Yftah Ziser

Assistant Professor · University of Groningen · Senior Research Scientist · Nvidia Research

Bernoulli Institute, University of Groningen, Nijenborgh 9, 9747 AG Groningen, The Netherlands

yftah89@gmail.com | https://yftah89.github.io/

Professional Experience

University of Groningen — Groningen, the Netherlands
Assistant Professor — 2025 - present

- Tenure-track assistant professor (research focus) working mainly on large language model alignment, reasoning, and low-resource language settings.
- Supervising MSc/PhD students and mentoring early-career researchers.

Nvidia Research — Groningen, the Netherlands (remotely)
Senior Researcher — 2024 - present

- Leading research on alignment, safety, and factuality of large language models.
- Driving projects from problem formulation to publication and product-facing impact.
- Leading and mentoring multiple research interns across alignment and NLP projects.
- Collaborating with product and applied research teams to translate research into deployable systems.

University of Edinburgh — Edinburgh, United Kingdom
Postdoctoral Researcher — 2021 - 2024

- Hosted by Dr. Shay Cohen.
- Conducted independent and collaborative research in NLP.
- Published in top-tier NLP and ML conferences and journals.
- Contributed to grant proposals and supervision of graduate students.

Meta — Tel-Aviv, Israel
Researcher — 2020 - 2021

- Applied deep learning and representation learning methods to real-world user behavior data at scale.
- Developed high-quality embedding representations for recommendation pipelines, improving signal quality for user interaction and relevance tasks.
- Collaborated with engineering teams to integrate research into production-facing recommendation systems.

Amazon Alexa — Haifa, Israel
Researcher — 2019 - 2020

- Designed and evaluated large-scale question answering systems using community data to improve Alexa responses.
- Led and mentored multiple research interns.

IBM — Haifa, Israel
Researcher — 2017 - 2018

- Designed and built machine learning algorithms for natural language problems.
- Designed and taught an internal ”Introduction to Deep Learning” course for IBM colleagues.

Publications

From Actions to Words: Towards Abstractive-Textual Policy Summarization in RL, Sahar Admoni, Assaf Hallak, **Yftah Ziser**, Omer Ben-Porat, Ofra Amir. To appear in AAMAS 2026.

Beyond Next Token Probabilities: Learnable, Fast Detection of Hallucinations and Data Contamination on LLM Output Distributions, Guy Bar-Shalom, Fabrizio Frasca, Derek Lim, Yoav Gelberg, **Yftah Ziser**, Ran El-Yaniv, Gal Chechik, Haggai Maron. To appear in AAAI 2026.

Beyond Token Probes: Hallucination Detection via Activation Tensors with ACT-ViT, Guy Bar-Shalom, Fabrizio Frasca, Yaniv Galron, **Yftah Ziser** and Haggai Maron. NeurIPS 2025.

Policy Optimized Text-to-Image Pipeline Design, Uri Gadot, Rinon Gal, **Yftah Ziser**, Gal Chechik, Shie Mannor, NeurIPS 2025.

Iterative Multilingual Spectral Attribute Erasure, Shun Shao, **Yftah Ziser**, Zheng Zhao, Yifu QIU, Shay B Cohen, Anna Korhonen, EMNLP 2025.

A Simple Yet Effective Method for Non-Refusing Context Relevant Fine-grained Safety Steering in LLMs, Shaona Ghosh, Amrita Bhattacharjee, **Yftah Ziser**, Christopher Parisien, EMNLP 2025.

Knowing Before Saying: LLM Representations Encode Information About Chain-of-Thought Success Before Completion, Anum Afzal, Florian Matthes, Gal Chechik and **Yftah Ziser**, Findings of ACL 2025.

TSPRank: Bridging Pairwise and Listwise Methods with a Bilinear Travelling Salesman Model, Waylon Li, **Yftah Ziser**, Yifei Xie, Shay B Cohen and Tiejun Ma, KDD 2025 (Research Track).

Spectral Editing of Activations for Large Language Model Alignment, Yifu Qiu, Zheng Zhao, **Yftah Ziser**, Anna Korhonen, Edoardo Ponti and Shay Cohen, NeurIPS 2024.

Layer by Layer: Uncovering Where Multi-Task Learning Happens in Instruction-Tuned Large Language Models, Zheng Zhao, **Yftah Ziser** and Shay Cohen, EMNLP 2024.

Are Large Language Models Temporally Grounded?, Yifu Qiu, Zheng Zhao, **Yftah Ziser**, Anna Korhonen, Edoardo Ponti and Shay Cohen, NAACL 2024.

Detecting and Mitigating Hallucinations in Multilingual Summarisation, Yifu Qiu, **Yftah Ziser**, Anna Korhonen, Edoardo Ponti and Shay Cohen, EMNLP 2023.

A Joint Matrix Factorization Analysis of Multilingual Representations, Zheng Zhao, **Yftah Ziser**, Bonnie Webber and Shay Cohen, findings of EMNLP 2023.

Rant or Rave: Variation over Time in the Language of Online Reviews, **Yftah Ziser**, Bonnie Webber and Shay Cohen, LRE 2023.

Erasure of Unaligned Attributes from Neural Representations, Shun Shao*, **Yftah Ziser***, and Shay Cohen, TACL 2023.

BERT Is Not The Count: Learning to Match Mathematical Statements with Proofs, Waylon Li, **Yftah Ziser**, Maximin Coavoux, and Shay Cohen, EACL 2023.

Gold Doesn't Always Glitter: Spectral Removal of Linear and Nonlinear Guarded Attribute Information, Shun Shao, **Yftah Ziser**, and Shay Cohen, EACL 2023.

Factorizing Content and Budget Decisions in Abstractive Summarization of Long Documents, Marcio Fonseca, **Yftah Ziser**, and Shay Cohen, EMNLP 2022.

Understanding Domain Learning in Language Models Through Subpopulation Analysis, Zheng Zhao, **Yftah Ziser**, and Shay Cohen, BlackBoxNLP@EMNLP 2022.

Customized Pre-Training for Improved Domain Adaptation with Category Shift, Tony Lechtman, **Yftah Ziser**, and Roi Reichart, EMNLP 2021.

WikiSum: Coherent Summarization Dataset for Efficient Human-Evaluation, Nachshon Cohen*, Oren Kalinsky*, **Yftah Ziser***, and Alessandro Moschitti, ACL 2021.

Answering Product-Questions by Utilizing Questions from Other Contextually Similar Products, Ohad Rozen, David Carmel, Avihai Mejer, Vitali Mirkis, and **Yftah Ziser**, NAACL 2021.

Humor Detection in Product Question Answering Systems, **Yftah Ziser**, Elad Kravi, and David Carmel, SIGIR 2020.

Task Refinement Learning for Improved Accuracy and Stability of Unsupervised Domain Adaptation, **Yftah Ziser**, and Roi Reichart, ACL 2019.

Deep Pivot-Based Modeling for Cross-language Cross-domain Transfer with Minimal Guidance, **Yftah Ziser**, and Roi Reichart, EMNLP 2018.

Pivot Based Language Modeling for Improved Neural Domain Adaptation, **Yftah Ziser** and Roi Reichart, NAACL 2018.

Neural Structural Correspondence Learning for Domain Adaptation, **Yftah Ziser** and Roi Reichart, CoNLL 2017.

Education

Technion – Israel Institute of Technology — Haifa, Israel
Ph.D. (Direct track) — 2015 - 2019

- Supervised by Prof. Roi Reichart.
- Thesis: "Domain Adaptation for Natural Language Processing - a Neural Network Based Approach."
- My Ph.D. focused on adapting models from domains and languages rich in labeled training data to domains and languages that are poor in such data through representation learning.
- Outstanding Ph.D. Student Award, the DDS faculty.

Technion – Israel Institute of Technology — Haifa, Israel
B.Sc. — 2011 - 2015

- Bachelor of Science in Computer Science

Mentoring

Mentored Ph.D. and M.Sc. students across academia and industry, including long-term academic supervision and industry research internships, resulting in publications at top-tier venues and successful academic and industrial placements.

Ph.D. Students

- Shun Shao (M.Sc., Edinburgh → Ph.D., Cambridge; academic collaboration with NVIDIA)
- Guy Bar-Shalom (Technion; academic collaboration with NVIDIA)
- Sahar Admoni (Technion; academic collaboration with NVIDIA)
- Yifu Qiu (University of Edinburgh)
- Zheng Zhao (University of Edinburgh)
- Marcio Fonseca (University of Edinburgh)

M.Sc. Students

- Waylon Li (University of Edinburgh)
- Entony Lekhtman (Technion)

Industry Research Interns (Ph.D.)

- Uri Gadot (Technion; NVIDIA)
- Ohad Rozen (Bar-Ilan University; Amazon)
- Eilon Sheetrit (Technion; Amazon)
- Ido Hakimi (Technion; Amazon)

Teaching Experience

Fall 2025	Natural Language Processing with Deep Learning,	Lecturer in Charge.
2019	Yearly Project in Data Science,	Project Mentor.
Fall 2018	Natural Language Processing for the Social Sciences,	Project Mentor.
Spring 2017	Learning Robust Representations for Natural Language Processing,	Project Mentor.
Fall 2017	Introduction to Natural Language Processing,	Teaching Assistant in Charge.
2015-2016	Introduction to Computer Science,	Teaching Assistant.

Academic Service

Invited Talks

2025 *Knowing Before Saying: LLM Representations Encode Information About Chain-of-Thought Success Before Completion*. Nvidia Research, "System 2" seminar, Santa Clara, USA.

2023 *Navigating Distribution Shifts in Natural Language Processing*. IBM Research, Haifa, Israel.

2023 *Navigating Distribution Shifts in Natural Language Processing*. Gong Research Center, Tel-Aviv, Israel.

2022 *Guarded Information Removal for Better Debiasing*. Language Technology Lab group, University of Cambridge, Cambridge, United Kingdom.

2022 *Guarded Information Removal for Better Debiasing*. Department for Computational Linguistics, Heidelberg University, Heidelberg, Germany.

2020 *Domain Bias in Humor Detection*. Amazon Research Center, Tel-Aviv, Israel.

2018 *Cross-language Cross-domain Transfer with Minimal Guidance*. Israel Seminar on Computational Linguistics, Haifa, Israel.

2018 *Contextualized Representations for Domain Adaptation*. Computer Science department, Tel-Aviv University, Tel-Aviv, Israel.

2017 *Harnessing the Power of Neural Networks for Domain Adaptation*. Technion Computational Data Science seminar, Haifa, Israel.

Organization

2021 Co-Organizer of the *Domain Adaptation for NLP Workshop*. At EACL, virtual event.

2018 Committee member of the *Seminar of Computational Linguistics (ISCOL)*. Haifa, Israel.

Reviewing

2018-2021 Reviewing for main NLP venues, including NAACL, ACL, EMNLP, EACL and CoNLL.

2023-present Reviewing for TACL.

2023 Area Chair, Efficient/Low-resource Methods in NLP, EACL 2024.

2021-2024 Reviewing for ACL Rolling Review (ARR).

2024–present Action Editor for ACL Rolling Review (ARR).

Ph.D. Admissions

2021-present Evaluating candidates for the European Laboratory for Learning and Intelligent Systems (ELLIS) Ph.D. program.

Community Engagement

Yotzeem Leshinuy
Jerusalem, Israel

Lecturer
2020 - 2021

- Designing and instructing a friendly programming course for former members of the Haredi Jewish community, who are among the most disadvantaged groups in Israel.

Educating for Excellence
Acre , Israel

Tutor
2019 - 2020

- Teaching underprivileged children programming and math.