Jonathan Lu

jonathan.lu31@gmail.com | (858)-610-9718 | www.linkedin.com/in/jonathanlu4b35 | https://github.com/jonathanlu31

Education

UNIVERSITY OF CALIFORNIA, BERKELEY | AUGUST 2024 - MAY 2025

M.S. Electrical Engineering and Computer Science | GPA: 3.95

- **Selected coursework**: Natural Language Processing, Applications of Parallel Computers, Advanced Language Model Agents

UNIVERSITY OF CALIFORNIA, BERKELEY | AUGUST 2020 - MAY 2024

B.A. Computer Science with Highest Distinction | GPA: 4.00

- **Selected coursework**: Deep Reinforcement Learning, Operating Systems, Deep Neural Networks, Computer Graphics
- **Activities**: UPE honor society industrial relations officer, CS 61A/CS189 TA, *The Daily Californian* mobile developer
- **Awards**: Arthur M Hopkin Award, Outstanding GSI

Experience

BAIR RESEARCHER, WAGNER LAB | OCTOBER 2022 - PRESENT

- Create synthetic datasets and train LLMs for adversarial robustness using SFT and DPO for a paper accepted at NeurIPS 2024 Safe Generative AI workshop
- Develop custom PyTorch fine-tuning library similar to torchtune
- Use LLM-as-a-Judge to set up a new benchmark to measure system instruction following robustness
- Trained classifiers on LLM activations to detect success of adversarial attacks
- Implement inference-time interventions like attention bias and classifier-free guidance to improve adversarial robustness of LLMs

ALGOVERSE RESEARCH MENTOR | JUNE 2024 - OCTOBER 2024

- Mentored five teams of highschoolers to implement research projects with some getting accepted to NeurIPS workshops
- Taught concepts like transformers, LLM training, alignment and latent-space steering

DEEPGRAM SOFTWARE ENGINEERING INTERN | MAY 2023 - AUGUST 2023

- Used LangChain and LLMs to develop a RAG chatbot for API documentation
- Integrated data sources and set up web scrapers to process the data for RAG
- Created a REST API using FastAPI to interact with the bot and integrated with Slack for internal onboarding usage
- Set up CI/CD workflows using GitHub Actions
- Created an evaluation set from past community issues to measure chatbot performance with LLM-as-a-Judge

CS 189 TEACHING ASSISTANT | JANUARY 2023 - MAY 2023

- Taught machine learning concepts like SVMs, logistic regression, neural networks, etc.
- Answered student questions in office hours and on the Edstem question forum
- Contributed to and used the unofficial Edstem API to develop a Slack bot to automatically assign unanswered questions to staff

META ENTERPRISE ENGINEERING INTERN | MAY 2022 - AUGUST 2022

- Added new workflows for a supply chain product to decrease time spent on repetitive purchase requests by 5x
- Worked full-stack with React, GraphQL, Relay, and Hack to enable users to reuse and share previous intake requests
- Iterated with designers and business stakeholders to improve UX and created unit tests to verify backend functionality

Projects

DEEP REINFORCEMENT LEARNING FROM AI FEEDBACK | DECEMBER 2023

- Queried GPT4-Vision for visual preference ratings to train a reward model in the style of DeepMind's Deep Reinforcement Learning From Human Preferences paper and created an RLHF pipeline as a baseline
- Used Stable Baselines3 to train a policy with the resulting reward model for tasks like doing the splits

Skills

Languages: Python, JavaScript, C, Java, TypeScript, Hack/PHP, Go

Frameworks/Technologies: 🤗Transformers, PyTorch, FastAPI, LangChain, Qdrant, React.js, Node.js, Express.js, React Native, GraphQL, Relay, Flask