

# Guy Kaplan

PhD Student, The Hebrew University of Jerusalem | Visiting Researcher, University of Southern California

✉ [guy.kaplan3@mail.huji.ac.il](mailto:guy.kaplan3@mail.huji.ac.il) · 📞 +1 (626) 354-9602 · 🌐 [guykap12.github.io](https://github.com/guykap12)  
🌐 LinkedIn · 🎓 Google Scholar · 📖 Semantic Scholar · 🐦 @GKaplan38844

---

## Education

**The Hebrew University of Jerusalem (HUJI) & University of Southern California (USC)** Jerusalem & LA  
*PhD Student, Computer Science* Aug. 2025 – Present

Jointly advised by Prof. Roy Schwartz (HUJI) and Prof. Swabha Swayamdipta (USC).  
Research: LLM representations, tokenization, hallucinations, and 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, DILL Lab* (Prof. Swabha Swayamdipta) 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 – Oct. 2026

Developed an AI method for identification from dental records, deployed by the Israeli Police after October 7.

**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** Oct. 2011 – Sep. 2019  
*Naval Officer (Major)*

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](https://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](https://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](https://arxiv.org/abs/2510.17001)
- [4] *Why Fine-Tuning Encourages Hallucinations and How to Fix It*  
**Guy Kaplan**, Zorik Gekhman, Zhen Zhu, Yuval Reif, Lotem Rozner, Swabha Swayamdipta, Derek Hoiem, Roy Schwartz  
*Conference on Language Modeling (CoLM)*, 2026  
[arxiv.org/abs/2604.15574](https://arxiv.org/abs/2604.15574)
- [5] *More Than Words: Compositional Tokenization for Efficient Language Models*  
 Yuval Reif, **Guy Kaplan**, Roy Schwartz  
*Conference on Language Modeling (CoLM)*, 2026 **Oral Spotlight** (top <1% of submissions)  
[openreview.net/pdf?id=Yw16kddexd](https://openreview.net/pdf?id=Yw16kddexd)

## 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 *Why Fine-Tuning Encourages Hallucinations and How to Fix It*.

## Invited Talks

---

|                                                                               |           |
|-------------------------------------------------------------------------------|-----------|
| <b>"The Inner Lexicon of LLMs"</b><br>University of Southern California (USC) | Jan. 2026 |
| <b>"The Inner Lexicon of LLMs"</b><br>Tel Aviv University (TAU)               | Dec. 2025 |
| <b>"The Inner Lexicon of LLMs"</b><br>Microsoft Israel Data Science Group     | Jan. 2025 |

## Awards & Honors

---

|                                                                                                                                                                                                                                                                                                                                                                                                                 |      |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------|
| <b>PBC (VATAT) Fellowship for Outstanding Doctoral Students in AI &amp; Data Science</b><br>Council for Higher Education of Israel, under the National Initiative for Artificial Intelligence and Data Science. Israel's national doctoral fellowship in AI and data science: up to 20 awards countrywide per cycle, with each university permitted at most six nominees; three years of full doctoral funding. | 2026 |
| Oral Spotlight, CoLM 2026 (top <1% of submissions) — <i>More Than Words</i>                                                                                                                                                                                                                                                                                                                                     | 2026 |
| 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**

*Year of Service Volunteer*

*Negev, Israel*

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, L<sup>A</sup>T<sub>E</sub>X

**Languages:** Hebrew (native), English (fluent)