KAIVU HARIHARAN

Website ⋄ 908 448 8004 ⋄ kaivu@mit.edu ⋄ Google Scholar

EDUCATION

Massachusetts Institute of Technology Class of 2024 SB, 2025 MEng
SB in Mathematics, Machine Learning Overall GPA: 5/5
MEng Candidate in AI, Supervised by Antonio Torralba and Roger Grosse

RELEVANT COURSEWORK

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

Math
Abstract Algebra
Real Analysis
Combinatorics
Project Lab in Mathematics

RESEARCH EXPERIENCE

Feature Level Adversaries (SNAFUE) May 2022 - Jan 2023
Algorithmic Alignment Lab

- Created automated pipeline for class-universal, targeted, copy-paste attacks/feature level natural patch adversaries as debugging tool using latent space of GANs
- Best paper award in NeurIPS ML Safety Workshop 2022 (Equal-First) - follow up work in NeurIPS 2023

Model Internals for LLM Jailbreaks December 2023 - Ongoing
Torralba Lab

- Studying phenomenology of gradient-based jailbreaks (e.g. GCG).
- Accepted to NeurIPs Attrib and Redteaming GenAI workshops 2024

SAD: Situational Awareness Detection in LLMs Feb 2024 - June 2024
Truthful AI (Owain Evans)

- Developing benchmarks for characterizing and evaluating situational awareness in LLMs
- Worked on Anti-imitation benchmark
- Paper accepted to NeurIPS 2024 in Benchmarks and Datasets Track

Understanding Competing Objectives in LLMs March 2023 - September 2023
Shavit Lab

- Understanding mechanisms involved in how Llama models reconcile competing objectives between correct factual recall and user request
- Paper (Equal-First) 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

ONGOING RESEARCH

Data Attribution Jan 2023 - Ongoing
Grosse Lab

- Studying out-of-context reasoning via influence functions
- Developing datamodels for chain of thought analysis

Semantic Contamination December 2023 - Ongoing
KASL Internship

- Taxonimizing and detecting when models are trained on paraphrased evaluation data

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

Awards

- Qualified for the USABO Nationals Camp: top 20 in USA (2019)
- Captained RHS Science Bowl Team to top 8 in country (2019)
- AIME Qualifier
- Robert A. Boit MIT Poetry Manuscript Prize ($500 prize, 2nd place) (2023)

Interests Ongoing

- Writing, Tennis, Designing Games, Literature