KAIVU HARIHARAN

908 448 8004 ◇ kaivu@mit.edu

EDUCATION

Massachusetts Institute of Technology Class of 2024
Double Major in Mathematics, Machine Learning Overall GPA: 5/5
Prospective Masters in AI

RELEVANT COURSEWORK

Computer Science
Advanced Algorithms (Grad.)
Deep Learning (Grad.)
Computer Vision
Theory of Computation

Math
Abstract Algebra
Real Analysis
Combinatorics, Probability
Project Lab in Mathematics

RESEARCH EXPERIENCE

Model Internals for NN Failures December 2023 - Ongoing
MIT Tang Family FinTech Undergraduate Research and Innovation Scholar

· Refining Features not Bugs theory of adversarial examples: aim to allows more concrete measurement of non-robust features

Feature Level Adversaries (SNAFUE) May 2022 - Jan 2023
Researcher in Hadfield-Menell Lab

· Created automated pipeline for class-universal, targeted, copy-paste attacks/feature level natural patch adversaries as interpretability/debugging tool using latent space of GANs
· Best paper award in NeurIPS ML Safety Workshop 2022, First Author
· Follow up work accepted to NeurIPS 2023

Understanding Competing Objectives in LLMs March 2023 - Ongoing
MIT Tang Family FinTech Undergraduate Research and Innovation Scholar

· Understanding mechanisms involved in how Llama models reconcile competing objectives between correct factual recall and user request
· Paper accepted to Attrib and SoLaR Workshops NeurIPS 2023)

Understanding Shortcuts for Automata in Transfomers Jan 2023
Remix Research Resident (Redwood Research)

· Working on mechanistic interpretability for shortcut solutions (sublinear depth) for simulating finite automata with Transformer, using Causal Scrubbing methodology
· Found that the networks can learn algorithms resembling theoretical shortcut solutions for Gridworld-9

Other Research March 2022 - Ongoing

· Independent research on Mechanistic Interpretability of Grokking in Group-theoretic Models
· Independent research on Statistical Signatures of Learning
· Implemented policy space explainable AI (Shah Lab - 2021)
· Worked as lab assistant in two bioinfomatics labs

LEADERSHIP

AI Safety 2021 - Ongoing

· Founder of MAIA (MIT AI Alignment): facilitating AI Safety reading group, graduate member meetings, lightning talks

TEACHING

Machine Learning 2021 - Ongoing

· TA'd @ ML Safety Scholars (2022 summer), taught introductory course on Machine Learning and AI safety
· Facilitated MAIAs graduate member meetings
· Facilitated AI Safety Fundamentals

Math 2020 - Ongoing

· Hosted Problem Solving in Probability Seminar
· Tutored High School Math

AWARDS AND INTERESTS

Robert A. Boit Poetry Manuscript Prize ($500 prize, 2nd place)
Qualified for the USABO Nationals Camp: top 20 in USA
Captained RHS Science Bowl Team to top 8 in country
AIME Qualifier
Interests: Writing, Tennis, Designing Games, Literature