike conventional question-answering tasks, riddles often rely on indirect clues and culturally shared knowledge, making them a useful instrument for evaluating reasoning beyond linguistic fluency.

In this paper, we present the Serbian contribution to the benchmark and describe the process of collecting, curating, and annotating Serbian riddles. Following the project guidelines, the collection consists of 100 riddles distributed across ten thematic categories. Eight categories were predefined by the organizing team to ensure comparability across languages, while two categories were selected locally in order to capture culturally specific aspects of Serbian language and society. Each riddle was independently assessed by a native-speaker evaluator who estimated its perceived difficulty for native speakers.

Since the benchmark is still being prepared for publication, we avoid evaluating models on the benchmark data itself. Instead, we conduct a preliminary pilot study using a separate held-out set of approximately 30 Serbian riddles that are not included in the benchmark. The goal of this pilot evaluation is not to provide statistically robust model rankings, but rather to obtain initial insights into how contemporary LLMs handle culturally grounded and figurative reasoning in Serbian. The pilot study aims to identify common strengths

and failure modes of contemporary LLMs in solving culturally grounded Serbian riddles. The evaluation considers answer correctness, reasoning quality, and the relationship between model behavior and human judgments of difficulty.

This initial study evaluated the performance of two models, Google Gemini 3.1 and OpenAI GPT-4o, using the same prompt and model settings. Gemini managed to fully or partially answer 26 out of 30 riddles, while OpenAI GPT-4o answered 19. Preliminary results suggest that Gemini solved a larger proportion of riddles than GPT-4o on the held-out set. Notably, the riddle 'Ima odraz kao ogledalo, ali se kreće kao voda.' (eng. *It has a reflection like a mirror, but it moves like water.*) presented a significant cognitive challenge, remaining unsolved by either model. Qualitative analysis of the models' explanations revealed frequent translation of Serbian riddles into English as an intermediate reasoning step before attempting a solution. In several cases, this appeared to contribute to incorrect answers when culturally grounded clues were interpreted outside their Serbian cultural context.

The main contribution of this work is the introduction of the Serbian component of a large-scale multicultural *benchmark* designed to evaluate reasoning beyond simple multilingual language understanding. The benchmark contributes a culturally grounded Serbian evaluation resource and forms part of a broader multilingual effort to assess reasoning capabilities across diverse linguistic and cultural contexts.

Keywords: Riddle Benchmark, Cultural Reasoning, Large Language Models

References:

- Boczka, K. (2025). *Hungarian riddles benchmark* [Data set]. Hugging Face. [Hungarian Riddles Benchmark](#)
- Chang, T. A., Arnett, C., Eldesokey, A., Sadallah, A., Kashar, A., Daud, A., Adelani, D. I., et al. (2025). *Global PIQA: Evaluating commonsense reasoning across 100+ languages and cultures*. arXiv preprint arXiv:2510.24081.
- Sayeedi, N. L., Sayeedi, M. F. A., Jahangir, K. B., Shatabda, S., & Preum, S. M. (2025). *Can LLMs Solve My Grandma's Riddle? Evaluating multilingual large language models on reasoning traditional Bangla tricky riddles*. arXiv preprint arXiv:2512.20324.
- Srivastava, A., Rastogi, A., Rao, A., Shueb, A. A. M., Abid, A., Fisch, A., Liang, P., et al. (2023). *Beyond the imitation game: Quantifying and extrapolating the capabilities of language models*. *Transactions on Machine Learning Research*, 2023.

Large Language Models and Emotion Annotation: From Open Vocabulary to Standardized Taxonomies

Ulfeta Marovac¹, Anida Vrcić Ammar², Ajla Gudžević³

¹ Department of Technical Sciences, State University of Novi Pazar, Serbia, umarovac@np.ac.rs

² Higher School of Vocational Studies for Teacher Education in Kikinda, Serbia, e-mail: avrcicamar@gmail.com

³ Department of Technical Sciences, State University of Novi Pazar, Serbia, ajlagudzevic1100@gmail.com

Automatic emotion recognition represents one of the key challenges in modern natural language processing. Most existing approaches rely on predefined emotion taxonomies, such as Ekman's model or the GoEmotions emotion framework (Demszky et al., 2020). The emergence of generative large language models (LLMs), such as Llama 3 (Dubey et al., 2024), has introduced the possibility of open-vocabulary emotion annotation, allowing models to identify and label emotions without being restricted to predefined categories. In this study, we investigate how the behavior of a generative model changes when it is allowed to freely name emotions compared to when it is constrained by a standardized emotion taxonomy.

Experiments were conducted on a corpus of 404 Serbian-language narratives previously annotated by two human annotators (Marovac & Vrcić Amar, 2025). Following the normalization process, the human annotations comprised a total of 146 distinct emotion labels. In the first experiment, the generative model Llama-3.3-70B-Versatile was not constrained by any predefined emotion taxonomy and was allowed to freely identify and name emotions expressed in the text. The analysis revealed that the model generated as many as 638 distinct emotion labels and emotion-related concepts, of which 92 overlapped with human annotations, while 546 labels were unique to the model. In addition to conventional emotions, the model frequently produced emotional states, relational concepts, and abstract notions, suggesting a substantially broader representation of the emotional space than that employed by human annotators.

In the second experiment, the GoEmotions taxonomy (Demszky et al., 2020), one of the most comprehensive resources for emotion analysis in natural language processing, was employed. GoEmotions comprises 27 emotion categories and one neutral category and was developed from a large corpus of user comments collected from Reddit. Previous experiments conducted on the same corpus employed a GoEmotions-based BERT model as a reference system for multi-label emotion classification. In the present study, human annotations were first normalized and mapped to the same 28 GoEmotions categories, enabling a direct comparison between human judgments, the GoEmotions model, and the generative LLM approach.

Based on the average scores of the two annotators, aggregated human labels were created and subsequently binarized using a threshold of 0.30 for the purpose of multi-label comparison with the model. The results show that the generative model achieved a Micro Precision of 0.4659, a Micro Recall of 0.4799, a Micro F1 score of 0.4728, and a Micro Jaccard index of 0.3096, while the average document-level Jaccard score reached 0.3256.

For comparison, a previously evaluated GoEmotions model, originally trained on the English GoEmotions dataset, was applied to machine-translated Serbian narratives without any additional adaptation and achieved a Micro F1 score of 0.06 and a Micro Jaccard score of 0.03. After fine-tuning on a corpus of Serbian narratives mapped to the GoEmotions categories, the performance of the same model improved

substantially, reaching a Micro F1 score of 0.50 and a Micro Jaccard score of 0.34 (Marovac & Vrcić Amar, 2026).

The findings suggest that one of the central challenges in emotion annotation may not be emotion recognition itself, but the establishment of a shared and standardized emotional vocabulary that enables reliable comparison across annotators and models. This is reflected in the fact that the unconstrained generative model produced a substantially broader and less consistent set of emotion labels than those used by human annotators. However, once the human annotations were mapped to the GoEmotions framework and the model was restricted to the same 28 emotion categories, a substantial level of agreement between human and model annotations was achieved (Micro F1 = 0.47; Micro Jaccard = 0.31). These results demonstrate that the generative model can approximate human emotion judgments within a standardized taxonomic framework, despite not being explicitly fine-tuned for the emotion annotation task. Overall, the results demonstrate that a general-purpose generative model, when guided by a standardized emotion taxonomy, can approximate the performance of a specialized GoEmotions-based classifier. Although no task-specific fine-tuning was performed, the model achieved a level of agreement with human annotations comparable to that of the fine-tuned reference system, highlighting the potential of large language models for emotion annotation tasks.

Keywords: emotion analysis, emotion annotation, generative large language models, multi-label classification, natural language processing

References:

- Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A. S., Nemade, G., & Ravi, S. (2020). *GoEmotions: A dataset of fine-grained emotions*. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 4040–4054). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.372>
- Dubey, A., Abhinandan, A., Agarwal, A., Arora, A., Athiwaratkun, B., Ayub, A., Balasingam, V., Bhatti, A., Brahma, B., Buraczynski, J., et al. (2024). *The Llama 3 herd of models* (arXiv No. a2407.21783). arXiv. <https://doi.org/10.48550/arXiv.2407.21783>
- Marovac, U., & Vrcić Amar, A. (2025, October 23–24). *When a machine interprets emotions: A comparative analysis of human and AI-based emotion annotation in text*. In *Social Aspects of the Application of Artificial Intelligence and Transhumanism: International Scientific Conference*. Belgrade, Serbia: Institute of Social Sciences & Research and Development Institute for Artificial Intelligence of Serbia.

A Hybrid Statistical-LLM Pipeline for Generating Synthetic Survey Respondents with Population-Aligned Distributions

Mladenović M.^{1[0000-0002-7973-452X]}, **Vujičić Stanković S.**^{2[0000-0002-7200-3724]}, **Kartelj A.**^{2[0000-0001-9839-6039]}

¹ University of Nis, Faculty of Education, Vranje, Serbia, ml.miljana@gmail.com

² University of Belgrade, Faculty of Mathematics, Belgrade, Serbia,
stasa.vujicic.stankovic@matf.bg.ac.rs, aleksandar.kartelj@gmail.com

Large language models (LLMs) have recently attracted attention as tools for simulating human survey respondents. By conditioning an LLM on a demographic profile (age, gender, education, etc.), researchers can generate synthetic responses to survey questionnaires at low cost, potentially replacing or augmenting expensive fieldwork. However, a fundamental limitation persists: LLM-generated response distributions often diverge significantly from real population distributions, especially for culturally specific or minority subgroups (Suh et al., 2025; Xie et al., 2026). Independent evaluations further confirm that demographic-only persona conditioning produces systematic positivity bias and reduced response variance (Kaiser et al., 2025). Most existing approaches address this either through post-hoc calibration of LLM outputs, fine-tuning on survey distribution data (Suh et al., 2025), or probabilistic prompt engineering without explicit statistical grounding.

In our recent work (Kartelj et al., 2026), we used LLM persona simulation on a Serbian survey of 1,000 respondents with 93 non-demographic questions and 7 demographic target variables. That study employed k-nearest neighbor (k-NN) augmented prompting, where the LLM received empirical response distributions from the most similar real respondents. We evaluate the method comparing the distributional fidelity of generated responses against the real population using KL divergence, Jensen-Shannon divergence, Total Variation distance, and maximum absolute per-question difference - metrics established as standard for this task (Xie et al., 2026). While effective, this approach had two limitations: the same k-NN neighborhood was used for all questions regardless of which demographics actually influenced the response, and the LLM generated answers question by question rather than as a coherent persona.

In this paper, we propose a hybrid two-phase pipeline that addresses both limitations. In the first phase, a statistical model computes conditional response distributions $P(\text{answer} \mid d)$ for each survey question individually, where d is a demographic variable and pairwise conditional response distributions $P(\text{answer} \mid d_1, d_2)$, where d_1 and d_2 are the two most influential demographic variables for that particular question, automatically selected via Jensen-Shannon divergence. For small demographic subgroups, weighted smoothing is applied to stabilize the estimates. These per-question distributions serve as structured behavioral priors. In the second phase, the complete set of priors for a given demographic profile is injected into the LLM prompt, and the model generates all survey responses at once as a coherent synthetic persona. Deploying a local open-weight model further enables token-level constrained decoding to enforce structural validity of generated responses, addressing the high invalid-output rates reported for open-weight models in comparable settings (Zhao et al., 2026).

No published work combines an explicit per-question statistical model as a pre-generation prior with whole-persona LLM generation. Existing approaches either calibrate LLM outputs post-hoc rather than guiding generation, fine-tune model weights rather than injecting distributions via prompts (Suh et al., 2025), or learn mixture weights without a separate statistical model per question. Our approach requires no model fine-tuning and is model agnostic, i.e. the same statistical priors can be used with any LLM. We evaluate the pipeline on the same Serbian survey dataset (1,000 respondents, 93 questions, 7 demographic targets) used in our prior work and using the same measures. We benchmark against three

baselines: (1) naïve LLM prompting with demographic profile only, (2) k-NN augmented prompting from our prior work, and (3) XGBoost as a purely statistical reference. Preliminary results suggest that injecting per-question statistical priors substantially reduces distributional divergence compared to prompt-only approaches, while preserving inter-question correlation structure, which purely statistical models cannot capture.

In summary, this work presents a hybrid two-phase pipeline that integrates per-question statistical priors with whole-persona LLM generation to produce synthetic survey respondents whose response distributions closely match those of the real population. Three promising avenues emerge for future investigation: (1) integrating grammar-guided and constrained decoding strategies to further reduce invalid output rates in local open-weight models, (2) systematic evaluation across additional survey instruments, languages, and model architectures to assess the generalizability of the approach, and (3) explicitly modeling and preserving inter-question correlation structure to ensure that dependencies among survey items in synthetic data reflect those observed in real respondents.

Keywords: LLM persona simulation, synthetic survey generation, distributional alignment, statistical priors, hybrid pipeline

References:

- Suh, J., Jahanparast, E., Moon, S., Kang, M., & Chang, S. (2025). Language model fine-tuning on scaled survey data for predicting distributions of public opinions. *Proceedings of the 63rd Annual Meeting of the ACL*. arXiv:2502.16761
- Xie, Y., Liang, L., Li, S., Lu, Y., Xiao, Z., Shi, M., ... & Xie, Y. (2026). Evaluating the statistical realism of LLM-generated social science data. *Proceedings of the National Academy of Sciences*, 123(19), e2538145123.
- Kartelj, A., Mladenovic, M., & Vujicic Stankovic, S. (2026). LLM-based persona simulation of survey respondents. *Proceedings of the 5th Serbian International Conference on Applied Artificial Intelligence (SICAAI)*.
- Kaiser, C., Kaiser, J., Manewitsch, V., Rau, L., & Schallner, R. (2025). Simulating human opinions with large language models: Opportunities and challenges for personalized survey data modeling. *UMAP Adjunct '25*. <https://doi.org/10.1145/3708319.3733685>
- Zhao, J., Yuan, C., Luo, W., Xie, H., Zhang, G., Quan, S. J., Yuan, Z., Wang, P., & Zhang, D. (2026). Large language models as virtual survey respondents: Evaluating sociodemographic response generation. *arXiv preprint*, arXiv:2509.06337.

LLM-Assisted Expansion of Serbian WordNet: Evaluating Prompting Strategies for Synset Candidate Generation

Saša Z. Petalinkar¹, Ranka Stanković², Cvetana Krstev³

¹ University of Belgrade, Faculty of Mining and Geology, Serbia, sasa.petalinkar@rgf.bg.ac.rs, ORCID: <https://orcid.org/0009-0007-9664-3594>

² University of Belgrade, Faculty of Mining and Geology, Serbia, ranka@rgf.rs, ORCID: <https://orcid.org/0000-0001-5123-6273>

³ Language Resources and Technologies Society - JeRTeh, Serbia, Cvetana@jerteh.rs, ORCID: <https://orcid.org/0000-0003-3328-9392>

The expansion of Serbian WordNet (SrpWN; <https://wn.jerteh.rs/>) from Princeton WordNet (PWN) cannot be reduced to translation. PWN provides a broad English synset inventory (Fellbaum, 1998), and SrpWN follows the EuroWordNet/BalkaNet model for Serbian (Krstev et al., 2004). However, differences in coverage and lexicalization mean that PWN concepts cannot always be transferred directly into Serbian. Previous studies have examined automatic synset construction and the use of LLMs to construct bilingual wordnets (Bergh et al., 2025; Lam et al., 2014). This experiment evaluates whether an LLM can generate Serbian synset candidates suitable for subsequent lexicographic review. Generated candidates are not intended for direct insertion into SrpWN.

Each source synset was represented by its identifier, English gloss and literals (lexical units); the required output was a candidate Serbian gloss and a list of candidate Serbian literals. All three workflows used the same OpenAI gpt-4.1-mini model. Using a single model kept model choice constant so that the comparison focused on prompting strategy; comparison between model families was outside the scope of the study. All prompts included Serbian-language instructions to return standard Latin-script dictionary lemmas (infinitives for verbs and nominative forms for nouns) and established lexical equivalents, while avoiding invented calques and nonstandard forms. The prompts and implementation are available in the public repository https://github.com/sasa5linkar/wordnet_autotranslate-.

Three prompting workflows were evaluated. The baseline directly translated the English gloss and literals. The multi-step workflow first analysed the source synset, then translated the gloss and literals, expanded the Serbian literal set, filtered candidates against the intended meaning, and reformulated the final gloss. The conceptual workflow constructed a structured representation of the source concept from the gloss, literals, examples, part of speech, domain information, and semantic relations. It then produced an expanded English definition, used it to formulate a Serbian definition, generated candidate literals, selected lexicalized Serbian expressions matching the intended sense, formulated the final Serbian gloss, and checked the result for circularity, semantic shift, and inappropriate broadening or narrowing. Accordingly, the workflows differed in how much semantic context informed literal selection.

The full generation series contained 1,734 candidate rows: 578 English synsets processed through the three workflows, with 4,721 Serbian candidate literal occurrences. The manually evaluated subset comprised 30 synsets and one output per workflow for each synset, yielding 90 reviewed rows. The evaluation was conducted by one expert lexicographer. The resulting scores are descriptive, and inter-annotator agreement was not calculated. Each criterion was scored from 1 (criterion not satisfied) to 5 (criterion fully satisfied); scores of 2, 3, and 4 indicated partial non-satisfaction, a neutral or partly satisfactory result, and substantial satisfaction, respectively. Glosses were assessed for naturalness, semantic adequacy, sense differentiation, part-of-speech-appropriate structure, and non-circularity.

Literals were assessed for semantic adequacy, with separate flags and corrections for redundant and missing lexical units.

The conceptual workflow obtained the highest mean across the five gloss criteria (4.71), followed by the multi-step (4.56) and baseline (4.36) workflows. The baseline and conceptual workflows both received a mean literal-adequacy score of 4.10, compared with 3.29 for the multi-step workflow, for which redundant literals were recorded in 26 of 29 evaluable outputs. Missing literals were marked in 53 of 90 rows: 18/30 baseline, 17/30 multi-step and 18/30 conceptual outputs. This flag did not mean that no literal was generated; it meant that the expert identified one or more appropriate Serbian lexical units absent from the candidate set. The expert's comments identified several causes related to Serbian lexicalization that were neither fully represented in the English source nor explicitly addressed by the prompts. All 15 evaluated verb outputs lacked at least one appropriate literal, frequently a perfective or imperfective pair; the comments also noted Ekavian and Ijekavian variants, additional synonyms, and appropriate candidates removed during filtering. In the reviewed subset, the generated candidate sets were often lexically incomplete despite generally high semantic-adequacy scores. The model output should therefore be treated as candidate material in an expert-supervised workflow rather than as ready-to-import lexical entries. Expert validation remains necessary before integration according to established SrpWN/PWN alignment and resource-based augmentation procedures (Stanković et al., 2018).

Keywords: Serbian WordNet, large language models, synset candidate generation, lexical-semantic resources, lexicographic evaluation

Acknowledgment: This research was conducted with the support of the Science Fund of the Republic of Serbia, project number: 7276 TESLA: "Text Embeddings - Serbian Language Applications" and ELEXAI project No 731015: "European Lexicographic Infrastructure for Artificial Intelligence" funded by the European Union's Horizon 2020 research and innovation programme.

References:

- Bergh, J., Waitelonis, J., & Siegel, M. (2025). Leveraging LLMs for constructing WordNets automatically as bilingual resources. In *Proceedings of the 13th Global Wordnet Conference* (pp. 203-211). Global Wordnet Association. <https://doi.org/10.18653/v1/2025.gwc-1.24>
- Fellbaum, C. (Ed.). (1998). *WordNet: An electronic lexical database*. MIT Press.
- Krstev, C., Vitas, D., Pavlović-Lažetić, G., & Obradović, I. (2004). Using textual and lexical resources in developing Serbian WordNet. *Romanian Journal of Information Science and Technology*, 7(1-2), 147-161.
- Lam, K. N., Al Tarouti, F., & Kalita, J. (2014). Automatically constructing Wordnet synsets. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 106-111). Association for Computational Linguistics. <https://doi.org/10.3115/v1/P14-2018>
- Stanković, R., Mladenović, M., Obradović, I., Vitas, M., & Krstev, C. (2018). Resource-based WordNet augmentation and enrichment. In *Proceedings of the Third International Conference on Computational Linguistics in Bulgaria (CLIB 2018)* (pp. 104-114). <https://aclanthology.org/2018.clib-1.14/>

Development of an Artificial Intelligence Chatbot Agent for Financial Investment

Sonja D. Radenković¹

¹ Belgrade Banking Academy – Faculty of Banking, Insurance and Finance, Union University, Serbia, sonja.radenkovic@bba.edu.rs

Artificial intelligence (AI) in finance is increasingly being used in many companies [1], while fundamental financial analysis still requires the collection, calculation, and interpretation of multiple company-level indicators [2,3]. This paper presents a step forward to both automated fundamental financial analysis and investment decision support. It describes the design and implementation of AI-based chatbot agent created to automate fundamental financial analysis in the process of identifying the liquidity and solvency of the companies. The main goal was to investigate whether integrating financial-data retrieval, indicator calculation, retrieval-augmented generation (RAG), and a large language model (LLM) can make company analysis faster and more understandable. Unlike traditional fundamental analysis which requires significant time and resources, especially in the collection, filtering and interpretation of data, this approach enables faster recognition of important information and patterns that are important to investors. The paper focuses on evaluating the advantages of the automated approach, as well as identifying opportunities for further development and practical application in a real investment environment.

The architecture of an intelligent chatbot agent is shown in the Figure 1 and includes several interconnected components. Each component has a special role, but the real quality of the system is achieved through their integration. The first component is the web-based user interface through which the user communicates with the chatbot. The second component is the message processing layer that receives user input and prepares it for further processing. The third component is a large language model (LLM), such as OpenAI, that is in charge for natural language queries by considering the context of the conversation and generating response in a user-friendly manner. The fourth component is the retrieval-augmented generation (RAG) layer that retrieves relevant contextual information from the knowledge base and provides it to the large language model before response generation [5]. The fifth component is a vector database that stores documents or document segments stored as the numerical representations that enable semantic search. The sixth component is a knowledge base that may include frequently asked questions, instructions, regulations, internal procedures and other relevant sources of information. The quality of the knowledge base directly affects the quality of the chatbot agent's response. The seventh component is the agent orchestrator that manages the overall interaction flow and decides which system components should be activated for a specific user request. Depending on the context, the orchestrator may decide whether to use a RAG-based search, call an application programming interface (API), request additional information from the user, generate a response, or escalate the request when necessary. In this way, the orchestrator functions as the main coordination layer connecting user requests, the LLM, external model as well as the knowledge base.

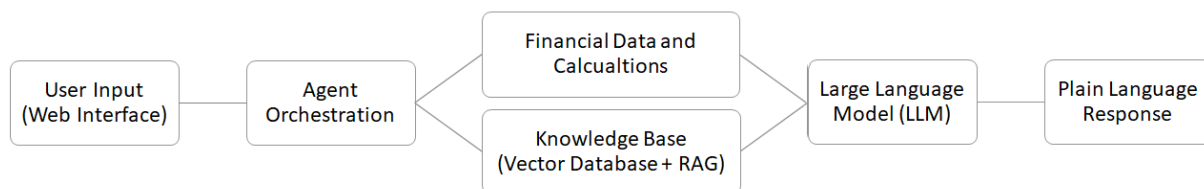


Figure 1. Components of the AI-chatbot agent

The user session begins by initiating the interaction by prompting the user to enter the ticker of the company to be analyzed (e.g. AAPL for Apple): Once the ticker name is obtained, the system retrieves the most recent financial indicators for the selected company, such as free cash flow (indicates the ability of the company to invest and pay dividends), total revenues (indicates the extent of the company's dominance in the market), net profit (indicates how operationally efficient the company is), total debt (represents the sum of all short-term and long-term liabilities), cash and cash equivalents (shows the short-term stability of the company). In addition to indicators extracted from financial reports, as a result of formulas and calculations, two additional financial indicators are obtained: discounted cash flows with a projection for 5 years (shows whether the company is undervalued or overvalued by the market) and net debt (shows the company's real exposure to liabilities). When the financial indicators are retrieved and calculated, the LLM then converts the numerical output into a plain-language interpretation, while the RAG component supplies relevant contextual information. Generative AI can support knowledge work by assisting with interpretation and communication [4]; however, the generated output is treated as decision support rather than autonomous investment advice. Therefore, the artificial intelligence (AI) component analyzes and interprets the results as a financial analyst, presenting them in a language that is understandable even to users without complex financial knowledge. By combining traditional financial indicators with advanced artificial intelligence algorithms, the proposed approach can provide users with a clearer understanding of company valuation, as well as facilitate the interpretation of complex financial information.

The presented user session completed end-to-end workflow from ticker input to the financial indicators and AI-generated interpretation. The main practical advantage is the integration of data retrieval, calculation, contextual retrieval, and explanation in a single interactive workflow, which reduces manual processing steps and improves the accessibility of fundamental-analysis results. These results demonstrate the technical feasibility of the proposed approach; they do not establish superior investment performance or the general reliability of AI-generated recommendations.

The study makes the following contributions: (1) an implemented modular architecture that integrates financial-data processing with an AI conversational interface; (2) an automated workflow that combines conventional liquidity, solvency, performance, indebtedness, and valuation indicators with AI-generated interpretation; and (3) a prototype demonstration showing how complex financial results can be communicated in plain language to users without advanced financial knowledge. The further direction of development will be aimed at including a larger volume of indicators for fundamental analysis, as well as combining it with a tool that specializes in technical analysis in order to complete a complete financial analysis.

Keywords: Artificial Intelligence, Financial Investment, Chatbot, LLM.

References:

- Bahoo, S., Cucculelli, M., Goga, X. et al. Artificial intelligence in Finance: a comprehensive review through bibliometric and content analysis. *SN Bus Econ* 4, 23 (2024). <https://doi.org/10.1007/s43546-023-00618-x>
- Brealey, R. A., Myers, S. C., & Allen, F. (2019). *Principles of Corporate Finance* (13th ed.). New York, McGraw-Hill Education.
- Brigham, E. F., & Houston, J. F. (2019). *Fundamentals of financial management* (15th ed.). Boston, MA: Cengage Learning.
- Brynjolfsson, E., Li, D., & Raymond, L. R. (2023). *Generative AI at Work*. National Bureau of Economic Research Working Paper No. 31161.
- 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.

Building a Knowledge Graph from the Historical Serbian Encyclopedia *Sveznanje*

Ranka Stanković¹, Biljana Rujević¹, Olivera Kitanović¹

¹ University of Belgrade, Faculty of Mining and Geology, Serbia,
{ranka.stankovic|biljana.rujevic|olivera.kitanovic}@rgf.bg.ac.rs

The digitization of historical encyclopedias is essential for preserving and accessing structured cultural knowledge, yet many projects remain limited to OCR-based text conversion. Andersson et al. (2026) presented a pipeline for restoring encyclopedia structure through headword extraction, entry identification, entity classification, cross-edition matching, and Wikidata linking. Applied to multiple editions of the Swedish encyclopedia “Nordisk familjebok”, the approach achieved high accuracy in headword extraction and classification, demonstrating the feasibility of automated processing of historical encyclopedic knowledge and its integration into linked data environments. Wu et al. (2020) presented a framework for constructing knowledge graphs from multiple online encyclopedias, addressing the limitations of sparse non-English Wikipedia data. The framework combines knowledge extraction and knowledge linking modules, enabling both periodic extraction of encyclopedia content and real-time processing of updated entries. Entity and property matching were performed using heuristic and semi-supervised methods, achieving better performance than existing approaches and enabling the creation of large-scale multilingual knowledge graphs.

The goal of this paper is to create a semantically enriched knowledge graph from the historical Serbian encyclopedia *Sveznanje* using TEI encoding, NER, relation extraction, and Wikidata linking. This paper presents the creation of a knowledge graph based on the Serbian encyclopedia *Sveznanje*, a Serbian general encyclopedia first published in 1937 in Belgrade (Petrović, 1937). It covered a wide range of topics from science, history, geography, culture, and technology, aiming to provide comprehensive knowledge to a broad readership. The text is printed in two columns with paginated columns. The encyclopedia was published in two volumes comprising 1500 pages (2720 columns). For this task illustrations were not used. The encyclopedia represents an important cultural and educational resource in Serbian publishing history and is today valuable for digital humanities and knowledge graph research. A comprehensive encyclopedic resource is partially available in digital form through Wikipedia¹ and additional digitized volumes were provided by the University Library “Svetozar Marković”.

The encyclopedia content available on the web was converted into a structured TEI-XML format, while the newly digitized material underwent OCR processing followed by extensive manual correction, structural enrichment, and normalization. The processing pipeline included the identification and annotation of encyclopedia entries, resolution of abbreviations, metadata structuring, and semantic enrichment, enabling the transformation of raw digitized text into a machine-readable and interoperable resource suitable for linked data and knowledge graph construction. The workflow included encyclopedia article structural annotation, partially automatic abbreviation resolution, named entity recognition (NER), and linking of entities to Wikidata, enabling integration with the Linked Open Data ecosystem. For semantic annotation, a very important step is NER that will rely on models developed within project TESLA (Text Embeddings – Serbian Language Applications) that recognize 7 entity types: persons, organizations, locations, professions, events, demonyms, works of art (Ikonić Nešić et al. 2026).

¹ <https://sr.wikisource.org/wiki/Свезнање>

A crucial part is relation extraction that will follow the approach established in (Stanković et al. 2026), where relation extraction was applied on the TeslaNER+ dataset, a large corpus of over 150,000 sentences annotated with named entities and linked to Wikidata. Nineteen semantic relations across multiple domains, including literary texts, were selected and extracted using both rule-based finite-state transducer (FST) methods and LLM-based prompting. The set of relations is predefined, and possible subject–object combinations are specified according to entity types. For this experiment, the following relations were selected: between persons (colleagueOf, managerOf, parentOf, relativeOf, siblingOf, marriedTo, spouseOf); between persons and organizations (educatedAt, managerOfOrg, affiliatedWith); between persons and locations (residesIn, bornAt, diedAt); between persons and works (authorOf); between organizations (partOfOrg); between organizations and locations (basedIn); between locations (partOfLoc); and relations connecting events with persons (eventPers) and locations (eventLoc). Experimental results on 23,000 sentences show that the FST approach achieves higher precision (0.77) but lower recall (0.39), while the LLM approach provides very high recall (0.97) at the cost of lower precision (0.65), greater computational expense, and slower processing.

The proposed knowledge graph combines textual, structural, and semantic information extracted from encyclopedia articles. In addition to entity linking, articles were automatically classified according to their domain and semantic type, including categories such as biographies, historical events, locations, organizations, concepts, and cultural artifacts. The resulting graph enables advanced semantic search, exploration of relationships between entities, and interoperability with external knowledge bases.

The paper describes the digitization and enrichment pipeline, TEI modeling decisions, semantic annotation procedures, and graph construction methodology. Special attention is given to the challenges of processing historical encyclopedic content in Serbian, including orthographic variation, abbreviation disambiguation, and entity normalization. The resulting resource represents an important contribution to Serbian digital humanities and language technologies, providing a reusable infrastructure for semantic search, corpus analysis, and knowledge-driven applications.

Keywords: NLP, Knowledge Graphs, NER, NEL, Serbian

Acknowledgments: This research was supported by the Science Fund of the Republic of Serbia under Project No. 7276, “Text Embeddings – Serbian Language Applications (TESLA)”.

References:]

- Andersson, A., Jonasson, S., Wastring, F., & Nugues, P. (2026). [ATLAS: Article Tracking, Linking, and Analysis of Swedish Encyclopedias](#). In Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026) (pp. 3566–3576). (ELRA).
- Wu, T., Wang, H., Li, C., Qi, G., Niu, X., Wang, M., ... & Shi, C. (2020). [Knowledge graph construction from multiple online encyclopedias](#). World Wide Web, 23(5), 2671-2698.
- Petrović, P.M., Свезнање : општи енциклопедиски лексикон (1937), уредник П. М. Петровић, Београд : Народно дело.
- Ikonić Nešić, M., Petalinkar, S., Kitanović, O., Stanković, R., Utvić, M. (2026). CNN-based named entity linking: Serbian use case. Poznan Studies in Contemporary Linguistics. DOI:10.1515/psicl-2025-0064.
- Stanković, R., Krstev, C., Ikonić Nešić, M. and Utvić, M. (2026) Relation Extraction from Text of Various Domains: A Serbian Case Study, The Seventh International Conference Computational Linguistics in Bulgaria (CLIB 2026) (in process).

Cross-lingual subword mapping from parallel corpora

Mihailo Škorić

University of Belgrade, Faculty of Mining and Geology, Serbia, mihailo.skoric@rgf.bg.ac.rs

Subword tokenization is the foundation of text representation for large language models (LLMs). This enables them to process diverse inputs with efficiency and scalability. However, most widely adopted tokenizers are trained on English or broad multilingual corpora, which often fail to capture the nuances of under-represented languages such as Serbian. Training LLMs entirely from scratch for these languages is not a practical solution either, as it would require trillions of tokens to achieve competitive performance. Instead, we rely on transfer learning, where knowledge acquired in high-resource languages is adapted to low-resource ones. A critical component of this process is embedding weight initialization: by starting with informed rather than random weights, models can converge faster and generalize more effectively. This is where cross-lingual subword mappings become important. By aligning subwords with those of resource-rich languages, we can bootstrap embeddings without massive monolingual corpora, ensuring that models benefit from shared knowledge while still respecting the unique lexical structure of the student model's language.

One example of cross-lingual subword mapping is *Subword Mapping and Anchoring* (SMALA), which introduces unsupervised alignment techniques to establish links between subwords of different languages (Vernikos, 2021). SMALA explicitly learns correspondences between subwords via processing learned embeddings, thereby improving zero-shot transfer in tasks such as machine translation. Empirical results demonstrate that SMALA significantly boosts BLEU scores and cross-lingual accuracy compared to baseline multilingual tokenizers, underscoring the importance of structured subword alignment for morphologically complex languages. However, the reliance on learned embeddings makes the approach somewhat convoluted: it works well for mappings within an existing vocabulary, but the method does not naturally extend to new, unknown tokens. This limitation reduces its usability for scenarios such as tokenizer vocabulary extension or the initialization of embeddings for newly introduced subwords, where flexible token-to-token mapping strategies are required.

Another influential resource in the space of cross-lingual mapping is *Word2Word*, an open-source project that provides cross-lingual word mappings covering over 100 languages, and the means to produce new them (Choe et al., 2020). Word pair candidates are extracted from large parallel corpora (Lison et al., 2018), based on cooccurrence and aligned into simple and lightweight word-to-word mappings that can be directly used for tasks such as cross-lingual embedding alignment, and low-resource machine translation. Its wide coverage and open availability have made it a common baseline in comparative studies. However, because the method operates strictly at the word level, it is not fully compatible with the LLM paradigm, fundamentally built on subword tokenization.

Building on *Word2Word*, we adapt its cross-lingual word-level mapping into a cross-lingual subword mapping framework. Instead of aligning surface words directly, we modify the tokenization procedure by splitting parallel text into subwords using *HuggingFace* tokenizers. This enables two complementary modes of operation: (1) both languages can be tokenized with the same tokenizer, producing a direct cross-lingual mapping of tokens within a single shared vocabulary; or (2) each language can be tokenized with a different tokenizer, yielding a vocabulary-to-vocabulary mapping across tokenizers. The first mode is useful for analyzing how a unified tokenizer handles morphologically rich languages, while the second mode is beneficial for practical applications such as initializing LLM embedding weights for a new tokenizer or tokenizer extension. By shifting the alignment from words to subwords, we hope to provide a

principled way to bootstrap cross-lingual transfer learning at the tokenizer level as a pathway for integrating new languages into LLMs without the need to train from scratch.

Using this approach, we construct two initial mappings: Serbian–Russian and Serbian–Chinese (zh). For parallel corpora, we rely on the *OpenSubtitles* dataset (same as *Word2Word*). Each parallel segment is tokenized using Hugging Face tokenizers, specifically the Qwen3.5 tokenizer (Qwen Team, 2026) for Russian and Chinese text, and a newly constructed Serbian tokenizer for Serbian text. The Qwen3.5 tokenizer was chosen for its strong support for Chinese and Russian, with the immediate goal of procuring reliable cross-lingual subword mappings and the secondary goal of extending the Qwen3.5 tokenizer with new Serbian tokens in the future.

These initial mappings serve as a constructive starting point for future research. Rather than presenting evaluation results at this stage, we focus on delivering the mappings themselves as a resource that can be built upon. They provide a basis for extending multilingual tokenizers, initializing embedding weights for new vocabularies, and exploring how subword alignment can enhance transfer learning in large language models, especially for under-represented languages such as Serbian.

Keywords: cross-lingual, subword mapping, tokenizers, transfer learning

References:

- Choe, Y. J., Park, K., & Kim, D. (2020, May). word2word: A collection of bilingual lexicons for 3,564 language pairs. In *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 3036-3045).
- Lison, P., Tiedemann, J., & Kouylekov, M. (2018, May). OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.
- Team, Q. (2026). Qwen3. 5-omni technical report. *arXiv preprint arXiv:2604.15804*.
- Vernikos, G., & Popescu-Belis, A. (2021, November). Subword mapping and anchoring across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2021* (pp. 2633-2647).

Exploring Narrative Frames via Emotional and Moral Signals in Text: A Knowledge Graph Approach to Analyze Social Discourse

dr Milena Šošić¹, prof. dr Ranka Stanković²

¹ Language Resources and Technologies Society (JeRTeh), milena.sosic@jerteh.rs

² Applied Mathematics and Informatics, Faculty of Mining and Geology, University of Belgrade, Serbia, ranka.stankovic@rgf.bg.ac.rs

This study investigates the role of emotional and moral signals in shaping narrative frames within social discourse, with a particular focus on propagandistic and misleading content. In the context of the rapid spread of misinformation across digital platforms, there is an increasing need for analytical frameworks that move beyond surface-level textual representations and provide deeper interpretability of manipulative communication strategies. Traditional Natural Language Processing (NLP) approaches to misinformation detection have primarily relied on lexical, syntactic, or statistical features, often overlooking the socio-cognitive mechanisms that make such narratives persuasive. This study argues that narrative operates through systematic emotional appeal and moral framing that influence audience perception and reinforce ideological positions. The proposed framework integrates two theoretical perspectives: Plutchik's model of basic emotions (Plutchik, 2001) and Moral Foundations Theory (MFT) of moral norms identified as shared across the cultures (Graham et al., 2009) into the knowledge graph structure to evaluate narrative frames. Emotional categories such as *anger*, *fear*, *sadness*, *joy*, *disgust*, *surprise*, *trust*, and *anticipation* are modeled as affective signals shaping the tone of discourse, while moral dimensions including *care/harm*, *fairness/cheating*, *loyalty/betrayal*, *authority/subversion*, *sanctity/degradation*, and *liberty/oppression*, capture normative judgments embedded within the text. By combining emotional and moral analysis, the framework identifies narrative structures frequently associated with propagandistic communication, including “*us versus them*” polarization, *victimization*, *heroism*, and *conspiracy framing* (Freeman, 2023).

The main contribution of the study is the development of a multidimensional annotation schema designed for the structured analysis of social discourse. Designed schema captures multiple interacting layers of communication, that includes truthfulness, emotional affect, moral foundations, propaganda techniques, narrative framing, audience targeting, and event domains. Proposed rich hierarchical representation enables a comprehensive modeling of conversation dynamics by linking rhetorical strategies with their emotional and moral underpinnings. In addition to narrative analysis, the schema provides a foundation for related tasks such as automated truthfulness assessment, misinformation detection, and early recognition of hate speech communication patterns. To operationalize this approach, the study introduces a multi-stage methodology consisting of data collection, annotation, knowledge graph construction, and graph evaluation on multiple classification tasks. A corpus containing Serbian conversational discourse related to socially significant events in Serbia was constructed from heterogeneous sources including news articles, Reddit discussions, and synthetically generated posts associated with politically polarized and propagandistic discourse. The dataset preserves the conversational structure by linking news articles and social media posts with their associated comments and discussion threads. Such hierarchical organization enables analysis of how emotional and moral signals propagate across interactions, and allows the study of narrative reinforcement, polarization dynamics, and the influence of conversational context on downstream classification tasks (Šošić, 2025). Semantic annotations were produced using the Large Language Model (LLM) GPT-5.1, guided by a predefined annotation schema, while human annotators validated and refined the emotional, moral, and rhetorical signals produced by the model. Human annotators were trained to identify mappings across multiple interconnected ontologies defined within the annotation schema, including emotion categories derived from the Plutchik model, moral foundations based on MFT, and novel ontological structures for propaganda, narratives, and misinformation introduced in this work.

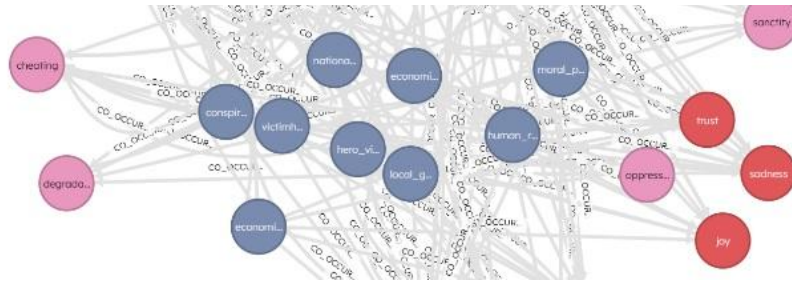


Figure 1: Excerpt from the SocialDiscourse.KG graph representing narrative frames (blue), moral (pink) and emotional (red) nodes with the cooccurrence relationships.

A central component of the framework is the construction of a knowledge graph, named *SocialDiscourse.KG*, with the nodes corresponding to discourse elements, and edges encoding co-occurrence patterns and semantic associations (Ji et al., 2021). Initial descriptive graph analysis indicates that *fear* and *anger* frequently co-occur with *authority* and *loyalty* moral values in *conspiracy* narratives, while *anger* combined with *fairness* predominantly appears in *accusatory* and *polarizing* conversations. Graph-derived features, such as node centrality, community detection, and recurrent subgraph frequencies, are integrated as additional context into machine learning classifiers for narrative frame, propaganda and misinformation detection. Compared to lexical baselines, the proposed approach improved classification accuracy on narrative frame tasks for +8.2% and F1-score for +7.3%. The same feature space also showed strong discriminatory potential for identifying propagandistic and misleading narrative patterns. Future work will focus on expanding the corpus with additional samples such as parliamentary debates, improving annotation robustness with Human-in-the-Loop (HITL) agentic workflows, and extending the framework toward multimodal and multilingual settings. Through the integration of multiple knowledge domains, knowledge graphs, and advanced machine learning methods, the study provides a comprehensive and interpretable framework for understanding different types of narratives in digital communication. Initial results demonstrate that emotional and moral features improve classification performance across narrative analysis tasks and provide valuable insights how the socio-cognitive mechanisms of underlying propaganda, misinformation, and competitive discourse operate.

Keywords: social discourse, narrative framing, emotional affect, moral foundation, knowledge graph

References:

- Freeman, L. et al. (2023). Conflicts, villains, resolutions: Towards models of narrative media framing. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 8712--8732).
- Jesse G., Haidt, J. and Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, 96(5), 1029.
- Ji, S., Pan, S., Cambria, E., Marttinen, P. and Yu, P.S., (2021). A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE transactions on neural networks and learning systems*, 33(2), 494-514.
- Plutchik, R. (2001). The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist*, 89(4), 344-350.
- Šošić, M. (2025). Modeling moral and emotional language aspects in conversational texts classification. *PhD Thesis*. University of Belgrade, Faculty of Mathematics.

Do Domain-Specific Financial Models Still Matter in the Era of Large Language Models?

Irena Vodenska¹, Ulfeta Marovac², Admir Hamzagić³, Dalila Pramenković⁴

¹ Administrative Sciences Department, Metropolitan College, Boston University, USA,
vodenska@bu.edu

² Department of Technical Sciences, State University of Novi Pazar, Serbia, umarovac@np.ac.rs

³ Department of Economic Sciences, State University of Novi Pazar, Serbia, ahamzagic@np.ac.rs

⁴ Department of Technical Sciences, State University of Novi Pazar, Serbia,
pramenkovicdalila@gmail.com

Recent advances in natural language processing have significantly improved the automatic analysis of financial texts. Financial news sentiment has long been considered an important source of information for understanding stock market behavior and predicting market trends. Previous studies have shown that sentiment extracted from financial news can contribute to the prediction of stock returns and market dynamics (Li et al., 2014). Furthermore, the development of deep learning and Transformer-based architectures has substantially improved the performance of financial sentiment analysis systems (Mishev et al., 2020). More recently, domain-specific models such as FinBERT have demonstrated strong performance in extracting financial information from textual data and predicting stock price movements (Gu et al., 2024). At the same time, the emergence of large language models (LLMs) has opened new possibilities for context-aware interpretation of financial news and market events (Dong et al., 2025). Recent studies have demonstrated that advanced general-purpose large language models can outperform both traditional dictionary-based methods and domain-specific financial models such as FinBERT in stock return prediction and sentiment-driven trading strategies, highlighting the benefits of richer contextual language understanding (Kirtac & Germano, 2024). In parallel, large language models, particularly domain-specific models such as FinBERT, have also demonstrated substantial improvements over conventional machine learning approaches in financial sentiment analysis and market forecasting, emphasizing the importance of contextual semantic representations in financial applications (Qi, 2025). However, it remains unclear whether general-purpose LLMs can fully replace financial-domain models in tasks involving both human interpretation of sentiment and market-oriented financial analysis.

This study investigates this question using a collection of 299 financial news articles related to companies included in the Dow Jones Industrial Average (DJIA), extracted from a broader dataset covering companies from the S&P 500 index. The study compares the performance of a large language model (LLM), a smaller language model (SLM), and the financial-domain model FinBERT in a company-oriented sentiment analysis task. In addition to human evaluation, the relationship between sentiment generated by the models and stock price movements before and after news publication is examined. The analysis is further motivated by the Efficient Market Hypothesis (EMH), which assumes that publicly available information is rapidly incorporated into stock prices, thereby limiting the ability of investors to predict future market movements and achieve abnormal returns based solely on newly released information (Fama, 1970). For human-centered evaluation, a subset of the dataset was independently annotated by three human evaluators using positive, neutral, and negative sentiment labels. A consensus annotation was used as the reference for model evaluation. The average inter-annotator agreement was 71.1%, corresponding to an average Cohen's κ of 0.57. The results show that the LLM achieved human-level agreement with the consensus annotations (73.3%, $\kappa = 0.60$), outperforming both the SLM (63.3%, $\kappa = 0.45$) and FinBERT (36.7%, $\kappa = 0.09$).

Analysis of inter-model agreement revealed an almost perfect alignment between the LLM and SLM approaches, achieving 87.6% agreement with a Cohen's κ of 0.808. In contrast, agreement between FinBERT and the other models was substantially lower, reaching 45.8% ($\kappa = 0.194$) for LLM versus FinBERT and 46.8% ($\kappa = 0.213$) for SLM versus FinBERT. These findings suggest that the LLM and SLM adopt a highly similar company-oriented interpretation of sentiment, whereas FinBERT captures a broader notion of financial sentiment that differs considerably from the interpretations produced by both language models. The relationship between sentiment and stock price movements was further examined over a period spanning five days before news publication (m1–m5) and five days after publication (p1–p5). The highest agreement with market trends was observed in the pre-publication period, reaching 46.2% for the LLM, 45.5% for the SLM, and 38.9% for FinBERT at the one-day horizon (m1). Agreement generally declined as the temporal distance from publication increased. In the post-publication period, FinBERT achieved the strongest alignment with market movements, reaching 40.6% agreement at both p1 and p2, whereas the LLM and SLM obtained lower values ranging between 30% and 37%. While the LLM and SLM more closely reproduced human sentiment judgments, FinBERT demonstrated comparatively stronger alignment with market behavior after news publication. This pattern is broadly consistent with the Efficient Market Hypothesis (Fama, 1970), which suggests that publicly available information is rapidly incorporated into stock prices.

Overall, the findings indicate that human-perceived sentiment and market-relevant sentiment do not necessarily represent the same phenomenon. Large language models more accurately reproduce the way humans interpret the impact of news on a specific company, whereas FinBERT appears to capture signals that are more closely associated with subsequent market behavior. These results suggest that domain-specific financial models continue to provide complementary value for financial analytics despite the rapid progress of generative AI systems. Although this study is limited to a relatively small dataset and a single company-oriented sentiment analysis task, the observed differences between general-purpose LLMs and domain-specific financial models may extend to other financial text analysis scenarios. Future research should validate these findings on larger datasets, additional financial domains, and other downstream tasks such as event detection, risk assessment, and investment recommendation. From a practical perspective, the results suggest that combining insights from general-purpose LLMs and domain-specific financial models may provide more reliable decision support for financial analysts and investors by integrating human-oriented sentiment interpretation with market-oriented financial signals.

Keywords: financial sentiment analysis; large language models; FinBERT; stock market prediction; Efficient Market Hypothesis.

References:

- Dong, Y., Wu, F., Zhang, K., Dai, Y., Zhang, S., Ye, W., Chen, S., & Cheng, Z.-Q. (2025). Large language model agents in finance: A survey bridging research, practice, and real-world deployment. *Findings of the Association for Computational Linguistics: EMNLP 2025*, 17889–17907.
- Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *Journal of Finance*, 25(2), 383–417.
- Gu, W., Zhong, Y., Li, S., Wei, C., Dong, L., Wang, Z., & Yan, C. (2024). Predicting stock prices with FinBERT-LSTM: Integrating news sentiment analysis. In *Proceedings of the 2024 8th International Conference on Cloud and Big Data Computing (ICCBDC 2024)* (pp. 67–72).
- Li, X., Xie, H., Chen, L., Wang, J., & Deng, X. (2014). News impact on stock price return via sentiment analysis. *Knowledge-Based Systems*, 69, 14–23.
- Kirtac, K., & Germano, G. (2024). Sentiment trading with large language models. *Finance Research Letters*, 62, 105227.
- Mishev, K., Gjorgjevikj, A., Vodenska, I., Chitkushev, L. T., & Trajanov, D. (2020). Evaluation of sentiment analysis in finance: From lexicons to transformers. *IEEE Access*, 8, 131662–131682.
- Qi, R. (2025). Financial news sentiment analysis and market sentiment prediction based on large language models. *Journal of Computer, Signal, and System Research*, 2(6), 22–31.

Table of Contents:

Program and Organizing Committee.....	I
Foreward	II
Legend	1
Keynote Talks.....	3
<i>Wei Shao</i> The Impact of Duplicate Data on Federated Learning: An Overlooked Issue in Data Quality and Privacy-Preserving Governance	5
Raka Jovanović Synergies Between Machine Learning and Metaheuristics: New Frontiers in Combinatorial Optimization	6
Contributed Talks	7
ARTIFICIAL INTELLIGENCE, ETHICS AND SOCIETY	9
<i>Ljiljana Branković and Olga Timcenko</i> The Dual Role of AI in Human Societies.....	11
<i>Maja Ilić</i> Comparing AI Incident Databases: Challenges in Classification and Interoperability	13
<i>Sara Kabić, Jovana Dedeić, Mladen Nikolić, Sana Stojanović Đurđević</i> Patient Privacy in Healthcare AI Partnerships: The DeepMind Case Revisited...15	
<i>Jovan Kojić</i> Green AI and Circular Economy Adoption in SMEs: Bias, Data Infrastructure, and Structural Constraints in the EU and Serbia.....	17
<i>Vladimir Pavlović, Bogdan Novaković, Olivera J. Bogunović</i> Transformative Challenges of AI-based Technologies on Education.....	19
<i>Dušan Stojanović, Anđelka Zečević</i> Intellectual Property Rights in the Era of AI: Challenges and Perspectives of the Contemporary Legal Framework.....	21

<i>Miljana Tanic</i> Concerns over the use of synthetic data in biomedicine – What could ever go wrong? ...22	22
<i>Olga Timcenko</i> AI-Assisted Learning in Practice: Github Copilot as a Programming Tutor for Non- Computer Science Students.....24	24
<i>Zubova Oksana, Khusnutdinova Margarita, Filipova Aleksandra</i> Ethical Challenges of Integrating AI into the Educational Process: Balancing Technological Progress with Academic Integrity.....26	26
COMPUTER VISION.....	29
<i>Dušan Denčić, Nataša Milosavljević, Tatjana Davidović</i> Usability analysis of Hu and Zernike moments for image classification.....31	31
<i>Jelena Jevremović, Boban Sazdić-Jotić, Rade Pavlović, Dušan Randelović</i> Dataset for UAV Visual Cross-View Geo-Localization in Mixed Forest-Urban Interfaces33	33
<i>Mina Stanković, Stefan Kostić, Dimitrije Gajin, Ognjen Antonijević, Branislav Bajat, Aleksandar Sekulić, Milan Kilibarda, Slobodan Jelić</i> Improving Soil Property Estimation from Spectroscopy Using Elastic-Net-CNN: An OSSL- Based Study with Transfer Learning.....35	35
<i>Daria Varga, Danijela Milekić, Milica Vasić, Sanela Slavković, Bojan Banjac, Lidija Krstanović</i> Machine Learning Based System for Real-Time Orofacial Motor Skills Evaluation.....37	37
DECISION MAKING AND PROBLEM SOLVING.....	39
<i>Anja Arapović</i> AP-8: When an AI System Must Not Decide Which Rule Wins Distinguishing Action Release from Decision-Authority Release in the Evaluation of Language Models.....41	41
<i>Slobodan Jelic, Matea Lukic, Tatjana Davidovic</i> Variable Neighborhood Descent Heuristic for Predictive Cluster-wise Linear Regression43	43
<i>Herwin Speckter, Velicko Vranes, Marko Radulovic</i> Optimising Deep Learning for Small Patient Cohorts: Predicting Gamma Knife Response in Meningioma.....45	45

<i>Nemanja Vučićević, Filip Andrić, Nenad Stojanović, Marina Svičević, Tatjana Stojanović</i> From Point-Valued to Interval-Valued Bipolar Neutrosophic Soft Sets: The Impact of Uncertainty Representation on Decision-Making.....	47
<i>Milica Vučićević, Nenad Stojanović</i> From Uncertain Assessments to Optimal Choice: An NHFSS-Based Decision-Making Algorithm.....	49
EXPLAINABLE AI	51
<i>Tushar Fnu, Anushka Thagle, Supanut Chindawan, Sanjna Sai Chintakunta, Faizan Raza, Maryam Roshanaei, Dušan Ramljak, Albert Elcock, Rich Moore, Praneeth Sunkaavalli, Chandan Shivalingaiah, Gaurav Prabhu, Lukas Sparer, Alexander Neulinger, Arturo Casasa, Sanjeev Sondur, Milan Todorović, Tatjana Davidović, Dragutin Ostojić, Natalija Ranđelović, Radoš Jojić</i> Blockchain in service of Responsible AI Governance: Integrating AI Sentinel with Blockchain-Inspired Accountability Mechanisms	53
<i>Pilar Ibañez</i> From Workplace Burnout to AI-Powered Emotional Support: Gratitude, Emotion Regulation, and the Development of an AI Emotional Companion.....	55
<i>Miloš Košprdić, Tijana Radović</i> What Makes Arguments Persuasive? An Interpretable Analysis of Alignment and Targeted Engagement in Online Discussions.....	57
<i>Marina Svičević, Aleksandar Milenković, Nemanja Vučićević, Branko Arsić, Lazar Krstić</i> Mathematical Foundations of AI and Machine Learning: A Pilot Study Among Graduates of Specialized Secondary School Students.....	59
GENERATIVE AI	61
<i>Xinju Liu, Min Lan</i> Addressing the Illusion of Understanding Through Multi-Agent Dialogue: A Socially Scaffolded Approach to AI-Assisted Academic Reading.....	63
<i>Dimitrije Matović, Aleksandar Milenković, Mladen Nikolić, Sana Stojanović Đurđević</i> Protecting Artistic Identity in the Age of Generative AI	65

<i>Ognjen Ristić</i> From No-Code to Orchestrated Notification Agent.....	67
<i>Milorad Tosic, Nenad Petrovic</i> Semantic Harness for LLM-driven Agents.....	69
<i>Dragiša D. Žunić</i> Market Mechanisms as Transparent Environments for AI Agents.....	71
KNOWLEDGE REPRESENTATION, REASONING AND PLANNING	73
<i>Milan Jovkić, Goran Savić, Milan Segedinac, Peter Steiner</i> Identifying Prerequisites among Competencies: a Large Language Models Approach..	75
<i>Robi Pritržnik, Pavle Boškoski, Tjaša Redek, Sagnika Sen, Biljana Mileva Boshkoska</i> Beyond Exact Skill Matching: Graph-Constrained AI Modeling of Competence Gaps....	77
<i>Muhammad Sohail Nisar, Muhammad Saleem, Axel-Cyrille Ngonga Ngomo, Sebastian Tramp, Nieves Sande, Milos Jovanovik, Mirko Spasić</i> Machine Learning in Optimized SPARQL Query Execution Over RDF Knowledge Graphs.....	79
<i>Vanja Stepanović, Andreja Tepavčević</i> Lattice-valued relational equations and inequations with complete residuated codomain lattice.....	81
<i>Biljana T. Stojanović, Saša N. Malkov, Miloš V. Beljanski, Nenad S. Mitić</i> Temporal Dynamics of Codon Usage Bias in SARS-CoV-2 Proteins	83
MACHINE LEARNING.....	85
<i>Obrad Aničić</i> Smart CNC Manufacturing: Application of Machine Learning for Predictive Monitoring of Tool Condition and Machining Quality.....	87
<i>Andjelija Ž. Ilić, Ivan D. Milovanovich, Vedrana Makević, Nevena Popovac, Radmila Janković, Jelena Trajković, Andrija Vuković, Jovan Jevtić, Silvio R. de Luka</i> Fractal and textural biomarkers in the diagnostics of pediatric inflammatory bowel disease (IBD).....	89

<i>Goran Keković, Vladimir Mikić, Miloš Ilić, Rade Božović</i> Wavelet-Based Optimization of Temporal Windows for LSTM Benzene Concentration Forecasting.....	91
<i>Stefan Nikolić, Vladan Devedžić</i> Take a Sad Song and Make It Better: Iterative Plugin Refinement in Song-Success Prediction.....	93
<i>Marina Popović Krneta, Miljana Tanić</i> Interpreting Machine Learning Model Performance in Biomedicine: From Statistical Performance to Clinical Decision-Making.....	95
NATURAL LANGUAGE PROCESSING	97
<i>Marjana Gligorijević, Milena Vujošević Janičić, Jelena Graovac</i> Same Pace, Different Path: An NLP-based Analysis of Gender Disparities in IT Careers	99
<i>Danka Jokić, Milica Andrić, Nebojša Damnjanović, Toma Knežević</i> Building a Serbian Riddle Benchmark for Multicultural Evaluation of Large Language Models.....	101
<i>Ulfeta Marovac, Anida Vrcić Ammar, Ajla Gudžević</i> Large Language Models and Emotion Annotation: From Open Vocabulary to Standardized Taxonomies.....	103
<i>Mladenović M., Vujičić Stanković S., Kartelj A.</i> A Hybrid Statistical-LLM Pipeline for Generating Synthetic Survey Respondents with Population-Aligned Distributions.....	105
<i>Saša Z. Petalinkar, Ranka Stanković, Cvetana Krstev</i> LLM-Assisted Expansion of Serbian WordNet: Evaluating Prompting Strategies for Synset Candidate Generation.....	107
<i>Sonja D. Radenković</i> Development of an Artificial Intelligence Chatbot Agent for Financial Investment.....	109
<i>Ranka Stanković, Biljana Rujević, Olivera Kitanović</i> Building a Knowledge Graph from the Historical Serbian Encyclopedia <i>Sveznanje</i>	111
<i>Mihailo Škorić</i> Cross-lingual subword mapping from parallel corpora.....	113

Milena Šošić, Ranka Stanković

Exploring Narrative Frames via Emotional and Moral Signals in Text: A Knowledge Graph Approach to Analyze Social Discourse.....115

Irena Vodenska, Ulfeta Marovac, Admir Hamzagić, Dalila Pramenković

Do Domain-Specific Financial Models Still Matter in the Era of Large Language Models?117