**Shawn Im**

Graduate Student
University of Wisconsin-Madison

**Email**: shawnim@cs.wisc.edu

**Interests** AI Safety, Learning theory, Interpretability

**Education**

**University of Wisconsin-Madison**
Doctor of Philosophy Computer Science — 2023 - Present

**Massachusetts Institute of Technology**
B.Sc in Mathematics, Computer Science — 2019 - 2023

**Research Experience**

**University of Wisconsin-Madison** — 2023 - Present
Advisor: Sharon Li
- Developed a theoretical framework to understand the generalization of preference learning and extending the framework to model the impact of noisy labels

**MIT CSAIL** — 2022-2023
Advisors: Yilun Zhou, Jacob Andreas
- Developed an evaluation method for saliency maps for image classification based on the saliency map's ability to improve user performance on a task representing a practical use case

**MIT Mathematics** — 2022-2023
Advisors: Sungwoo Jeong, Alan Edelman
- Studied the spectral properties of Neural Tangent Kernels using Random Matrix Theory in a feature learning regime

**Julia Lab** — 2020-2021
Advisors: Chris Rackauckas, Alan Edelman
- Developed models for chemical reactions for batteries and for pollutants using surrogate models and Neural ODEs

**Media Lab** — 2019-2020
Advisors: Takatoshi Yoshida, Hiroshi Ishii
- Developed a model to classify a person's activity (e.g. walking, spinning) through force sensors embedded in the floor

**Industry Experience**

**Amazon Software Engineer Intern** — Summer 2021
- Developed an end-to-end AWS framework to delete user data upon request integrating SNS, Lambda, EMR, S3, API Gateway

**Publications**

**Shawn Im**, Yixuan Li. On the Generalization of Preference Learning with DPO. Preprint, 2024.

**Shawn Im**, Yixuan Li. Understanding the Learning Dynamics of Alignment with Human Feedback. In Proceedings of International Conference on Machine Learning (ICML), 2024.

**Shawn Im**, Jacob Andreas, Yilun Zhou. Evaluating the Utility of Model Explanations for Model Development. NeurIPS Workshop on Attributing Model Behavior at Scale (ATTRIB), 2023.

**Activities**

| | |
|---|---|
| Wisconsin AI Safety Initiative, Research Team | Fall 2024 - Present |
| Wisconsin AI Safety Initiative, Safety Scholars | Spring 2024 - Present |
| Wisconsin AI Safety Initiative, AI Safety Fundamentals Facilitator | Fall 2023 |
| Grader, Theory of Probability (18.675) | Fall 2022 |
| MIT Math Learning Center Tutor | Fall 2021 |