Tal Haklay – CV

Email: talhaklay535@gmail.com | **Scholar:** Tal Haklay | **Linkedin:** Tal Haklay

PhD student in the NLP Lab at the Technion, advised by Yonatan Belinkov.
My research focuses on developing interpretability methods for large language models.

EDUCATION

Technion

Feb 2025 - Present	PhD Candidate in Computer science. Advised by Yonatan Belinkov.
Mar 2022 - Feb 2025	MSc in Computer science. GPA: 94.5.
Oct 2018 – Mar 2022	BSc in Computer science. President's List - 1 semester. Dean's List - 3 semesters.

PUBLICATIONS

- **Linearity of Relation Decoding in Transformer Language Models** [Arxiv]
  Evan Hernandez, Arnab Sen Sharma, Tal Haklay, Kevin Meng, Martin Wattenberg, Jacob Andreas, Yonatan Belinkov, David Bau
  In International Conference on Learning Representations (ICLR 2024). **Spotlight paper.**

- **Fine-Tuning Enhances Existing Mechanisms: A case study on Entity Tracking** [Arxiv]
  Nikhil Prakash, Tamar Rott Shaham, Tal Haklay, Yonatan Belinkov, David Bau
  In International Conference on Learning Representations (ICLR 2024).

- **MIB: A Mechanistic Interpretability Benchmark** [Arxiv]
  Aaron Mueller, Atticus Geiger, Sarah Wiegreffe, Dana Arad, Iván Arcuschin, Adam Belfki, Yik Siu Chan, Jaden Fiotto-Kaufman, Tal Haklay, Michael Hanna, Jing Huang, Rohan Gupta, Yaniv Nikankin, Hadas Orgad, Nikhil Prakash, Anja Reusch, Aruna Sankaranarayanan, Shun Shao, Alessandro Stolfo, Martin Tutek, Amir Zur, David Bau, Yonatan Belinkov
  Forty-Second International Conference on Machine Learning (ICML 2025)

- **Position-aware Automatic Circuit Discovery** [Arxiv]
  Tal Haklay, Hadas Orgad, David Bau, Aaron Mueller, Yonatan Belinkov
  The 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025)

PROFESSIONAL EXPERIENCE

Teaching Assistant

Technion	Introduction to Artificial Intelligence (236501).
Mar 2022 – Mar 2024	Introduction to Natural Language Processing (236299).

Student SW Engineer

Intel Jun 2020 – Feb 2022	Collaborating with algorithm researchers in the image processing field and implementing algorithms in C++. Developing scripts in Python to support and increase automation.

PRIZES

- 2024 Final Prize for Outstanding Master's Students in Computer Science, Electrical Engineering, Mathematics, or Physics.
- 2020 Excellence Award for Social Involvement on behalf of the Faculty of Computer Science at the Technion.

ACADEMIC SERVICES

- Co-organizer of the first Actionable Interpretability Workshop at ICML 2025.

- Reviewer for the BlackBoxNLP Workshop at EMNLP 2024.
- Reviewer for the Mechanistic Interpretability for Vision Workshop at CVPR 2025.
- Reviewer for NuerIPS 2025.

SOCIAL ACTIVITIES

- Manager, *Baot – Researchers* community, the largest network of female researchers in Israel.
- Co-Founder of “She S – Woman in CS @ Technion”, he official community for female students at the Technion's Computer Science Faculty.
- Initiated and led a mentorship program for freshmen, recruiting and managing over 150 volunteers. Received the Excellence Award for Social Involvement for this initiative.
- Co-organize , “CS Hackathon 2022 – Doing Good”.