

University of Zagreb
martin.tutek@gmail.com

MARTIN TUTEK

<https://mttk.github.io/>
[Google Scholar](#)

RESEARCH INTERESTS

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

EMPLOYMENT

Postdoctoral Researcher, University of Zagreb - Substituting for prof. Jan Šnajder in summer semester 24/25	Feb 2025 – Present
Postdoctoral Researcher, Technion	Feb 2024 – Feb 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	2016 – 2022
M.Sc. In Computer Science, University of Zagreb	2012 – 2014
B.Sc. In Computer Science, University of Zagreb	2009 – 2012

PUBLICATIONS

Selected Publications

- [*] **Tutek, M.**, Hashemi Chaleshtori, F, Marasović, A, Belinkov, Y. (2025). [Measuring Faithfulness of Natural Language Explanations by Unlearning Reasoning Steps](#). EMNLP 2025. **Under review for Outstanding Paper Award.**
- [*] Mueller*, A., Geiger*, A., Wiegrefe*, S., ..., **Tutek, M.**, Zur, A, Bau, D., Belinkov, Y. (2025). MIB: [A Mechanistic Interpretability Benchmark](#). ICML 2025.
- [*] **Tutek, M.** & Šnajder, J. (2020). [Staying True to Your Word:\(How\) Can Attention Become Explanation?](#). Representation Learning for NLP workshop @ ACL 2020.

Under review or awaiting decision

- [1] Kukić, M. L, Čuljak, M., Dukić, D., **Tutek, M.**, Šnajder, J. (2025). Improving Decoder-only Language Models for Sequence Labeling through Sequence Repetition. → under review @ ARR October 2025
- [2] Simhi, A., Herzig, J., **Tutek, M.**, Itzhak, I., Szpektor, I., Belinkov, Y. (2025). [ManagerBench: Evaluating the Safety-Pragmatism Trade-off in Autonomous LLMs](#). → under review @ ICLR 2026.
- [3] Jukić, J., **Tutek, M.**, Šnajder, J. (2025). [Context Parametrization with Compositional Adapters](#) → under review @ ICLR 2026
- [4] Bibhuti, A., Vashishtha, S., Naik, A., **Tutek, M.**, Aditya, S. (2025). PragWorld: A Benchmark Evaluating LLMs' local world model under Perturbations and Conversational Dynamics. → awaiting decision @ AAAI 2026.
- [5] Ashuach, T., Arad, D., Mueller, A., **Tutek, M.**, Belinkov, Y. (2025). [CRISP: Persistent Concept Unlearning via Sparse Autoencoders](#). reviewed at ARR July 2025 → commit to ACL 2026.

Papers in Rigorously Reviewed Journals and Conferences

- [6] **Tutek, M.**, Hashemi Chaleshtori, F, Marasović, A, Belinkov, Y. (2025). [Measuring Faithfulness of Natural Language Explanations by Unlearning Reasoning Steps](#). EMNLP 2025. **Under review for Outstanding Paper Award.**
- [7] Mueller*, A., Geiger*, A., Wiegrefe*, S., ..., **Tutek, M.**, Zur, A, Bau, D., Belinkov, Y. (2025). [MIB: A Mechanistic Interpretability Benchmark](#). ICML 2025.
- [8] 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.

- [9] Puerto, H., **Tutek, M.**, Aditya, S., Zhu, X., & Gurevych, I. (2024). [Code Prompting Elicits Conditional Reasoning Abilities in Text+ Code LLMs](#). EMNLP 2024.
- [10] Jelenić, F., Jukić, J., **Tutek, M.**, Puljiz, M., & Šnajder, J. (2024). [Out-of-Distribution Detection by Leveraging Between-Layer Transformation Smoothness](#). ICLR 2024.
- [11] Sachdeva, R., **Tutek, M.**, & Gurevych, I. (2024). [CATFOOD: Counterfactual Augmented Training for Improving Out-of-Domain Performance and Calibration](#). EACL 2024.
- [12] Jukić, J.*, **Tutek, M.***, & Šnajder, J. (2023). [Easy to Decide, Hard to Agree: Reducing Disagreements Between Saliency Methods](#). Findings of ACL 2023 *Equal contribution.
- [13] **Tutek, M.**, & Šnajder, J. (2022). [Toward Practical Usage of the Attention Mechanism as a Tool for Interpretability](#). IEEE Access.
- [14] **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

- [15] Soker, Y., **Tutek, M.**, & Belinkov, Y. (2025). [Predicting Success of Model Editing Through Intrinsic Features](#). Interplay workshop @ COLM 2025.
- [16] 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.
- [17] 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.
- [18] **Tutek, M.** & Šnajder, J. (2020). [Staying True to Your Word:\(How\) Can Attention Become Explanation?](#). Representation Learning for NLP Workshop @ ACL 2020.
- [19] **Tutek, M.** & Šnajder, J. (2018). [Iterative Recursive Attention Model for Interpretable Sequence Classification](#). In BlackBoxNLP workshop @ EMNLP 2018.
- [20] 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.
- [21] Rotim, L., **Tutek, M.**, & Šnajder, J. (2017, August). [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.
- [22] 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.
- [23] **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](#). IWorkshop on Semantic Evaluation (SemEval) @ NAACL 2016.

INVITED TALKS

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

- *University of Rijeka* Oct 2025
- *“The Cross-Disciplinary Impact of Transformers and LLMs” lecture series*, Jun 2025
FER, University of Zagreb

On interpretability: attention, saliency and beyond

- Technion Jun 2022

TEACHING

- 209719: Introduction to Artificial Intelligence, University of Zagreb Summer 2025
- *Lecturer (50% of course; 580 students)*
- 20-00-0947: Deep Learning for Natural Language Processing, TU Darmstadt Summer 2023

- Lecturer (50% of course; 130 students)	
252377: Deep Learning 1, University of Zagreb	Summer 2018-2021
- Course design, lab assignments and lectures (2 lectures)	
222925: Text Analysis and Retrieval, University of Zagreb	Autumn 2018-2021
- Lecturer (2 lectures)	
209719: Introduction to Artificial Intelligence, University of Zagreb	Summer 2016-2022
- Lab assignments and lectures (1 lecture)	

MENTORING

University of Zagreb PhD students (with Jan Šnajder)	
- Josip Jukić (now Postdoc at University of Zagreb)	Aug 2021 – Jul 2025
- Papers @ ACL 2023 Findings, ICLR 2024, preprint	
Technion PhD Students (with Yonatan Belinkov)	
- Tomer Ashuach (now PhD at Technion)	Feb 2024 – present
- Papers @ ACL 2025 Findings, preprint	
Darmstadt PhD Students (with Iryna Gurevych)	
- Haritz Puerto	Dec 2022 – Dec 2023
- Paper @ EMNLP 2024 Main	
- Rachneet Sachdeva	Sep 2022 – Oct 2023
- Paper @ EACL 2024 Main	

SERVICE

Organizing experience:

- **BlackBoxNLP 2025** Shared task: *benchmarking new techniques for localizing circuits and causal latent variables in language models* (<https://blackboxnlp.github.io/2025/task/>)

Area Chair: Interpretability and Analysis of Models for NLP

- COLING 2024
- ACL Rolling Review
Dec 2023–current
- 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

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

Summer School Lecturer

Intl' Summer School of Data Science in Split, practical sessions. 2016, 2017

OPEN-SOURCE SOFTWARE AND DATASETS

- **Podium**: A framework agnostic python NLP library for data loading and preprocessing
<https://github.com/TakeLab/podium>
- **RNN-classifier**: A minimalistic RNN classifier sample
<https://github.com/mttk/rnn-classifier>
- **pytorch**
Contributor
<https://github.com/pytorch/pytorch>
- **pytorch-text**
Contributor, maintainer
<https://github.com/pytorch/text>

REFERENCES

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