Ido Hakimi

(+972) 54-288-1136 | idohakimi@gmail.com | idohakimi

Academic Background

ETH Zürich
Zürich, Switzerland

Postdoctoral Researcher in Computer Science — *2024 - Present*

- **ETH AI Center**. Postdoctoral researcher at the ETH AI Center under the supervision of Professors Andreas Krause, Mrinmaya Sachan, and Ryan Cotterell. Research focuses on advancing LLM capabilities with strategic compute allocation during inference.
- **SwissAI**. Contributing to the SwissAI core LLM initiative, developing a large language model from scratch on a 10,000 GPU GH200 Nvidia cluster.

Technion - Israel Institute of Technology
Haifa, Israel

PhD (direct track) in Computer Science — *2017 - 2022*

- Research in distributed neural networks under the supervision of Prof. Assaf Schuster.
- Received a PhD scholarship from the Hasso-Plattner Institute.

Technion - Israel Institute of Technology
Haifa, Israel

BSc in Computer Science — *2013 - 2017*

- Won first place at the Technion's ACM ICPC Competition.
- Represented the Technion universoty at the international 2015 ACM ICPC SWERC coding competition.

Publications

2025	**From Problem-Solving to Teaching Problem-Solving: Aligning LLMs with Pedagogy using Reinforcement Learning** David Dinucu-Jianu, Jakub Macina, Nico Daheim, **Ido Hakimi**, Iryna Gurevych, Mrinmaya Sachan	*Under-Review*
2025	**Local mixtures of experts: Essentially free test-time training via model merging** Ryo Bertolissi, Jonas Hübotter, **Ido Hakimi**, Andreas Krause	*COLM*
2025	**Mathtutorbench: A benchmark for measuring open-ended pedagogical capabilities of llm tutors** Jakub Macina, Nico Daheim, **Ido Hakimi**, Manu Kapur, Iryna Gurevych, Mrinmaya Sachan	*Under-Review*
2025	**Efficiently Learning at Test-Time: Active Fine-Tuning of LLMs** Jonas Hübotter, Sascha Bongni, **Ido Hakimi**, Andreas Krause	*ICLR*
2023	**q2d: Turning Questions into Dialogs to Teach Models How to Search** Yonatan Bitton, Shlomi Cohen-Ganor, **Ido Hakimi**, Yoad Lewenberg, Roee Aharoni, Enav Weinreb	*EMNLP*
2022	**SMEGA^2: Distributed Asynchronous Deep Neural Network Training With a Single Momentum Buffer** Rafi Cohen*, **Ido Hakimi***, Assaf Schuster	*ICPP*
2021	**Faster neural network training with approximate tensor operations** Menachem Adelman, Kfir Yehuda Levy, **Ido Hakimi**, Mark Silberstein	*NeurIPS*
2021	**Asynchronous Distributed Learning: Adapting to Gradient Delays without Prior Knowledge** Rotem Zamir Aviv, **Ido Hakimi**, Assaf Schuster, Kfir Yehuda Levy	*ICML*
2021	**Fine-tuning giant neural networks on commodity hardware with automatic pipeline model parallelism** Saar Eliad, **Ido Hakimi**, Alon De Jagger, Mark Silberstein, Assaf Schuster	*USENIX-ATC*
2021	**LAGA: Lagged AllReduce with Gradient Accumulation for Minimal Idle Time** **Ido Hakimi**, Rotem Zamir Aviv, Kfir Yehuda Levy, Assaf Schuster	*ICDM*
2020	**Gap-Aware Mitigation of Gradient Staleness** Saar Barkai*, **Ido Hakimi***, Assaf Schuster	*ICLR*
2019	**Taming momentum in a distributed asynchronous environment** **Ido Hakimi***, Saar Barkai*, Moshe Gabel, Assaf Schuster	*arXiv*

* equal contribution.

Program Committees

KDD · 2021, 2022, 2023, 2024 **ICLR** · 2022[1], 2023,2024 **NeurIPS** · 2022[1], 2023,2024,2025

ICML · 2023, 2024 **IJCAI** · 2024 **AAAI** · 2024 **COLM** · 2024,2025 **ACL** · 2024,2025

[1] top reviewer

Work Experience

Google DeepMind
Tel-Aviv, Israel

Research Engineer — *2024 - 2024*

- **LLM Planning Complex Task Decomposition**. Developed an algorithm for handling complex task that require iterative planning with multiple information retrievals. For example, "What are the top ten US based oil comapnies by net value? List their ticker symbol, net worth, stock price, and CEO name." This query requires breaking down into smaller tasks, which are then run in parallel to retrieve the required information.
- **LLM Creative Writing**. We found that LLMs perform poorly in creative writing (based on drama writing experts), when the LLM is not prompted properly. For example, writing short drama stories or plays have certain rules that are typically followed such as "show don't tell". Prompting the LLM to first research about the subject and allow itself to learn about what to emphasize greatly improves the writing results.
- Served as a reviewer for NLP grant proposals

Google Research
Tel-Aviv, Israel

Research Engineer — *2022 - 2024*

- **LLM In-Context Hallucination Reduction**. Developed a hallucination detection model that utilized a novel iterative adversarial synthetic data generation method to identify in-context hallucinations more effectively.
- **Bard Powered Personal Assistant (Page-1)**. Developed a new technique for LLM user property retrieval. Furthermore, we developed a new technique for conversational bot personalization using prompt-tuning as compressed memory.
- **LaMDA Query Generation**. Developed a new technique for conversational query generation using synthetically generated data by large language models. As part of the research, we wrote the q2d academic paper.
- Volunteered to mentor AI interns from under-represented countries.

Toga Networks
Hod Hasharon, Israel

Research Scientist — *2021 - 2022*

- Developed new algorithms for large scale distributed neural network training with efficient communications.
- Researched in large batch distributed settings for fast convergence and high final model quality.

Amazon - Alexa Shopping Research
Haifa, Israel

Applied Scientist Intern — *2020*

- Researched on conservative reformulation of user utterances to improve question answering.

Facebook - Data AI
Tel-Aviv, Israel

Machine Learning Intern — *2019*

- Developed and deployed a personalized data artifact retriever using user activity usage.

Technion - Israel Institute of Technology
Haifa, Israel

AI Project Instructor — *2018 - 2020*

- Mentored various AI based projects, some of which later became research paper.

Apple - Turi
Seattle, WA

Deep Learning Research Intern — *2018*

- Research on neural network ensemble technique which utilizes parallel training of neural networks without any communications during training.

Technion - Israel Institute of Technology
Haifa, Israel

Head Teaching Assistant — *2017 - 2019*

- Served as Head TA for the "Concurrent and Distributed Programming for Data processing and Machine Learning" course (236370).
- Transformed the course to be more data-oriented for ML.
- Responsibilities included writing and grading exams, as well as writing the homework assignments.

Dell EMC - ScaleIO
Haifa, Israel

Software Engineer Student — *2015 - 2017*

- Wrote low-level high-performance C code for ScaleIO software-defined storage solution.
- Developed a new efficient algorithm to load-balance server storage blocks for high performance and scalability.
- Received an acknowledgment of excellence.