SHAULI RAVFOGEL

(+1) 6463358177 shauli.ravfogel@gmail.com shauli-ravfogel.netlify.app

CURRENT POSITION

Faculty Fellow, New York University Center for Data Science 2025–present

RESEARCH INTERESTS

Representation learning for NLP; interpretability; unsupervised learning; causal analysis and attribution in neural models.

AWARDS

- IAAI Best PhD Thesis Award
- Fulbright Postdoctoral Fellowship (2024; declined after accepting 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 the U.S.)
- Israeli Council for Higher Education (Vatat) Fellowship for Outstanding PhD Students in Data Science (declined due to 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)

EDUCATION

- **Ph.D. in Computer Science**, Bar-Ilan University 2020–2024
  Dissertation: Analyzing the representation space of transformer-based language models
  Supervisor: Prof. Yoav Goldberg
- **M.Sc. in Computer Science**, Bar-Ilan University 2018–2020
  Supervisor: Prof. Yoav Goldberg
- **B.Sc. in Computer Science**, Bar-Ilan University 2015–2018
- **B.Sc. in Chemistry**, Bar-Ilan University 2010–2013

PROFESSIONAL EXPERIENCE AND APPOINTMENTS

- **Faculty Fellow**, NYU Center for Data Science 2025–present
- **Research Visitor**, ETH Zürich Sep 2024–Dec 2024
- **Student Researcher**, Google Research Oct 2023–Sep 2024
- **Research Intern**, Bloomberg (London) Jun 2023–Sep 2023
- **Research Intern**, FAIR (Meta AI), Israel Jun 2022–Sep 2022
- **Ph.D. Research Intern**, Allen Institute for AI (AI2) Israel Jan 2020–Sep 2023

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, Bar-Ilan University — 2019–2021

PUBLICATIONS

[*] *Equal contribution.*

Preprints / Under Review

- **Shauli Ravfogel**, Gilad Yehudai, Tal Linzen, Joan Bruna, Alberto Bietti. *Emergence of Linear Truth Encodings in Language Models.*
- Yu Fan, Yang Tian, **Shauli Ravfogel**, Mrinmaya Sachan, Elliott Ash, Alexander Hoyle. *The Medium Is Not the Message: Deconfounding Text Embeddings via Linear Concept Erasure.*
- Aviya Maimon, Amir DN Cohen, Gal Vishne, **Shauli Ravfogel**, Reut Tsarfaty. *IQ Test for LLMs: An Evaluation Framework for Uncovering Core Skills in LLMs.*
- Floris Holstege[*], **Shauli Ravfogel**[*], Bram Wouters. *Preserving Task-Relevant Information Under Linear Concept Removal.*
- Nhi Nguyen, **Shauli Ravfogel**, Rajesh Ranganath. *Do LLMs Lie About What They Use? Benchmark for Metacognitive Truthfulness in Large Language Models.*
- Jackson Petty, Michael Y. Hu, Wentao Wang, **Shauli Ravfogel**, William Merrill, Tal Linzen. *RELIC: Evaluating Compositional Instruction Following via Language Recognition.*
- Yihuai Hong, Lei Yu, Haiqin Yang, **Shauli Ravfogel**, 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.*

Peer-Reviewed (First/Last Author)

- **Shauli Ravfogel**[*], Anej Svete[*], Vésteinn Snæbjarnarson, Ryan Cotterell. *Gumbel Counterfactual Generation from Language Models.* ICLR 2025
- Matan Avitan, Ryan Cotterell, Yoav Goldberg, **Shauli Ravfogel**. *Natural Language Counterfactuals through Representation Surgery.* 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, Ponnurangam Kumaraguru. *Representation Surgery: Theory and Practice of Affine Steering.* ICML 2024
- 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.* EMNLP 2023
- **Shauli Ravfogel**, Yoav Goldberg, Ryan Cotterell. *Log-linear Guardedness and its Implications.* ACL 2023
- **Shauli Ravfogel**, Yoav Goldberg, Jacob Goldberger. *Conformal Nucleus Sampling.* Findings of ACL 2023

- **Shauli Ravfogel**, Michael Twiton, Yoav Goldberg, Ryan D. Cotterell. *Linear Adversarial Concept Erasure*. ICML 2022
- **Shauli Ravfogel**, Francisco Vargas, Yoav Goldberg, Ryan Cotterell. *Kernelized Concept Erasure*. EMNLP 2022
- Royi Rassin, **Shauli Ravfogel**[*], Yoav Goldberg. *DALLE-2 is Seeing Double: Flaws in Word-to-Concept Mapping in Text-to-Image Models*. BlackboxNLP 2022
- **Shauli Ravfogel**, Grusha Prasad, Tal Linzen, Yoav Goldberg. *Counterfactual Interventions Reveal the Causal Effect of Relative Clause Representations on Agreement Prediction*. CoNLL 2021
- **Shauli Ravfogel**, Yanai Elazar, Jacob Goldberger, Yoav Goldberg. *Unsupervised Distillation of Syntactic Information from Contextualized Word Representations*. BlackboxNLP 2020
- **Shauli Ravfogel**, Yanai Elazar, Hila Gonen, Michael Twiton, Yoav Goldberg. *Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection*. ACL 2020
- **Shauli Ravfogel**, Yoav Goldberg, Tal Linzen. *Studying the Inductive Biases of RNNs with Synthetic Variations of Natural Languages*. NAACL 2019
- **Shauli Ravfogel**, Yoav Goldberg, Francis Tyers. *Can LSTM Learn to Capture Agreement? The Case of Basque*. BlackboxNLP 2018
- Carlo Meloni[*], **Shauli Ravfogel**[*], Yoav Goldberg. *Ab Antiquo: Neural Proto-language Reconstruction*. NAACL 2021

Peer-Reviewed (Co-author)

- Royi Rassin, Eran Hirsch, Daniel Glickman, **Shauli Ravfogel**, Yoav Goldberg, Gal Chechik. *Linguistic Binding in Diffusion Models: Enhancing Attribute Correspondence through Attention Map Alignment*. NeurIPS 2023
- Nora Belrose, David Schneider-Joseph, **Shauli Ravfogel**, Ryan Cotterell, Edward Raff, Stella Biderman. *LEACE: Perfect Linear Concept Erasure in Closed Form*. 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, Yanai Elazar. *Few-shot Fine-tuning vs. In-context Learning: A Fair Comparison and Evaluation*. RepL4NLP 2024 Workshop
- Rita Sevastjanova, Eren Cakmak, **Shauli Ravfogel**, Ryan Cotterell, Mennatallah El-Assady. *Visual Comparison of Language Model Adaptation*. IEEE TVCG 2022
- Hila Gonen, **Shauli Ravfogel**, Yoav Goldberg. *Analyzing Gender Representation in Multilingual Models*. RepL4NLP 2022 (Best Paper)
- Yanai Elazar, Nora Kassner, **Shauli Ravfogel**, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, Yoav Goldberg. *Measuring and Improving Consistency in Pretrained Language Models*. TACL 2021
- Elad Ben-Zaken, **Shauli Ravfogel**, Yoav Goldberg. *BitFit: Simple Parameter-Efficient Fine-Tuning for Transformer-based Masked Language Models*. ACL 2022 (Short)
- Yanai Elazar, **Shauli Ravfogel**, Alon Jacovi, Yoav Goldberg. *Amnesic Probing: Behavioral Explanation with Amnesic Counterfactuals*. TACL 2021

- Alon Jacovi, Swabha Swayamdipta, **Shauli Ravfogel**, Yanai Elazar, Yejin Choi, Yoav Goldberg. *Contrastive Explanations for Model Interpretability.* EMNLP 2021
- Hila Gonen, **Shauli Ravfogel**, Yanai Elazar, Yoav Goldberg. *It's Not Greek to mBERT: Inducing Word-Level Translations from Multilingual BERT.* BlackboxNLP 2020

SERVICE & COMMUNITY

- **Area Chair**: ACL Rolling Review (ARR)
- **Reviewer**: ACL, EMNLP, NAACL, NeurIPS, ICML
- **Workshops**: Co-organizer, RepL4NLP 2023 at ACL (250+ attendees)
- **Grant reviewing**: U.S.–Israel Binational Science Foundation (BSF)
- **Advising**: Co-advised master's students

INVITED TALKS

- Technion — Winter 2024
- EPFL (Switzerland) — Autumn 2023
- DeepMind; J.P. Morgan (London) — Summer 2023
- IBM Israel; 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 COLLABORATION)

- Prof. Tal Linzen's group, Johns Hopkins University — 2020
- Prof. Ryan Cotterell's group, ETH Zürich — 2021–2023 (three visits)