AMIR ZUR

amirzur@cs.stanford.edu

EDUCATION

Stanford University Stanford, CA

M.S., Computer Science June 2023

Advisor: Dr. Omer Reingold
GPA 4.00

Stanford University Stanford, CA

B.A. with Honors, Linguistics, Minor in Education June 2023

Undergraduate Thesis: *Causal Abstraction for Interpretable and Debiased Language Models*
Advisor: Dr. Chris Potts
GPA 4.00

RESEARCH EXPERIENCE

Pr(Ai)^{2}R Group Stanford, CA

Research Intern Fall 2023 – Now

- Investigating representation bias in short story generation and how language models represent narrative structure.
- Collaborated on generalizing the theoretical framework of causal abstraction, and on unifying mechanistic interpretability methods such as path patching and causal mediation analysis under causal abstraction.

Stanford NLP Group Stanford, CA

Assistant Researcher Spring 2022 – Spring 2023

- Applied causal analysis to interpret and de-bias large language models, resulted in an award-winning undergraduate thesis.
- Induced interpretable causal structure in CLIP for more accessible image descriptions, resulted in a submission to ACL Rolling Review.
- Collaborated on benchmarking causal explanations of neural networks, resulted in an ICML publication.

Stanford Theory Group Stanford, CA

Assistant Researcher Summer 2021– Spring 2023

- Investigated definitions and guarantees of fairness in prediction and classification tasks.
- Co-developed a lexicon on algorithmic fairness (http://wiki-loaf.org/), connecting the values and perspectives of different disciplines within algorithmic fairness.

WORK EXPERIENCE

Microsoft, Applied Deep Learning Team Redmond, WA

Data Scientist Intern Summer 2022 and Summer 2023

- Trained a multi-task model to classify support tickets based on level of complexity.
- Developed infrastructure to host state-of-the-art LLMs a GPU cluster, providing integrating with DSPy to programmatically develop complex systems with LLMs.

PUBLICATIONS

Published

Zur, Amir, Elisa Kreiss, Karel D'Oosterlinck, Christopher Potts, and Atticus Geiger. "Updating CLIP to Prefer Descriptions Over Captions." *arXiv preprint arXiv:2406.09458*, 2024.

Wu, Zhengxuan, Karel D'Oosterlinck, Atticus Geiger, **Amir Zur**, and Christopher Potts. "Causal Proxy Models for concept-based model explanations." In *International Conference on Machine Learning*, pp. 37313-37334. PMLR, 2023.

Zur, Amir, Isaac Applebaum, Jocelyn E. Nardo, Dory DeWeese, Sameer Sundrani, and Shima Salehi. "Meta-Learning for Better Learning: Using Meta-Learning Methods to Automatically Label Exam Questions with Detailed Learning Objectives." In *International Conference on Educational Data Mining*, pp. 224-233. International Educational Data Mining Society, 2023.

In review

Geiger, Atticus, Duligur Ibeling, **Amir Zur**, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. "Causal Abstraction: A Theoretical Foundation for Mechanistic Interpretability." [Paper in review]. 2024.

DeWeese, Dory, Jocelyn Nardo, Isaac Applebaum, Sameer Sundrani, **Amir Zur**, Robert Waymouth, Jennifer Schwartz Poehlmann, and Shima Salehi. "The STEMentors Program: Promoting the Belonging of Historically Marginalized Students within Introductory Chemistry." [Accepted into *Journal of Chemical Education*, 2023]. Stanford University, 2023.

Unpublished

Zur, Amir. "Causal Abstraction for Interpretable, Debiased, and Accessible Language Models". [Unpublished undergraduate honors thesis]. Stanford University, 2023.

Su, Kein, **Amir Zur**, Jade Lintott, and Omer Reingold. "More Impossibilities between Calibration and Balance." [Poster presentation]. In *Undergraduate Research in CS (CURIS)*. Stanford University, 2020.

TEACHING

2019 (Winter), 2021 (Spring). Section leader for CS106A Programming Methodologies.

2019 (Spring), 2020 (Fall, Winter, Spring). Section leader for CS 106B Programming Abstractions.

2021 (Fall). Section leader for CS 105 Introduction to Computers, offered to Title I high school students by the National Equity Lab.

2020, 2021, 2022. Volunteer Section Leader for Code in Place, a live online course offered to over 10,000 students free of charge.

2024. Volunteer Teaching Intern for IB Math and Algebra I Lab in Rainier Beach High School, a Title I high school in Seattle.