Yoav Gur-Arieh

MS Student and software engineer with 10 years of experience, researching interpretability in LLMs

⚲ Tel-Aviv, Israel   ✉ yoavgurarieh@gmail.com   🔗 yoav.ml   **in** yoavgurarieh   ○ yoavgur

Education

Tel-Aviv University *Oct 2024 to Jun 2025*
MS in Computer Science

- NLP lab, researching interpretability in LLMs under advisor Mor Geva.
- Done as part of the direct-to-masters interdisciplinary program.
- Average: 96.44

Tel-Aviv University *Oct 2022 to Jun 2026*
Interdisciplinary Studies

- Studying in the Adi Lautman Interdisciplinary Program for Outstanding Students.
- Took courses in computer science, biology, neuroscience, physics and history.
- Average: 93.71

Experience

Graduate Researcher (*Advisor: Mor Geva*) *Sep 2024 to present*
Tel-Aviv University

- Working on research in **interpretability and control** in LLMs
- Published research on improvements to large-scale automated **feature interpretability pipelines**, generating output-centric feature descriptions to improve understanding of features.
- Researching novel scalable **knowledge adaptation** and injection techniques for LLMs.
- Working on building large-scale data pipelines for precise **knowledge attribution for pre-training**.

Research Intern (*Supervisor: Atticus Geiger*) *Jun to Sep 2025*
Pr(Ai)^{2}R Group

Senior Software Engineer *Mar 2021 to present*
Laminar Security / Rubrik

- Joined the startup at its inception, spearheading the development of front-end, back-end, and agent-based systems from scratch, including a document sensitivity classification engine.
- Led technical research and implemented solutions for extracting decrypted data from encrypted cloud traffic.

Senior Developer & Researcher *Mar 2016 to Mar 2021*
Stealth

- Carried out months-long solo research projects into esoteric and opaque technologies that have little or no publicly available documentation, culminating in the development of technical solutions.
- Led through to completion highly complex, high-risk projects involving multiple teams.

Software Developer *Aug 2015 to Mar 2016*
Checkpoint Software Technologies

- Contributed to the development of a log aggregator and analyzer product tailored for large enterprises. Designed and implemented automated reports featuring visualization of insights.

Publications

LMEnt: A Suite for Analyzing Knowledge in Language Models from Pre-training Data to Representations *September 2025*

Daniela Gottesman, Alon Gilae-Dotan, Ido Cohen, **Yoav Gur-Arieh**, Marius Mosbach Ori Yoran, Mor Geva

Submitted to TACL - https://arxiv.org/abs/2509.03405

Precise In-Parameter Concept Erasure in Large Language Models *May 2025*

Yoav Gur-Arieh, Clara Suslik, Yihuai Hong, Fazl Barez, Mor Geva

Accepted to EMNLP Main 2025 - arxiv.org/abs/2505.22586

Enhancing Automated Interpretability with Output-Centric Feature Descriptions *Jan 2025*

Yoav Gur-Arieh, Roy Mayan, Chen Agassy, Atticus Geiger, Mor Geva

Accepted to ACL Main 2025 - arxiv.org/abs/2501.08319

Invited Talks

Max Planck Institute for Security and Privacy – *From Description to Erasure:* *Sep 2025*
Feature-Based Control of LLMs

Projects

SAE Knowledge Erasure Project *Code* ⬈

- Leveraged output-centric feature descriptions to identify MLP SAE features associated with specific concepts in the Gemma-2 2B model. Ablated these features to effectively erase the corresponding concepts, demonstrating targeted knowledge manipulation - explained in detail here.

Avian Neuronal Response Project *Code* ⬈

- Analyzed avian neuronal activity in response to varied bird calls (authentic and artificial), and used classical machine learning techniques to classify stimuli.

Parkalot - Parking Finder App *Link* ⬈

- Developed an app that displays parking lots with free spots around the user.
- Created the scraper (python), backend (python), database (SQL) and app (JS).

Hoppa - Platform Jumper Android Game *Link* ⬈

- Developed an android platform jumper game using Unity in C#, from conception to deployment.

Technologies

Languages: Python, C, Golang, C#, Java, JavaScript

Technologies: PyTorch, HF Transformers, TransformerLens, SAELens, eBPF, Linux Internals, Cybersecurity

Language Proficiency

English, Hebrew: Native