# 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** | Evaluation design, mechanism design, adversarial robustness |
| **Math** | Analysis, Linear Algebra, Statistical Learning Theory, Optimization, Statistics |

## Publications and Presentations

### Presentations and External Impact

1. "Paying for Information: Incentive-Compatible LLM Evaluation via f-Mutual Information", Invited Lightning Talk, *Building an Aligned AI Future* (Fifty Years @ Stanford), September 2025
2. "Information-Theoretic Evaluation Methods", Featured Implementation, Jina AI Production Systems ($30M Series A), 2024
3. "Exploring the Complex Ethical Challenges of Data Annotation", Invited Talk, FloodGate, San Francisco, September 2024
4. "Towards Scalable Information Elicitation for Oversight in Human-AI Systems", Invited Talk, Max Planck Institute for Intelligent Systems, Tübingen, July 2024

### Selected Publications

1. Farnaz Jahanbakhsh, Dora Zhao, Tiziano Piccardi, Zachary Robertson, Ziv Epstein, Sanmi Koyejo, and Michael S Bernstein. Value alignment of social media ranking algorithms. *arXiv preprint arXiv:2509.14434*, 2025
2. Zachary Robertson and Sanmi Koyejo. Let's measure information step-by-step: Llm-based evaluation beyond vibes. *arXiv preprint arXiv:2508.05469*, 2025
3. Zachary Robertson and Sanmi Koyejo. Implict regularization in feedback alignment mechanisms. *ICML*, 2024
4. 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
5. Zachary Robertson, Hantao Zhang, and Sanmi Koyejo. Cooperative inverse decision theory for uncertain preferences. *AISTATS*, 2023