

University of Zagreb
martin.tutek@gmail.com

MARTIN TUTEK

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

RESEARCH INTERESTS

Natural Language Processing, Faithful Explainability, Safe and Reliable AI

EMPLOYMENT

Postdoctoral Researcher, University of Zagreb	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](#). Under review.
- [*] 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?](#). In Proceedings of the 5th Workshop on Representation Learning for NLP (pp. 131-142).

PREPRINTS OR UNDER REVIEW

- [1] Bibhuti, A., Vashishtha, S., Naik, A., **Tutek, M.**, Aditya, S. (2025). PragWorld: A Benchmark Evaluating LLMs' local world model under Perturbations and Conversational Dynamics. Under review.
- [2] Dukić, D., Barić, A., Čuljak, M., Jukić, J., **Tutek, M.** (2025). Characterizing Linguistic Shifts in Croatian News via Diachronic Word Embeddings. Under review.
- [3] **Tutek, M.**, Hashemi Chaleshtori, F, Marasović, A, Belinkov, Y. (2025). [Measuring Faithfulness of Natural Language Explanations by Unlearning Reasoning Steps](#). Under review.
- [4] Soker, Y., **Tutek, M.**, & Belinkov, Y. (2025). Predicting Success of Model Editing Through Intrinsic Features. Under review.

PEER-REVIEWED PUBLICATIONS

- [5] Mueller*, A., Geiger*, A., Wiegrefe*, S., ..., **Tutek, M.**, Zur, A, Bau, D., Belinkov, Y. (2025). [MIB: A Mechanistic Interpretability Benchmark](#). ICML 2025.

- [6] Ashuaach, T., **Tutek, M.**, & Belinkov, Y. (2024). [REVS: Unlearning Sensitive Information in Language Models via Rank Editing in the Vocabulary Space](#). Findings of ACL 2025.
- [7] Puerto, H., **Tutek, M.**, Aditya, S., Zhu, X., & Gurevych, I. (2024). [Code Prompting Elicits Conditional Reasoning Abilities in Text+ Code LLMs](#). EMNLP 2024.
- [8] Jelenić, F., Jukić, J., **Tutek, M.**, Puljiz, M., & Šnajder, J. (2024). [Out-of-Distribution Detection by Leveraging Between-Layer Transformation Smoothness](#). ICLR 2024.
- [9] Sachdeva, R., **Tutek, M.**, & Gurevych, I. (2024). [CATFOOD: Counterfactual Augmented Training for Improving Out-of-Domain Performance and Calibration](#). EACL 2024.
- [10] Jukić, J.*, **Tutek, M.***, & Šnajder, J. (2023). [Easy to Decide, Hard to Agree: Reducing Disagreements Between Saliency Methods](#). Findings of ACL 2023 ***Equal contributions**.
- [11] **Tutek, M.**, & Šnajder, J. (2022). [Toward Practical Usage of the Attention Mechanism as a Tool for Interpretability](#). IEEE Access.
- [12] Obadić, L., **Tutek, M.**, & Šnajder, J. (2022). [NLPOP: a Dataset for Popularity Prediction of Promoted NLP Research on Twitter](#). In Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (pp. 286-292).
- [13] **Tutek, M.** & Šnajder, J. (2020). [Staying True to Your Word:\(How\) Can Attention Become Explanation?](#). In Proceedings of the 5th Workshop on Representation Learning for NLP (pp. 131-142).
- [14] **Tutek, M.** & Šnajder, J. (2018). [Iterative Recursive Attention Model for Interpretable Sequence Classification](#). In Proceedings of the 2018 EMNLP Workshop: Analyzing and interpreting neural networks for NLP.
- [15] 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](#). In 2017 EMNLP Workshop on Natural Language Processing Meets Journalism (pp. 1-6).
- [16] 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](#). In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017) (pp. 866-871).
- [17] 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](#). LAW XI 2017, 82.
- [18] **Tutek, M.**, Glavas, G., Šnajder, J., Milić-Frayling, N., & Dalbelo Basic, B. (2016, October). [Detecting and Ranking Conceptual Links between Texts Using a Knowledge Base](#). In Proceedings of the 25th ACM International on Conference on Information and Knowledge Management (pp. 2077-2080).
- [19] **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](#). In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016) (pp. 464-468).

INVITED TALKS

Using, Improving and Calibrating (Large) Language Models

- CroAI Meetup Mar 2024

ChatGPT: What can large pre-trained language models say about the future?

- CroAI Meetup May 2023
- IEEE Croatia, University of Zagreb Dec 2022

On interpretability: attention, saliency and beyond

- Technion Jun 2022

TEACHING

209719: Introduction to Artificial Intelligence, University of Zagreb	Summer 2025
20-00-0947: Deep Learning for Natural Language Processing, TU Darmstadt	Summer 2023
252377: Deep Learning 1, University of Zagreb	Summer 2018-2021
222925: Text Analysis and Retrieval, University of Zagreb	Autumn 2018-2021
209719: Introduction to Artificial Intelligence, University of Zagreb	Summer 2016-2022

MENTORING

Technion MA Researchers (*with Yonatan Belinkov*)

- Tomer Ashuach (*now PhD at Technion*) Feb 2024 – present
- Yanay Soker Feb 2024 – Sep 2024

Darmstadt PhD Students (*with Iryna Gurevych*)

- Haritz Puerto Dec 2022 – Dec 2023
- Rachneet Sachdeva Sep 2022 – Oct 2023

Darmstadt MA Researchers (*with Iryna Gurevych*)

- Dorian Arnouts Feb 2023 – Aug 2023

University of Zagreb PhD students (*with Jan Šnajder*)

- Josip Jukić Aug 2021 – present

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

- ICLR 2025
- 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

- Mechanistic Interpretability Workshop @ICML 2024
- BlackboxNLP @EMNLP 2018

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