Shawn Hu

shawnghu@gmail.com
512-296-4617
shawnghu.github.io

EDUCATION

Stanford University — Sept. 2016 - June 2018
MS Computer Science with specialization in AI, GPA 3.7

The University of Texas at Austin — August 2014 - May 2016
BS Mathematics, GPA 3.9, graduated age 18

WORK AND RESEARCH PROJECTS

Founding Researcher at d_model — April 2025 - present

- Basic research on interpretability and reward hacking
- Various tasks of SWE/MLE flavor for customers as needed, most recently developing RL envs at scale for big labs

Senior Research Scientist at PlusAI — July 2018 - July 2023

- First 20 engineers; company develops self-driving trucks, is now series C
- Developed core algorithms for lidar perception, sensor fusion, and object tracking, using a combination of deep learning and classical robotics approaches
- Wrote the system's core "classical" lidar perception algorithm and core lidar sensor fusion logic
- Owned the pipeline for the 1st-gen DL lidar model (sourcing labels, developing and maintaining data pipelines, training/evaluating models (PyTorch), deployment with C++/TensorRT)
- Also designed/contributed to sensor driver applications, simulators and visualizers, root-cause analysis and debugging tools, optimization/deployment/benchmarking in many areas.
- Project managed the company's L4 autonomy demo to investors/partners

Machine Learning Intern at PlusAI — June 2017 - Sept. 2017

- Used deep reinforcement learning to train agents to execute control for simulated cars
- Wrote the car dynamics simulator and environment for the above agents

ACADEMICS

Collaboration as independent with Dawn Song's lab, UC Berkeley — March 2025 - May 2025

- OMEGA: Can LLMs Reason Outside the Box in Math? Evaluating Exploratory, Compositional, and Transformative Generalization; second author

MATS 1 Scholar — Sept. 2021 - Dec. 2021

- Disentangling Perspectives On Strategy-Stealing in AI Safety on the Alignment Forum

SKILLS AND KNOWLEDGE

- General, longstanding familiarity with AI safety and alignment and the surrounding technical and philosophical debates; awareness of high-level agendas, their assumptions and value propositions
- Understanding of modern ML in theory and practice, e.g, the limitations of classical learning theory, core techniques/workflows in mechanistic interpretability, scaling laws, optimization of LLM inference and fine-tuning, how to implement basic LLM agents, modern RL setups
- Thorough knowledge of reinforcement learning, optimization and approximation theory, numerical analysis, classical ML/statistics/stochastic methods, (SVMs, MCMC, Bayesian networks), information theory, classical techniques in CV/NLP.
- Experience w/ DL applications in CV/3D obj. detection/NLP/RL, incl. deployment w/ TensorRT
- High-level knowledge of most of an autonomous vehicle's stack, including control and planning, mapping/localization/odometry, stereo vision, sensor calibration, general robotics algorithms
- Adept with Python and C++; naturally, familiar with the modern Python ML stack
- Misc. software tech's, e.g, Docker, AWS, various frontend tech's, various client-server arch's