Zachary **Robertson**

GRADUATE STUDENT

📧 zroberts@stanford.edu | 🏠 zrobertson466920.github.io/ | Phone: (832)-318-3039 | Mail: 51 Dudley Ln., Apt. 329 Stanford, CA 94305

"Play is the highest form of research."

Education

Stanford
Stanford, U.S.

PH.D. IN COMPUTER SCIENCE
Sept. 2022 -

- School of Engineering Fellowship
- EDGE Fellowship
- NSF Honorary Mention

University of Illinois at Urbana-Champaign
Champaign, U.S.

MS IN COMPUTER SCIENCE
Aug. 2020 - May. 2022

- Thesis on Performance Metric Elicitation
- GEM fellowship
- Wing Kai Cheng Fellowship

University of Chicago
Chicago, U.S.

B.S. IN COMPUTATIONAL AND APPLIED MATH
Oct. 2016 - June 2020

- Graduated with Honors
- Jackie Robinson Scholarship

LoneStar Community College
Spring, TX, U.S.

ASSOCIATE DEGREE IN SCIENCE
Oct. 2013 - May. 2016

- President's List
- Home schooled with dual enrollment at local community college

Experience

STAIR Lab at Stanford (advised by Sanmi Koyejo)
Stanford, U.S.

RESEARCH ASSISTANT
Sept. 2022 - Present

- Developing scalable oversight mechanisms and aligning AI systems with human preferences
- Contributing to the lab's work on AI superalignment, supported by an OpenAI award

Google
Mountain View, U.S.

STUDENT RESEARCHER
Apr. 2022 - Aug. 2022

- Designed tractable surrogates for welfare maximization
- Applications to the training of predicted click-through-rates models used in Ads

Lam Research
Fremont, U.S.

RESEARCH INTERN
June 2020 - March 2021

- Lam is a major player in the design, manufacture, marketing, and service of semiconductor processing equipment used in the fabrication of integrated circuits
- Applied knowledge of reinforcement learning and machine learning to optimize the production flow of semi-conducting wafers
- Used Python, Open AI Gym, and Tensorflow

Robot Intelligence through Perception Lab at TTIC
Chicago, U.S.

RESEARCH ASSISTANT
Jun. 2019 - June 2020

- Implemented a model to do sparse-depth completion on a robot using PyTorch, Docker, and SLURM.
- Trained a simulated UR5 arm to pick up blocks from a table with natural language using PyTorch, OpenGym, and Mujoco.
- Engaged in theoretical study of reinforcement learning using regret based approach for Honor's Thesis

Allen and Company LLC.
New York, U.S.

ANALYST
June 2018 - Aug. 2018

- Allen and Co. is a boutique investment banking firm that has advised on deals such as Facebook's acquisition of WhatsApp (2014) and hosts the annual Sun Valley Conference in Idaho
- Brought domain knowledge to help evaluate quantum computing startups and cryptocurrencies, automated formulaic internal reporting, and analyzed investor sentiment

Honors & Awards

2024	**Idea and Writing**, Lead author, Superalignment Fast Grant proposal ($500k, OpenAI) awarded to STAIR lab	*Stanford*
2023	**Idea and Writing**, Accelerating Foundation Models Research (Microsoft) awarded to STAIR lab	*Stanford*
2022	**Recipient**, School of Engineering Fellowship	*Stanford*
2022	**Recipient**, EDGE Fellowship	*Stanford*
2021	**Honorary Mention**, NSF GRFP Fellowship	*UIUC*
2021	**Recipient**, Wing Kai Cheng Fellowship	*UIUC*
2020	**Recipient**, GEM Fellowship	*UChicago*
2018	**Recipient**, Dean's List	*UChicago*
2017	**Winner**, Illinois Blockchain Hackathon	*Chicago/Urbana*
2016	**Scholar**, Jackie Robinson Foundation	*New York*

Service

WellLabeled Affinity Group *HAI*

Lead *Sept. 2023 - May. 2024*

- Led an advocacy group for data annotation worker rights
- Interviewed stakeholders from Turkopticon, Scale AI, and OpenAI
- Featured article and video testimonial. See here.

Alignment Newsletter *CHAI*

Contributor *Sept. 2019 - Sept. 2023*

- Wrote summaries and opinions on advances in machine learning for the Alignment Newsletter
- Over 1k subscribers

Skills

Programming	Python, LaTeX, C++, Haskell
Machine Learning	Prompt Engineering, PyTorch
Math	Analysis, Linear Algebra, Statistical Learning Theory, Optimization, Statistics

Publications and Presentations

Presentations

1. "Exploring the Complex Ethical Challenges of Data Annotation", Invited Talk, FloodGate, San Francisco, September 2024

2. "Towards Scalable Information Elicitation for Oversight in Human-AI Systems", Invited Talk, Max Planck Institute for Intelligent Systems, Tubingen, July 2024

Selected Publications

1. Zachary Robertson and Sanmi Koyejo. Let's measure information step-by-step: Llm-based evaluation beyond vibes. *arXiv preprint arXiv:2505.10573*, 2025

2. Zachary Robertson and Sanmi Koyejo. Implict regularization in feedback alignment mechanisms. *ICML*, 2024

3. Boxiang Lyu, Zhe Feng, Zachary Robertson, and Sanmi Koyejo. Pairwise ranking losses of click-through rates prediction for welfare maximization in ad auctions. *ICML*, 2023

4. Zachary Robertson, Hantao Zhang, and Sanmi Koyejo. Cooperative inverse decision theory for uncertain preferences. *AISTATS*, 2023

Additional Works

1. Olawale Salaudeen, Anka Reuel, Ahmed Ahmed, Suhana Bedi, Zachary Robertson, Sudharsan Sundar, Ben Domingue, Angelina Wang, and Sanmi Koyejo. Measurement to meaning: A validity-centered framework for ai evaluation. *arXiv preprint arXiv:2505.10573*, 2025

2. Zachary Robertson, Hannah Cha, Andrew Sheha, and Sanmi Koyejo. Implementability of information elicitation mechanisms with pre-trained language models. *ICML 2024 Workshop on Theoretical Foundations of Foundation Models*, 2024

3. Zachary Robertson and Oluwasanmi Koyejo. No bidding, no regret: Pairwise-feedback mechanisms for digital goods and data auctions. *Workshop on Information-Theoretic Methods for Trustworthy Machine Learning at IEEE ISIT*, 2024

4. Rylan Schaeffer, Zachary Robertson, Akhilan Boopathy, Mikail Khona, Ila R Fiete, Andrey Gromov, and Sanmi Koyejo. Divergence at the interpolation threshold: Identifying, interpreting & ablating the sources of a deep learning puzzle. *NeurIPS 2023 Workshop on Mathematics of Modern Machine Learning*, 2023

5. Zachary Robertson. Gpt4 is slightly helpful for peer-review assistance: A pilot study. *arXiv preprint arXiv:2307.05492*, 2023

6. Zachary Robertson, Hantao Zhang, and Sanmi Koyejo. Probabilistic performance metric elicitation. *1st Workshop on Human and Machine Decisions (WHMD 2021) at NeurIPS 2021*, 2022

7. Zachary Robertson and Matthew Walter. Concurrent training improves the performance of behavioral cloning from observation. *arXiv preprint arXiv:2008.01205*, 2020

8. Julian Stürmer, Andreas Seifahrt, Zachary Robertson, Christian Schwab, and Jacob L Bean. Echelle++, a fast generic spectrum simulator. *Publications of the Astronomical Society of the Pacific*, 131(996):024502, 2018