Roi Cohen

NATURAL LANGUAGE PROCESSING · LARGE LANGUAGE MODELS · FACTUALITY, UNCERTAINTY & SAFETY
Berlin, Germany · roi3993@gmail.com
Homepage · Google Scholar · LinkedIn · GitHub · X

RESEARCH SUMMARY

Ph.D. researcher in natural language processing focused on the factuality, uncertainty, and safety of large language models — developing methods that reduce hallucination and enable models to recognize and express the limits of their own knowledge, across pretraining, alignment, and inference. First author of widely-read work on hallucination detection and knowledge in LLMs (LM vs LM, Ripple Effects of Knowledge Editing, the [IDK] token), with publications at NeurIPS, EMNLP, TACL, and EACL. Currently a visiting researcher in Prof. Monica S. Lam's group at Stanford University and a research intern at Superhuman, with prior research experience at IBM Research, Microsoft, and Israel's Unit 8200.

EDUCATION

Ph.D. in Computer Science — Hasso Plattner Institute (HPI), Artificial Intelligence Group 2023 – 2027 (expected)
Berlin / Potsdam, Germany. Advisor: Prof. Gerard de Melo. Co-supervising two M.Sc. students.
Research: factuality, uncertainty, information retrieval, question answering, alignment, reasoning, interpretability, and LLM safety.

M.Sc. in Computer Science — Tel Aviv University 2021 – 2023
Advisors: Prof. Amir Globerson, Prof. Jonathan Berant, Prof. Mor Geva. Teaching Assistant, NLP course.
Focus: knowledge and factuality in language models, IR, QA, knowledge graphs, reasoning.

B.Sc. in Computer Science — Tel Aviv University 2018 – 2021
GPA 91.78 / 100. Member of the Computer Science Excellence Program (honors track).
Final project: a deep encoder for tabular data containing free-text cells.

PUBLICATIONS

Peer-Reviewed

1. **InFact: Informativeness Alignment for Improved LLM Factuality. Roi Cohen**, Russa Biswas, Gerard de Melo. *Findings of EMNLP 2025.* [link]
2. **I Don't Know: Explicit Modeling of Uncertainty with an [IDK] Token. Roi Cohen**, Konstantin Dobler, Eden Biran, Gerard de Melo. *NeurIPS 2024.* [link]
3. **Evaluating the Ripple Effects of Knowledge Editing in Language Models. Roi Cohen**, Eden Biran, Ori Yoran, Amir Globerson, Mor Geva. *TACL 2024.* [link]
4. **LM vs LM: Detecting Factual Errors via Cross-Examination. Roi Cohen**, May Hamri, Mor Geva, Amir Globerson. *EMNLP 2023.* [link]
5. **Crawling the Internal Knowledge-Base of Language Models. Roi Cohen**, Mor Geva, Jonathan Berant, Amir Globerson. *Findings of EACL 2023.* [link]

Preprints & Under Review

6. **Unifying Latent Uncertainty Signals in Large Language Models for Improved Factual Precision. Roi Cohen**, Omri Fahn, Gerard de Melo. *Under review, ICLR 2026.* [link]
7. **Uncertainty Beyond Geometry: Kernel-Based Functional Routing in LLMs. Roi Cohen**, Omri Fahn, Gerard de Melo. *Under review, COLM 2026.*
8. **Knowledgeless Language Models: Decoupling Linguistic Competence and Factual Knowledge. Roi Cohen**, Yvan Carre, Nick Lechtenborger, Hendrik Droste, Lucas Kerschke, Russa Biswas, Gerard de Melo, Jan Buys. *Under review, NeurIPS 2026.*

9. **StyleAT: Defending Face Recognition Against Semantic Attacks.** Ben Shapira, **Roi Cohen**, Shang-Tse Chen, Mahmood Sharif. *Under review, ICML 2026.* [link]

Workshop Papers

10. **Metacognition and the Source of Competence in Large Language Models. Roi Cohen** et al. *ICML 2026 Workshop; extended version in preparation.*

Research & Industry Experience

Stanford University — Visiting Researcher, group of Prof. Monica S. Lam 2026 – Present
Visiting research project in Prof. Lam's group.

Superhuman — Research Intern 2026 – Present
Metacognition of LLM-as-judge systems in the text-revision setting.

IBM Research — Research Scientist, Language Department Jun 2023 – Dec 2024
Instruction alignment and tuning, tool-use training, retrieval-augmented generation (RAG), in-context learning, and active learning.

Second Nature AI — NLP Researcher 2022 – Jan 2023
Developed NLP for a virtual assistant aimed at sales and marketing teams.

Microsoft — Data Science & Research Intern 2020 – 2022
Improved deep architectures for tabular data containing free text.

Tel Aviv University — Data Scientist, Department of Computational Biology 2019 – 2020
Analysis of genetic data.

Unit 8200, Israel Defense Forces — Researcher & Developer (team lead, final year) 2014 – 2019
Network-security and vulnerability research; designed and built small- and large-scale software systems.

Honors & Awards

Celia & Marcos Maus Prize for Excellence in Research.
Computer Science Excellence Program, Tel Aviv University (2018 – 2021).

Professional Service & Media

Reviewer for NeurIPS and ACL.
Featured in *The Wall Street Journal*, "Why Do AI Chatbots Have Such a Hard Time Admitting 'I Don't Know'" [article]

Technical Skills

Programming: Python (10+ years), C, C++, Java, R.
ML frameworks & tools: PyTorch, Hugging Face Transformers.
Languages: English (fluent), Hebrew (native).

References

Available upon request.