SHAULI RAVFOGEL

(+972) 0549508315 ⋄ shauli.ravfogel@gmail.com ⋄ shauli-ravfogel.netlify.app

Faculty Fellow, New York University Centr of Data Science. My research interests encompass representation learning, unsupervised machine learning, causal attribution, and interpretability of neural models for NLP.

EDUCATION

- **2024-** Faculty Fellow, NYU
- **2020-2024** PhD student in Computer Science, Bar-Ilan University.
  Dissertation title: Analyzing the representation space of transformer-based language models.
  Supervisor: Prof. Yoav Goldberg.
- **2018-2020:** MSc in Computer Science, Bar-Ilan University, supervised by Prof. Yoav Goldberg.
- **2015-2018:** BSc in Computer Science, Bar-Ilan University.
- **2010-2013:** BSc in Chemistry, Bar-Ilan University.

PROFESSIONAL EXPERIENCE AND APPOINTMENTS

- **September-December 2024**: Research visit, ETH Zurich.
- **October 2023-September 2024** Student Researcher, Google Research
- **June-September 2023:** Research internship, Bloomberg London.
- **June-September 2022:** Research internship, Facebook AI Research, Israel.
- **January 2020-summer 2022 & Autumn 2022-summer 2023:** Ph.D. Research Intern, AI2 Israel.

TEACHING EXPERIENCE

- **Lecturer:** Causality and Interpretability of Language Models, NYU (Winter 2025)
- **Lecturer:** Introduction to Machine Learning, Bar-Ilan University (2020–2021)
- **Teaching Assistant:** Graduate Machine Learning Course, Bar-Ilan University (2019–2021)

PUBLICATIONS

Under submission

Yihuai Hong, Lei Yu, **Shauli Ravfogel**, Haiqin Yang, Mor Geva. "Intrinsic Evaluation of Unlearning Using Parametric Knowledge Traces".

Royi Rassin, Aviv Slobodkin, **Shauli Ravfogel**, Yanai Elazar, Yoav Goldberg. GRADE: Quantifying Sample Diversity in Text-to-Image Models.

Accepted

First/Last Author:

- **Shauli Ravfogel***, Anej Svete*, Vésteinn Snæbjarnarson, Ryan Cotterell. Gumbel Counterfactual Generation from Language Models. Accepted in ICLR 2025.

- Matan Avitan, Ryan Cotterell, Yoav Goldberg, and **Shauli Ravfogel**. "Natural Language Counterfactuals through Representation Surgery". Accepted in Findings of NAACL 2025.
- **Shauli Ravfogel**, Valentina Pyatkin, Amir David Nissan Cohen, Avshalom Manevich, Yoav Goldberg. "Description-based Text Similarity.". COLM 2024.
- Shashwat Singh*, **Shauli Ravfogel***, Roee Aharoni, Jonathan Herzig, and Ponnurangam Kumaraguru. "Representation Surgery: Theory and Practice of Affine Steering". Forty-first International Conference on Machine Learning.
- Aviv Slobodkin, Omer Goldman, Avi Caciularu, Ido Dagan, **Shauli Ravfogel**. "The Curious Case of Hallucinatory Unanswerability: Finding Truths in the Hidden States of Over-Confident Large Language Models." Accepted in EMNLP 2023.
- **Shauli Ravfogel**, Yoav Goldberg, and Ryan Cotterell. "Log-linear Guardedness and its Implications." In Proceedings of ACL 2023.
- **Shauli Ravfogel**, Yoav Goldberg, and Jacob Goldberger. "Conformal Nucleus Sampling." Findings of ACL 2023.
- **Shauli Ravfogel**, Michael Twiton, Yoav Goldberg, and Ryan D. Cotterell. "Linear adversarial concept erasure." In International Conference on Machine Learning 2022.
- **Shauli Ravfogel**, Francisco Vargas, Yoav Goldberg and Ryan Cotterell. "Kernelized Concept Erasure." In Proceedings of EMNLP 2022.
- Royi Rassin, **Shauli Ravfogel***, and Yoav Goldberg. "DALLE-2 is seeing double: Flaws in word-to-concept mapping in Text2Image models." In Proceedings of BlackBoxNL 2022.
- **Shauli Ravfogel**, Grusha Prasad, Tal Linzen, and Yoav Goldberg. "Counterfactual interventions reveal the causal effect of relative clause representations on agreement prediction." In Proceedings of CONLL 2021.
- **Shauli Ravfogel**, Yanai Elazar, Jacob Goldberger, and Yoav Goldberg. "Unsupervised Distillation of Syntactic Information from Contextualized Word Representations." In Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP 2020.
- **Shauli Ravfogel**, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. "Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection." In Proceedings of ACL 2020.
- **Shauli Ravfogel**, Yoav Goldberg, and Tal Linzen. "Studying the Inductive Biases of RNNs with Synthetic Variations of Natural Languages." In Proceedings of the NACCL 2019.
- **Shauli Ravfogel**, Yoav Goldberg, and Francis Tyers. "Can LSTM Learn to Capture Agreement? The Case of Basque." In Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP.
- Carlo Meloni*, **Shauli Ravfogel***, and Yoav Goldberg. "Ab Antiquo: Neural Proto-language Reconstruction." In Proceedings of NAACL 2021.

Co-author:

- Royi Rassin, Eran Hirsch, Daniel Glickman, **Shauli Ravfogel**, Yoav Goldberg, and Gal Chechik. "Linguistic Binding in Diffusion Models: Enhancing Attribute Correspondence through Attention Map Alignment." Accepted in Neurips 2023.
- Nora Belrose, David Schneider-Joseph, **Shauli Ravfogel**, Ryan Cotterell, Edward Raff, and Stella Biderman. "LEACE: Perfect linear concept erasure in closed form." Accepted in Neurips 2023.
- Mosh Levy, **Shauli Ravfogel**, Yoav Goldberg. "Guiding LLM to Fool Itself: Automatically Imbuing Multiple Shortcut Triggers in Question Answering Samples." Findings of EMNLP 2023.

- Marius Mosbach, Tiago Pimentel, **Shauli Ravfogel**, Dietrich Klakow, and Yanai Elazar. "Few-shot Fine-tuning vs. In-context Learning: A Fair Comparison and Evaluation." Proceedings of Repl4NLP 2024 workshop.
- Rita Sevastjanova, Eren Cakmak, **Shauli Ravfogel**, Ryan Cotterell, and Mennatallah El-Assady. "Visual comparison of language model adaptation." IEEE Transactions on Visualization and Computer Graphics 2022.
- Hila Gonen, **Shauli Ravfogel**, and Yoav Goldberg. "Analyzing Gender Representation in Multilingual Models." Proceedings of the 7th Workshop on Representation Learning for NLP 2022 (Best Paper Award).
- Yanai Elazar, Nora Kassner, **Shauli Ravfogel**, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. "Measuring and improving consistency in pretrained language models." Transactions of the Association for Computational Linguistics 2021.
- Elad Ben-Zaken, **Shauli Ravfogel**, and Yoav Goldberg. "Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models." In Proceedings of ACL 2022.
- Yanai Elazar, **Shauli Ravfogel**, Alon Jacovi, and Yoav Goldberg. "Amnesic probing: Behavioral explanation with amnesic counterfactuals." Transactions of the Association for Computational Linguistics 2021.
- Alon Jacovi, Swabha Swayamdipta, **Shauli Ravfogel**, Yanai Elazar, Yejin Choi, and Yoav Goldberg. "Contrastive Explanations for Model Interpretability." In Proceedings of EMNLP 2021.
- Hila Gonen, **Shauli Ravfogel**, Yanai Elazar, and Yoav Goldberg. "It's not Greek to mBERT: Inducing Word-Level Translations from Multilingual BERT." In Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP 2020.

AWARDS

- IAAI Best PhD Thesis Award.
- Fulbright Postdoctoral Fellowship (2024; I declined after accepting an offer from the NYU CDS Faculty Fellowship).
- Blavatnik Prize for Outstanding Israeli Doctoral Students in Computer Science (2024)
- DAAD AInet Postdoctoral Fellowship (2023).
- Bloomberg Data Science PhD Fellowship (2022-present; first recipient outside of the US).
- Israeli Council for Higher Education (Vatat) Fellowship for outstanding PhD Students in Data Science (declined due to a conflict).
- Outstanding Paper Award, Bar-Ilan Data Science Institute (2023).
- Bar-Ilan Presidential Scholarship for Outstanding PhD Students (2021-2024).
- Best paper award, RepL4NLP workshop (2022)

ACADEMIC ACTIVITY

- **Academic service:**
  - Co-organized the Representation Learning for NLP (Repl4NLP) 2023 workshop, a leading workshop within the Association for Computational Linguistics (ACL), with over 250 attendees.
  - Co-advised 2 master's students.

  - Reviewing for ACL, EMNLP, NACCL, Neurips, and ICML. Area chair in ACL Rolling Review (ARR).

- **Invited talks:**
  - Technion (Winter 2024)
  - EPFL, Switzerland (Autumn 2023).
  - DeepMind and JP Morgan London (summer 2023).
  - IBM Israel and the Hebrew University NLP group (2023).
  - SIGTYP Lecture Series (2021).
  - Prof. Roi Reichart's group, Technion (2020).
  - Prof. Robert Frank's group, Yale (2020).
- **Research visits (leading to sustained academic collaboration):**
  - 2020: Prof. Tal Linzen group (Johns Hopkins University).
  - 2021-2023 (three different stays): Prof. Ryan Cotterell's group, ETH Zurich.