Eric Wang

732-492-2378 | eric_wang@college.harvard.edu | linkedin.com/in/ericliwang/ | https://github.com/ericwang1409

Education

Harvard University Cambridge, MA

AB in Computer Science and Mathematics. GPA: 4.0/4.0 September 2022 - May 2026

Coursework: Data Structures & Algorithms, Computer Systems, Machine Learning, ML Interpretability, Probability

Activities: Harvard AI Safety, Harvard Tech for Social Good, Harvard Computer Society, Harvard Climbing

Experience

Research Intern February 2025 - Present

Redwood Research Berkeley, CA

- Conducted research on early signs of collusion abilities in Large Language Models between agents and monitors
- Developed evaluation methodologies that identified subtle signs of strategic behavior in AI agents, highlighting potential vulnerabilities in oversight mechanisms

Computer Science Teaching Assistant January 2024 – Present

Harvard University Cambridge, MA

- Held office hours and taught weekly recitations for 40+ students, supporting problem sets and theoretical concepts
- Teaching Assistant for CS 121 Introduction to Theoretical Computer Science -Fall 2024
- Teaching Assistant for CS 51 Abstraction and Design in Computation - Spring 2024

Machine Learning Engineer Intern June 2024 - August 2024

Analog Devices San Jose, CA

- Built an end-to-end human activity classification model using PyTorch, pre-trained with self-supervised learning techniques, and fine-tuned on open-source accelerometer data, achieving an average accuracy of 99.3%
- Created a data pipline to download, process, and clean over 24 TB of accelerometer data, utilizing AWS S3 for storage, EC2 for cloud computing, and parallel processing to speed up data processing sevenfold
- Coordinated closely with team members and customers through biweekly meetings to ensure the model met the desired requirements such as low latency and desired memory size

Full Stack Engineer Intern October 2023 - January 2024

Beaver Health Boston, MA

- Developed and maintained a full stack application using TypeScript, PostgreSQL, and OpenAI API, significantly streamlining interactions with an online dementia chatbot, resulting in 100% reliability in user interactions
- Designed and implemented a user-friendly flashcard feature using TypeScript and integrated it with the PostgreSQL database, providing users with an effective tool for memory enhancement and cognitive exercises

Projects

Sparse Autoencoders Thresholding for Safety | *Python, PyTorch* September 2024 - Present

- Developed and analyzed sparse autoencoders (SAEs) trained on open-source LLMs, establishing them as a competitive method for real-time model response oversight compared to residual stream activation probing
- Demonstrated SAEs to be as effective (98%) and more robust than probing SAE and residual stream activations

Mechanistic Interpretability of Maximum of Lists | *Python, PyTorch* January 2024 - May 2024

- Trained a single-layer, attention only transformer written from scratch to take the maximum number across variable length lists, achieving an average 97.4% test accuracy across all variations
- Performed mechanistic interpretibility on the attention patterns to analyze and write a paper on the results

TaiYo! Solver | *Python, Pymunk, Deep Q-Learning, PyTorch* November 2023

- Engineered a custom game from the ground up, utilizing PyGame for engaging gameplay interfaces and PyMunk for accurate physics simulations utilizing object-oriented programming
- Implemented a Deep-Q Learning algorithm to train a model to autonomously play the game using optimal actions over the state space, outperforming over 80% of human players based on comprehensive gameplay metrics
- Won most ambitious/best idea hack at HackWellesley with a team of 3

Technical Skills

Languages: Python, Java, Javascript, Typescript, Rust, C/C++, C#, Swift, OCaml, PostgreSQL, HTML/CSS
Frameworks: PyTorch, React, Node.js, Next.js, Electron
Developer Tools: Git, Docker, Amazon AWS, Gradle, VS Code, Linux