

MARTIN TUTEK

FER, University of Zagreb · martin.tutek@gmail.com · mttk.github.io · Google Scholar

RESEARCH INTERESTS

Knowledge Localization and Editing, Explainability and Controllability of Large Language Models

EMPLOYMENT

Postdoctoral Researcher, University of Zagreb	Feb 2025 – Present
Postdoctoral Researcher, Technion	Feb 2024 – Jan 2025
Postdoctoral Researcher, UKP Lab, Technische Universität Darmstadt	Sep 2022 – Dec 2023
Research Consultant, European Commission, Joint Research Centre	Sep 2014 – Sep 2015

EDUCATION

Ph.D. in Computer Science, University of Zagreb <i>Honors: Magna cum laude</i>	2016 – 2022
M.Sc. in Computer Science, University of Zagreb	2012 – 2014
B.Sc. in Computer Science, University of Zagreb	2009 – 2012

PUBLICATIONS

Preprints or under review

- [1] Majer, L, Šnajder, J., **Tutek, M.** (2026). [Evaluating Pluralism in LLMs through Latent Perspectives](#). Preprint & Pluralistic Alignment workshop @ ICML 2026.
- [2] Jukić, J., **Tutek, M.**, Šnajder, J. (2025). [Context Parametrization with Compositional Adapters](#). Under review @ NeurIPS 2026.

Papers in rigorously reviewed journals and conferences

- [3] Nikankin Y., **Tutek M.**, Ashuach T., Rosenfeld J., Belinkov Y. (2026). [Reasoning Models Know What's Important, and Encode It in Their Activations](#). COLM 2026.
- [4] Simhi, A., Barez, F., **Tutek, M.**, Belinkov, Y., Cohen, S.B. (2026). [Old Habits Die Hard: How Conversational History Geometrically Traps LLMs](#). ICML 2026.
- [5] Ashuach, T., Arad, D., Mueller, A., **Tutek, M.**, Belinkov, Y. (2025). [CRISP: Persistent Concept Unlearning via Sparse Autoencoders](#). ACL 2026.
- [6] Simhi, A., Herzig, J., **Tutek, M.**, Itzhak, I., Szpektor, I., Belinkov, Y. (2025). [ManagerBench: Evaluating the Safety-Pragmatism Trade-off in Autonomous LLMs](#). ICLR 2026.
- [7] Kukić, M. L., Čuljak, M., Dukić, D., **Tutek, M.**, Šnajder, J. (2025). [Improving Decoder-only Language Models for Sequence Labeling through Sequence Repetition](#). Findings of EACL 2026.
- [8] Bibhuti, A., Vashishtha, S., Naik, A., **Tutek, M.**, Aditya, S. (2025). [PragWorld: A Benchmark Evaluating LLMs' local world model under Perturbations and Conversational Dynamics](#). AAAI 2026.
- [9] **Tutek, M.**, Hashemi Chaleshtori, F, Marasović, A, Belinkov, Y. (2025). [Measuring Faithfulness of Natural Language Explanations by Unlearning Reasoning Steps](#). **Outstanding Paper @ EMNLP 2025** — one of 7 out of 1811 accepted papers.
- [10] Mueller*, A., Geiger*, A., Wiegrefe*, S., ..., **Tutek, M.**, Zur, A, Bau, D., Belinkov, Y. (2025). [MIB: A Mechanistic Interpretability Benchmark](#). ICML 2025.
- [11] Ashuach, T., **Tutek, M.**, & Belinkov, Y. (2024). [REVS: Unlearning Sensitive Information in Language Models via Rank Editing in the Vocabulary Space](#). Findings of ACL 2025.
- [12] Puerto, H., **Tutek, M.**, Aditya, S., Zhu, X., & Gurevych, I. (2024). [Code Prompting Elicits Conditional Reasoning Abilities in Text+ Code LLMs](#). EMNLP 2024.
- [13] Jelenić, F., Jukić, J., **Tutek, M.**, Puljiz, M., & Šnajder, J. (2024). [Out-of-Distribution Detection by Leveraging Between-Layer Transformation Smoothness](#). ICLR 2024.

- [14] Sachdeva, R., **Tutek, M.**, & Gurevych, I. (2024). CATfOOD: Counterfactual Augmented Training for Improving Out-of-Domain Performance and Calibration. EACL 2024.
- [15] Jukić, J. *, **Tutek, M. ***, & Šnajder, J. (2023). Easy to Decide, Hard to Agree: Reducing Disagreements Between Saliency Methods. Findings of ACL 2023. **Equal contribution.*
- [16] **Tutek, M.**, & Snajder, J. (2022). Toward Practical Usage of the Attention Mechanism as a Tool for Interpretability. IEEE Access.
- [17] **Tutek, M.**, Glavas, G., Šnajder, J., Milić-Frayling, N., & Dalbelo Basic, B. (2016). Detecting and Ranking Conceptual Links between Texts Using a Knowledge Base. CIKM 2016.

Workshop papers

- [18] Soker, Y., **Tutek, M.**, & Belinkov, Y. (2025). Predicting Success of Model Editing Through Intrinsic Features. Interplay workshop @ COLM 2025.
- [19] Dukić, D., Barić, A., Čuljak, M., Jukić, J., **Tutek, M.** (2025). Characterizing Linguistic Shifts in Croatian News via Diachronic Word Embeddings. Slavic NLP workshop @ ACL 2025.
- [20] Obadić, L., **Tutek, M.**, & Šnajder, J. (2022). NLPOP: a Dataset for Popularity Prediction of Promoted NLP Research on Twitter. Computational Approaches to Subjectivity, Sentiment & Social Media Analysis workshop @ ACL 2022.
- [21] **Tutek, M.** & Šnajder, J. (2020). Staying True to Your Word: (How) Can Attention Become Explanation?. Representation Learning for NLP Workshop @ ACL 2020.
- [22] **Tutek, M.** & Šnajder, J. (2018). Iterative Recursive Attention Model for Interpretable Sequence Classification. BlackBoxNLP workshop @ EMNLP 2018.
- [23] di Buono, M. P., Šnajder, J., Dalbelo Bašić, B., Glavaš, G., **Tutek, M.**, & Milic-Frayling, N. (2017). Predicting News Values from Headline Text and Emotions. Natural Language Processing Meets Journalism workshop @ EMNLP 2017.
- [24] Rotim, L., **Tutek, M.**, & Šnajder, J. (2017). TakeLab at SemEval-2017 Task 5: Linear aggregation of word embeddings for fine-grained sentiment analysis of financial news. Workshop on Semantic Evaluation (SemEval) @ ACL 2017.
- [25] di Buono, M. P., **Tutek, M.**, Šnajder, J., Glavaš, G., Bašić, B. D., & Milic-Frayling, N. (2017). Two Layers of Annotation for Representing Event Mentions in News Stories. Linguistic Annotation Workshop @ EACL 2017.
- [26] **Tutek, M.**, Sekulić, I., Gombar, P., Paljak, I., Čulinović, F., Boltužić, F., Karan, M., Alagić, D. and Šnajder, J. (2016). TakeLab at SemEval-2016 Task 6: Stance classification in tweets using a genetic algorithm based ensemble. Workshop on Semantic Evaluation (SemEval) @ NAACL 2016.

FUNDED PROPOSALS

Coefficient Giving, research grant, \$98,340

February 2026

Title: Uncovering True Reasons of LMs Predictions

HONORS, AWARDS AND FELLOWSHIPS

- Outstanding paper award @ EMNLP 2025
- Outstanding reviewer @ ARR October 2023
- DAAD-STIBET fellowship for visiting doctoral students, 2018 (visited University of Mannheim, autumn 2018)

INVITED TALKS

From Internals to Integrity: How Insights into Transformer LMs Improve Safety and Faithfulness

- Lamarr Institute, University of Bonn Mar 2026
- Data and Web Science Group, University of Mannheim Feb 2026
- RTG Neuroexplicit models, University of Saarland Feb 2026
- IMS, University of Stuttgart Feb 2026
- Bocconi University, Milan Oct 2025
- University of Rijeka Jun 2025
- “The Cross-Disciplinary Impact of Transformers and LLMs” lecture series, FER, University of Zagreb

On interpretability: attention, saliency and beyond

- Technion, Israel Institute of Technology Jun 2022

TEACHING

222925: Text Analysis and Retrieval, University of Zagreb <i>Lecturer</i>	Summer 2026 – present
209719: Introduction to Artificial Intelligence, University of Zagreb <i>Lecturer (~500 students)</i>	Summer 2025 – present
20-00-0947: Deep Learning for NLP, TU Darmstadt <i>Lecturer (50% of course; 130 students)</i>	Summer 2023
252377: Deep Learning 1, University of Zagreb <i>Course design, lab assignments and lectures (2 lectures)</i>	Summer 2018 – 2021
222925: Text Analysis and Retrieval, University of Zagreb <i>Lecturer (2 lectures)</i>	Fall 2018 – 2021
209719: Introduction to Artificial Intelligence, University of Zagreb <i>Lab assignments and lectures (1 lecture)</i>	Summer 2016 – 2022

MENTORING

University of Zagreb PhD students (with Jan Šnajder)

- **Marko Čuljak** Feb 2025 – present
 - Papers @ Slavic NLP, Findings of EACL 2026
- **Josip Jukić** (now Postdoc at University of Amsterdam) Aug 2021 – Jul 2025
 - Papers @ Findings of ACL 2023, ICLR 2024, preprint

Technion PhD students (with Yonatan Belinkov)

- **Tomer Ashuach** Feb 2024 – present
 - Papers @ Findings of ACL 2025, ACL 2026
- **Adi Simhi** Jun 2025 – present
 - Papers @ ICLR 2026, ICML 2026

Darmstadt PhD students (with Iryna Gurevych)

- **Haritz Puerto** (now Postdoc at ELLIS Institute Tübingen) Dec 2022 – Dec 2023
 - Paper @ EMNLP 2024
- **Rachneet Sachdeva** Sep 2022 – Oct 2023
 - Paper @ EACL 2024

SERVICE

Organizing experience

- BlackBoxNLP 2026 — Member of organizing committee (blackboxnlp.github.io/2026)
- BlackBoxNLP 2025 — Shared task: benchmarking new techniques for localizing circuits and causal latent variables in language models (blackboxnlp.github.io/2025/task)

Area Chair: Interpretability and Analysis of Models for NLP

- ACL Rolling Review Dec 2023 – Jan 2026
- COLING 2024
- EMNLP 2023
- ACL 2023

Conference reviewer

- AAAI 2026
- ICLR 2025, 2026
- COLM 2024, 2025
- NeurIPS 2024, 2025
- ARR Nov 2021 – Oct 2023 (outstanding reviewer Oct 2023)
- EACL 2023
- EMNLP 2018 – 2022

- ACL 2018 – 2022

Journal reviewer

- Automatika 2020, 2021
- Artificial Intelligence 2021, 2022

Workshop reviewer

- Pluralistic Alignment Workshop @ ICML 2026
- Mechanistic Interpretability Workshop @ ICML 2024, 2026
- BlackboxNLP @ EMNLP 2018, EMNLP 2025
- L2M2 Workshop @ ACL 2025
- Actionable Interpretability Workshop @ ICML 2025

Summer school lecturer

- Int'l Summer School of Data Science in Split, practical sessions. 2016, 2017

REFERENCES

- **Yonatan Belinkov**, belinkov@technion.ac.il
- **Jan Šnajder**, jan.snajder@fer.hr
- **Ana Marasović**, ana.marasovic@utah.edu