

Ofek Avivi

ofek@ofekavivi.com · ofekavivi.com · [GitHub](#) · [LinkedIn](#)

Researcher working on making large language models cheaper and faster to run, with 1.5 years of industry experience building computer-vision and full-stack software.

Research

Researcher · Technion and INSIGHT Lab, Ben-Gurion University

Jun 2026 – present

With Avi Mendelson and Chaim Baskin.

- Studied how compressing diffusion language models to low precision hurts their accuracy, and found the damage depends on the task and on the stage of generation.
- Built T²Quant, which gives each stage the precision it needs for each task. At the same memory budget, it improved three recent compression methods on every model tested, by up to 6.7 points.
- Released the code and an interactive project page.

Publications

Different Tasks, Different Timesteps: Temporal Task-Aware Quantization for Diffusion Language Models

Ofek Avivi*, Eliron Rahimi*, Avi Mendelson, Chaim Baskin

arXiv:2610.31198, 2026. *Equal contribution. [Paper](#) · [Project page](#) · [Code](#)

Experience

Junior Software Engineer · EyeContact DX

Aug 2024 – Jan 2026

- Developed applications in JavaScript, C++, C# and Python for an eye-tracking tool that offers an alternative autism (ASD) screener for toddlers aged 6 months to 3 years.
- Built computer-vision components with MediaPipe, including image processing and facial recognition, and increased facial-data processing speed by 50%.
- Implemented Python data pipelines for processing and real-time analysis.

Education

B.Sc. in Computer Science · The Open University of Israel

2020 – 2025

GPA 92.

Bar-Ilan Youth Advancement Program · Bar-Ilan University

2020 – 2022

Completed the 5-unit math Bagrut in May 2022 with a score of 100.

Skills

ML and research	PyTorch, Hugging Face (LLaDA, Dream), quantization, multi-GPU evaluation, reinforcement learning (DQN), MediaPipe, NumPy, Pandas, Matplotlib, eye tracking (EyeDid)
Languages	Python, C, C++, C#, JavaScript, Java, MIPS assembly
Web	React, Next.js, Tailwind, Node.js, Express, PostgreSQL, MongoDB
Tools	AWS (S3, EC2, Lambda, Amplify), Git, NGINX

Projects

Unigrade	Live GPA tracker and degree planner for university students (Next.js).
Connect 4 AI	PyGame game with minimax, alpha-beta and DQN agents.
Sorting Visualizer	React app that animates six sorting algorithms.
Two-pass assembler	C assembler for a made-up 16-instruction machine.

More

Languages	Hebrew (native), English (native; lived four years in the US and the UK), Spanish (conversational)
Volunteering	Beit Chabad: delivered food and groceries to elderly people and Holocaust survivors in my area, and kept them company.
Hobbies	Playing and watching football. Visca el Barça i visca Catalunya!