Guy Kaplan

PhD Student, Computer Science · The Hebrew University of Jerusalem

guy.kaplan3@mail.huji.ac.il · +972 54 432 5657 · LinkedIn · Google Scholar

Education

The Hebrew University of Jerusalem — *Jerusalem, Israel*
PhD Student, Computer Science — Aug. 2025 – Present

Advisor: Prof. Roy Schwartz, School of Computer Science & Engineering.
Research: inner lexicon & token-level representations of LLMs; mechanistic interpretability.

The Hebrew University of Jerusalem — *Jerusalem, Israel*
M.Sc., Computer Science (GPA: 95/100) — Oct. 2022 – Aug. 2025

Advisor: Prof. Roy Schwartz, School of Computer Science & Engineering.
Thesis: *The Inner Lexicon of Large Language Models* — published at ICLR 2025.

The Open University of Israel — *Israel*
B.Sc., Computer Science (GPA: 90/100, Dean's Honours 2021) — Oct. 2019 – Jun. 2022

Tel Aviv University — *Tel Aviv, Israel*
B.A., Economics — Oct. 2019 – Jun. 2022

Concurrent with B.Sc. studies at the Open University of Israel.

University of Haifa — *Haifa, Israel*
B.A., General Studies, Magna Cum Laude (GPA: 93/100) — 2014

Research & Professional Experience

University of Southern California (USC) — *Los Angeles, CA*
Visiting Researcher — Sep. 2025 – Present

- Full-year official research visit; active collaborations on reward hacking in LLMs and mechanistic interpretability.

Microsoft — *Israel*
Data Scientist — Mar. 2024 – Jan. 2025

- Built large-scale NLP and anomaly-detection models for risky-user detection.

Bright Forensic Innovations — *Tel Aviv, Israel*
Chief Scientist Officer — Oct. 2023 – Present

- Developed AI method for cadaver identification from dental records; operationally deployed by the Israeli Police for October 7 victim identification; raised top-10 accuracy from 10% to 89%.
- Lectured the Israeli Police forensic unit on the AI methodology.

Microsoft — *Israel*
Software Engineer — Mar. 2022 – Mar. 2024

Built EventHub & Cosmos DB features for the Azure Defender UEBA product; drove CI/CD rollout.

Israeli Navy — Air Patrol Unit
Naval Officer (Major) — Oct. 2011 – Sep. 2019

Commanded 50 personnel; earned Palmachim AFB Commander's Award.

Publications

[1] *From Tokens to Words: On the Inner Lexicon of LLMs*
Guy Kaplan, Matanel Oren, Yuval Reif, Roy Schwartz
International Conference on Learning Representations (ICLR), 2025 (presented in Singapore).
arxiv.org/abs/2410.05864

[2] *Follow the Flow: On Information Flow Across Textual Tokens in Text-to-Image Models*
Guy Kaplan[*], Michael Toker[*], Yuval Reif, Yonatan Belinkov, Roy Schwartz ([*]equal contribution)
Annual Meeting of the Association for Computational Linguistics (ACL), 2026
`arxiv.org/abs/2504.01137`

[3] *Vocab Diet: Reshaping the Vocabularies of LLMs with Vector Arithmetic*
Yuval Reif, **Guy Kaplan**, Roy Schwartz
Findings of the Association for Computational Linguistics (ACL Findings), 2026
`arxiv.org/abs/2510.17001`

Pre-prints

[1] *SFT-Induced Hallucinations as a Continual Learning Problem*
Guy Kaplan, Zorik Gekhman, Zhen Zhu, Yuval Reif, Lotem Rozner, Swabha Swayamdipta, Derek Hoiem, Roy Schwartz
Submitted to the Conference on Language Modeling (CoLM), 2026

[2] *More Than Words: Compositional Tokenization for Efficient Language Models*
Yuval Reif, **Guy Kaplan**, Roy Schwartz
Submitted to the Conference on Language Modeling (CoLM), 2026

Teaching & Mentoring

- Mentored Omer Ben Shahar (undergraduate, 2025) on training-data frequencies and the inner lexicon of LLMs; manuscript in preparation for ARR submission.
- Mentored Lotem Rozner (M.Sc.) on LLM personalization, extending the inner-lexicon research program; her work resulted in co-authorship on *SFT-Induced Hallucinations as a Continual Learning Problem*.

Invited Talks

"The Inner Lexicon of LLMs" Jan. 2026
University of Southern California (USC)

"The Inner Lexicon of LLMs" Dec. 2025
Tel Aviv University (TAU)

"The Inner Lexicon of LLMs" Jan. 2025
Microsoft Israel Data Science Group

Awards & Honors

Excellence Graduate Student Scholarship, KLA 2025
Dean's Honours in B.Sc. Computer Science, Open University of Israel 2021
Commander's Award, Palmachim AFB, Israeli Navy 2019
Magna Cum Laude, B.A. General Studies, University of Haifa 2014

Academic Service

Reviewer: ACL Rolling Review (ARR), Conference on Language Modeling (CoLM), ICML 2024 Workshop on Actionable Interpretability

Research Collaborations: Technion, UIUC, Google Research, ETH Zürich (official research visit), University of Southern California

Volunteering

Israeli Scouts — Bedouin Youth at Risk & Jewish–Arab Coexistence *Negev, Israel*
Year of Service Volunteer 2010 – 2011

Full-time educational & social engagement with at-risk Bedouin youth; promoted Jewish–Arab coexistence.

Skills

ML & NLP: Transformers, mechanistic interpretability, LLM fine-tuning (LoRA/QLoRA), contrastive learning

Programming: Python (PyTorch, TensorFlow), Java, Bash, LaTeX

Languages: Hebrew (native), English (fluent)