Irena Saracay

irena@cs.stanford.edu · Google Scholar · Academic Website

Research interests: user modeling, personal agents, and human oversight of AI systems

Education

Ph.D. in Computer Science, *Stanford University.* 2023–Present

M.S. in Statistics, *Stanford University.* 2021–2023

B.S. in Computer Science with Honors and Distinction, *Stanford University.* 2018–2023

Industry Experience

Research Intern, Microsoft Research 2022

Advised by Marco Tulio Ribeiro.

- **AdaVision** (ICCV 2023, oral): Built a human-in-the-loop testing tool where users and foundation models collaboratively discover failure modes of vision models in natural language, surfacing bugs with 2–3x higher failure rates than automatic clustering.
- Ran user studies on SOTA classification, detection, and captioning models, and fixed discovered bugs by fine-tuning while preserving in-distribution accuracy.

Academic Experience

PhD Researcher, Stanford Artificial Intelligence Lab 2023–Present

Advised by Carlos Guestrin.

- **OpenFlamingo** (2023; 4.1K GitHub stars, 1K+ citations): Co-led an open-source replication of DeepMind's Flamingo; owned the PyTorch multimodal pretraining pipeline for 3B–9B vision-language models, reaching 80–89% of Flamingo's performance.
- **Model equality testing** (ICLR 2025): Designed a statistical test for inference API oversight to detect model drift from ∼10 samples per prompt. Audited 31 commercial APIs, flagging 11 as modified.
- **CoShop** (COLM 2026): Built an economics-grounded user simulator whose preferences change based on agent actions. Used it to evaluate personal agents, showing shopping agents fail to aid non-expert users.

Research Assistant, Stanford NLP Group 2020–2023

Advised by Percy Liang and Pang Wei Koh.

- **WILDS** (ICLR 2022 oral, ICML 2021; 2.8K+ citations): Extended a benchmark of real-world distribution shifts with unlabeled data for 10 datasets, and evaluated the robustness gains of popular adaptation methods.
- **Targeted augmentations** (ICML 2023): Studied how to design data augmentations for robustness to distribution shifts, combining theory with experiments that set a new state of the art on 3 real-world datasets by 3.2–15.2 points.

Honors

Crosby Intelligence Fellowship 2026

ICCV Oral (top 6.72% of accepted papers), *AdaVision* 2023

CVPR Highlight (top 10% of accepted papers), *CREPE* 2023

ICLR Oral (top 5% of accepted papers), *Extending WILDS* 2022

Publications

* denotes equal contribution. Irena Gao is a former name.

[1] **Irena Saracay**, Ludwig Schmidt, and Carlos Guestrin. Beyond expert users: agents should help users construct preferences, not just elicit them. In *Conference on Language Modeling*, 2026.

[2] **Irena Gao**, Percy Liang, and Carlos Guestrin. Model equality testing: which model is this api serving? In *International Conference on Learning Representations*, 2025.

[3] Anas Awadalla*, **Irena Gao***, Joshua Gardner, Jack Hessel, Yusuf Hanafy, et al. Openflamingo: an open-source framework for training large autoregressive vision-language models, 2023.

[4] **Irena Gao**, Gabriel Ilharco, Scott Lundberg, and Marco Tulio Ribeiro. Adaptive testing of computer vision models. In *International Conference on Computer Vision*, 2023. **Oral presentation (top 6.72% of accepted papers).**

[5] **Irena Gao***, Shiori Sagawa*, Pang Wei Koh, Tatsunori Hashimoto, and Percy Liang. Out-of-domain robustness via targeted augmentations. In *International Conference on Machine Learning*, 2023

[6] Nicholas Carlini, Milad Nasr, Christopher A Choquette-Choo, Matthew Jagielski, **Irena Gao**, et al. Are aligned neural networks adversarially aligned? In *Neural Information Processing Systems*, 2023.

[7] Zixian Ma*, Jerry Hong*, Mustafa Omer Gul*, Mona Gandhi, **Irena Gao**, et al. Crepe: can vision-language foundation models reason compositionally? In *Conference on Computer Vision and Pattern Recognition*, 2023. **Highlight (top 10% of accepted papers).**

[8] Shiori Sagawa*, Pang Wei Koh*, Tony Lee*, **Irena Gao***, Sang Michael Xie, et al. Extending the wilds benchmark for unsupervised adaptation. In *International Conference on Learning Representations*, 2022. **Oral presentation (top 5% of accepted papers).**

[9] Pang Wei Koh*, Shiori Sagawa*, ..., **Irena Gao**, et al. Wilds: a benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pp. 5637–5664. PMLR, 2021.

[10] Tommy Bruzzese*, **Irena Gao***, Christina Ding*, Alyssa Romanos*, and Griffin Dietz. Effect of confidence indicators on trust in ai-generated profiles. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–8, 2020.

[11] Mark E Whiting, **Irena Gao***, Michelle Xing*, N'godjigui Junior Diarrassouba*, Tonya Nguyen*, et al. Parallel worlds: repeated initializations of the same team to improve team viability. *proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1):1–22, 2020.

Service

Reviewer, ICLR, ICML, NeurIPS, CVPR, ICCV, COLM	
Organizer, Stanford NLP Seminar	2025–2026
Organizer, CURIS: departmental research program for undergraduates	2024–2026
Organizer, Stanford Student Application Support Program	2023–2025

Teaching

Course Assistant, CS224N: *Natural Language Processing with Deep Learning*	2023
Course Assistant, CS221: *Artificial Intelligence: Principles and Techniques*	2022–2023
Section Leader, CS106B: *Programming Abstractions*	2020
Section Leader, CS106A: *Programming Methodology*	2019