

Yourui Shao

yourui@caltech.edu | github.com/iuruoy-shao

Hello! I'm Yourui.

I have a background in applicational machine learning, full-stack software engineering, and math. Currently, I'm broadly interested in AI algorithm development (foundational models, inference-time scaling, multimodality) and the computational side of brain-computer interfaces, but I enjoy solving problems of all sorts, and I love debugging.

Education

California Institute of Technology

Grad. 2029

B.S. in Computer Science, minor in Neurobiology

- Courses: Machine Learning & Data Mining, Learning Systems, Intro to Computing Systems, Software Design in C, Analytical Linear Algebra, Analytical Multivariable Calculus, Differential Equations, Combinatorial Analysis

Experience

Summer Undergraduate Research Fellow

Jun 2026 – Sep 2026

@ Stanford Neural Prosthetics Translational Lab

- Task-driven modeling of concept-to-phoneme translation during speech production to verify neuroscientific theories. Neural interpretability for BCIs.
- Analyzing effects of structural memory and delay constraints on RNN model geometry. Probing RNN hidden states with linear decoders for phoneme identification, slot similarity and rotation score comparisons.
- Building large scale word embedding, sentence embedding, & token to phoneme training pipeline with model variants.

Undergraduate Researcher

Mar 2026 – Present

@ Mazumdar Group (Rigorous Systems Research Group), Caltech

- Optimizing multi-agent systems inference-time scaling of general problem-solving domains with information theory.
- Developed custom, GEPA-optimized domain-specific skills template generation & embeddings-based retrieval pipelines to generate subdomain experts for math problem solving. Analyzing leverage and efficacy of template expertise.
- Designing & running experiments. Developed comprehensive evaluation modules for inference, template, & retrieval with ablationary baseline testing, template & expert contribution metrics, & LLM answer judging.

Projects

Relay

Nov 2025

@ HackPrinceton 2025 | Python, LangChain, FastAPI, TypeScript, React

- In 20 hours, built functioning full-stack, multi-agent AI web application to look up and cite, write articles, and notify users on state legislative activity using RAG system capable of self-updating via live scraping congressional sites.
- Developed conversational agent with custom source retrieval tooling and adaptive to user interest and reading level.

Reinforcement Learning for *Fish* | Python, PyTorch

Jan 2025 – May 2025

- Senior research project. Reinforcement learning model for multiplayer, team-based card game *Fish*.
- Built and optimized custom modular two-part (1) LSTM hand prediction and (2) Q-learning decision-making neural networks in PyTorch generalizing beyond card, suit, and player order. Designed model from studying human gameplay.
- Programmed custom state-action converter, data parser, and reward functions selected via testing.
- Self-collected and parsed real-world gameplay data. Programmed custom multiplayer self-play gameplay pipeline to generate simulated data for exploration and training across many iterations.

Publications

Predicting College Admission Results with Machine Learning on Unstructured Online Data, 2024 IEEE FMLDS

An Accurate Classification and Recommendation Method of Competitive Math Problems, 2024 IEEE BigDataService

Skills

- **Languages & Tools:** Python, DSPy, LangChain, CrewAI, FastAPI, Flask, PyTorch, TensorFlow, Numpy, Pandas, JavaScript, TypeScript, React, NodeJS, Vercel SDK, HTML/CSS, SQL (Postgres, MySQL), git, Docker
- **Concepts:** Natural Language Processing, Transformers, Reinforcement Learning, Deep Learning, AI Agent Systems, Feature Engineering, Statistical Analysis, REST APIs, Embedded Systems, Full Stack, Databases