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 -*

- Google Graduate Fellowship in Computer Science
- 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

Stanford University — *Stanford, U.S.*

Instructor — *2025 - Present*

- Instructor for Introduction to AI Safety (CS120); previously taught core Artificial Intelligence curriculum (CS221)

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

Research Assistant — *Sept. 2022 - Present*

- Developed mutual evaluation, an information-theoretic framework for evaluating AI systems without ground truth; first-author work published in TMLR
- Conceived, wrote, and submitted a $500K OpenAI Superalignment grant proposal; advisor served as institutional PI

Google — *Mountain View, U.S.*

Student Researcher — *Apr. 2022 - Aug. 2022*

- Designed tractable surrogates for welfare maximization
- Applied these methods to predicted click-through-rate models used in Ads

Lam Research — *Fremont, U.S.*

Research Intern — *June 2020 - March 2021*

- Applied reinforcement learning and machine learning to optimize semiconductor wafer production flows
- 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*

- Evaluated quantum-computing startups and cryptocurrencies, automated 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**, Google Graduate Fellowship in Computer Science	*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 (Organized by Rohin Shah) *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

Presentations and Talks

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

Publications

Peer-Reviewed Journal and Conference Papers

1. Farnaz Jahanbakhsh, Dora Zhao, Tiziano Piccardi, **Zachary Robertson**, Ziv Epstein, Sanmi Koyejo, and Michael S. Bernstein. Value alignment of social media ranking algorithms. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, pages 1401:1–1401:26. ACM, 2026
2. **Zachary Robertson** and Sanmi Koyejo. Let's measure information step-by-step: AI-based evaluation beyond vibes. *Transactions on Machine Learning Research*, 2026
3. **Zachary Robertson** and Sanmi Koyejo. Implicit regularization in feedback alignment learning mechanisms for neural networks. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 42601–42619. PMLR, 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. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 23239–23263. PMLR, 2023

5. **Zachary Robertson**, Hantao Zhang, and Sanmi Koyejo. Cooperative inverse decision theory for uncertain preferences. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 5854–5868. PMLR, 2023

6. 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, 2019

Workshop Papers

7. 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. In *NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling*, 2025

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

9. **Zachary Robertson** and Oluwasanmi Koyejo. No bidding, no regret: Pairwise-feedback mechanisms for digital goods and data auctions. In *Workshop on Information-Theoretic Methods for Trustworthy Machine Learning at IEEE ISIT*, 2024. Preprint released in 2023; workshop presentation in 2024

10. 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. In *NeurIPS 2023 Workshop on Mathematics of Modern Machine Learning*, 2023. Also presented at the NeurIPS 2023 Workshop on Attributing Model Behavior at Scale

11. **Zachary Robertson** and Oluwasanmi Koyejo. Layer-wise feedback alignment is conserved in deep neural networks. In *ICML 2023 Localized Learning Workshop*, 2023

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

Preprints

13. **Zachary Robertson**. Mutual evaluation and supervision without peers. *arXiv preprint arXiv:2609.20789*, 2026

14. **Zachary Robertson**. A global characterization of f-divergences yielding PSD mutual-information matrices. *arXiv preprint arXiv:2601.08929*, 2026

15. **Zachary Robertson**. Identity-link IRT for label-free LLM evaluation: Preserving additivity in TVD-MI scores. *arXiv preprint arXiv:2510.14966*, 2025

16. **Zachary Robertson**. GPT4 is slightly helpful for peer-review assistance: A pilot study. *arXiv preprint arXiv:2307.05492*, 2023

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

Thesis

18. **Zachary Robertson**. Probabilistic performance metric elicitation. Master's thesis, University of Illinois Urbana-Champaign, 2022. Distinct from the coauthored WHMD 2021 workshop paper

Other Scholarly Writing and Software

19. Rylan Schaeffer, **Zachary Robertson**, Akhilan Boopathy, Mikail Khona, Kateryna Pistunova, Jason W. Rocks, Ila R. Fiete, Andrey Gromov, and Sanmi Koyejo. Double descent demystified. In *ICLR Blogposts 2024*, 2024. Subtitle: Identifying, Interpreting & Ablating the Sources of a Deep Learning Puzzle

20. **Zachary Robertson**. You can probably amplify GPT3 directly. LessWrong, July 2020

21. Julian St”urmer, Andreas Seifahrt, **Zachary Robertson**, Christian Schwab, and Jacob L. Bean. Echelle++: Generic spectrum simulator. Astrophysics Source Code Library, 2018. Astrophysics Source Code Library, ascl:1810.006; identifier supplied by Google Scholar