Roi Cohen

roi3993@gmail.com • twitter.com/roicohen9 • scholar.google.com/citations?hl=en&user=TcPUXJYAAAAJ

Education

HPI **2023 – 2026**
PhD in Computer Science **Berlin, Germany**

- Main fields: Knowledge of LMs, Factuality, Interpretability, Alignment, Tool-Use, IR, QA, KGs, Reasoning, Bias, Responsible AI
- Research Advisors: Prof. Gerard de Melo

Tel Aviv University **2021 – 2023**
M.Sc. in Computer Science **Tel Aviv, Israel**

- Focus domains: NLP. Specifically – Knowledge of LMs, Factuality, IR, QA, KGs, Reasoning, Responsible AI
- Research Advisors: Prof. Amir Globerson, Prof. Jonathan Berant, Prof. Mor Geva
- Was the NLP course teaching assistant.

Tel Aviv University **2018 – 2021**
B.Sc. in Computer Science **Tel Aviv, Israel**
GPA: 90.78/100

- Was a member of the Excellence program of the Computer Science School.
- Final project focused on implementing a deep encoder for tabular data containing textual cells.

Unit 8200 **2015 (for five months)**
Networking and Binary Communication analysis course

- Proficiency in Python, SQL, Data Science and Computer Networking Algorithms.
- Chosen to be the outstanding student of that course, out of 14 students.

Industry Experience

IBM Research, Language Department **June 2023 – Sep 2023, June 2024 – Sep 2024**
Research Intern

- Main research topics: Factuality, In-Context Learning, Active Learning
- Was working on developing active learning methods to improve factuality.

Second Nature AI **2022 – Jan 2023**
NLP researcher

- Was working on a virtual assistant for marketing people.

Microsoft **2020 – 2022**
Data Science/Research Intern

- Projects focused on improving current deep architectures for handling tabular data contain text.

Tel Aviv University, Department of Computational Biology **2019 – 2020**
Data Scientist

- Worked on projects mainly focus on analyzing genetic data.

IDF, R&D Team, Unit 8200 **2015 – 2019**
Algorithm Researcher & Team Leader

- Research and development of algorithms that analyze networking data.

First Author Publications

- Roi Cohen, Mor Geva, Jonathan Berant and Amir Globerson. *Crawling the internal Knowledge-Base of Language Models. Findings of EACL 2023.* https://arxiv.org/abs/2301.12810

- Roi Cohen, May Hamri, Mor Geva and Amir Globerson. *LM vs LM: Detecting Factual Errors via Cross-Examination. EMNLP 2023.*
  https://arxiv.org/abs/2305.13281
- Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson and Mor Geva. *Evaluating the Ripple Effects of Knowledge Editing in Language Models. TACL 2024.*
  https://arxiv.org/abs/2307.12976

- Roi Cohen, Konstantin Dobler, Eden Biran and Gerard de Melo I Don't Know: Explicit Modeling of Uncertainty with an [IDK] Token *Neurips 2024.*
  https://proceedings.neurips.cc/paper_files/paper/2024/file/14c018d2e72c521605b0567029ef0efb-Paper-Conference.pdf

- Roi Cohen, Russa Biswas and Gerard de Melo InFact: Informativeness Alignment for Improved LLM Factuality *submitted to EMNLP 2025.*
  https://arxiv.org/abs/2505.20487

- Roi Cohen, Omri Fahn and Gerard de Melo Pretrained LLMs Learn Multiple Types of Uncertainty *submitted to Neurips 2025.*
  https://arxiv.org/abs/2505.21218

AWARDS

- CELIA & MARCOS MAUS PRIZE in Computer Science for Excellence in Research

APPEARANCE IN JOURNALISM

- "Why Do AI Chatbots Have Such a Hard Time Admitting "I Don't Know" "- The Wall Street Journal.
  https://www.wsj.com/tech/ai/ai-halluciation-answers-i-dont-know-738bde07?st=ykKM29

ADDITIONAL NOTES

- Serve as an official Neurips Reviewer

PROGRAMMING LANGUAGES

- Python (10 years of experience), Java, C++, C, R

LANGUAGES

- English (Fluent), Hebrew (Native)