Andy Liu

+1 510-648-7681 | a.liu27568@gmail.com | ascl1u.github.io | linkedin.com/in/andy-liu-sic

Experience

Amazon May 2025 – July 2025
Software Development Engineer Intern *Seattle, WA*

- Reduced manual compliance audit time by 75% by architecting a full-stack audit feature serving comprehensive logs to 60+ developers and 500+ admins via a scalable REST API.
- Cut data-loading latency by 96% (from 10s to 400ms) by refactoring a critical authorization page, consolidating 5 REST endpoints into a single GraphQL query and removing 1,500+ lines of legacy code.

Bambu Lab June 2024 – September 2024
Machine Learning Engineer Intern *Shenzhen, China*

- Developed a video recognition model to identify filament clumping, achieving over 97% accuracy and preventing costly hardware damage.
- Engineered a high-throughput inference pipeline to process 1,000,000+ print-monitoring events daily, using Kafka and Redis to enable real-time AI-driven defect detection.
- Benchmarked model architectures for real-time edge inference to optimize performance on resource-constrained hardware.

Lawrence Berkeley National Laboratory July 2023 – September 2023
Research Assistant *Berkeley, CA*

- Applied physics-informed neural networks to predict spectra from nonlinear amplifiers, increasing simulation accuracy by 10%.

Education

University of California San Diego September 2022 – December 2024
Bachelor of Science in Mathematics-Computer Science *La Jolla, CA*

- GPA: 3.81/4.00

Projects & Open Source

Prime Intellect (Open Source) October 2025 – Present

- Built a containerized Next.js codebase search environment enabling LLM agents to navigate and modify repositories via sandboxed Bash tools (awarded $500 open-source bounty).
- Developing RL training environments for evaluating agentic capabilities across browser automation, codebase navigation, and constrained memory scenarios.
- Researching long-horizon reasoning via RL-trained active state compression, training LLMs to maintain coherent state without lossless context windows.

nano-slime | github.com/ascl1u/nano-slime March 2026 – Present

- Built a 4-process (Actor → Workers → Buffer → Trainer) asynchronous RL pipeline using Redis Streams to decouple GPU generation from CPU/Docker sandbox evaluation.
- Reimplemented GLM-5's GRPO loss (Equation 1) from scratch in PyTorch, featuring asymmetric clipping, IcePop MoE masking, and a replay buffer with staleness filtering.
- Increased pipeline throughput by overlapping environment execution with LLM generation, validating performance gains against a synchronous baseline via micro-HumanEval tasks.

MangaFuse | mangafuse.com August 2025 – October 2025

- Architected and deployed a production AI pipeline serving 10+ active users, automating scanlation workflows from hours to minutes.
- Built an asynchronous job queue with Redis and Docker to decouple heavy GPU inference from the FastAPI backend.

Technical Skills

Languages: Python, C++, TypeScript, SQL
ML/AI: PyTorch, Hugging Face, vLLM, TensorRT, Ollama, Weights & Biases
Agent Systems: Playwright, LlamaIndex, OpenCV, FastAPI
Infrastructure: Docker, AWS, Ray, Redis, PostgreSQL, ChromaDB
Frontend: React, Next.js