

Hank Sha

San Jose, CA | (510) 320-1116 | hankssha@gmail.com | linkedin.com/in/hank-sha | hanksha.com | Github

EDUCATION

University of California, Santa Barbara

B.S. Statistics and Data Science

Goleta, CA

Sep 2022 – June 2026

EXPERIENCE

UC Riverside

Machine Learning Research Assistant

Remote

Jun 2025 – Mar 2026

- Established which graph-traversal strategies survive LLM reasoning at scale, the evidence behind second-author findings on node classification and graph unlearning, via an eval harness spanning 12+ experiments, datasets, and prompting regimes.
- Architected a single-graph prototype into an unattended benchmark suite over 10K+ topological graphs, removing manual supervision from every full-dataset run.
- Cut experiment runtime 75% and GPU memory 20%, keeping full-dataset runs on existing hardware, by profiling Hugging Face Transformers for KV-cache thrash and tokenizer overhead, then applying batched inference, quantization, and CUDA tuning.

ABVC BioPharma

Clinical Data Scientist Intern

Remote

Aug 2025 – Nov 2025

- Translated FDA protocol constraints and Phase II endpoints into production modeling workflows for an MDD trial, contributing independently within 2 weeks of zero clinical background.
- Delivered a patient-dropout risk model at 83% recall (10-fold CV) from 15+ engineered clinical features, SMOTE on a imbalanced cohort, and XGBoost selected on held-out folds.
- Recommended against deployment and surfaced a data-leakage defect in my own pipeline after cross-fold variance exposed an underpowered dataset, confirming unreliable signal.

GitRoll

AI/ML Engineer Intern

San Francisco, CA

Sep 2024 – Apr 2025

- Shipped a RAG pipeline serving 45K+ users, owning the GitHub-analysis path end to end against Claude, GPT, and Gemini APIs to produce the product's core developer metrics.
- Drove PR-classification precision to 92% and throughput 20%, shipping the NLP impact-scoring system as a core production metric, by replacing a word-bank heuristic with NLTK categorization against 216+ hand-labeled issues.
- Cut LLM inference spend 13% by benchmarking the three models on cost-versus-latency, then applying prompt compression and batched inference.

PROJECTS

Xeno CLI — open-source auto coding workflow

2026 – present

Python · LangGraph · Docker · Tree-sitter

- Orchestrated 6 specialized agents through a LangGraph state machine with Docker sandboxing and Tree-sitter parsing, enabling autonomous multi-step coding fully locally.
- Built prompt caching across system prompts and codebase maps with hybrid local/API routing, keeping source code off third-party APIs.

Biotech Intelligence Engine — data platform for biotech-catalyst trading

2026 – present

Python · FastAPI · Celery · TypeScript · Docker

- Unified 137K+ entries from 5 heterogeneous sources: SEC filings, trial registries, insider transactions, financials and news into one data feed on FastAPI, Celery, and GitHub Actions CI/CD.
- Ran a small language model fully in-browser with MedRAG retrieval over drug mechanisms and trial outcomes, eliminating per-query API cost. Supabase and Upstash for auth and state.

Global PID Steering — undergraduate senior thesis

2026

Python · PyTorch · Transformers · GCP

- Collapsed 26 layer-specific PID controllers into one averaged refusal vector, cutting compute and storage 96% with no loss in coherence or jailbreak-defense efficacy on Gemma-2.
- Validated the simplification under adversarial attack with no regression on benign refusals, automating paired adversarial/benign benchmarking on GCP GPUs.

TECHNICAL SKILLS

Languages: Python, SQL, TypeScript, R, bash

Systems: Docker, AWS, GCP, CI/CD (GitHub Actions), FastAPI, REST APIs, Redis, Vercel

Data: Postgres, Supabase, MongoDB, Pinecone, Neo4j, Pandas, NumPy

AI/ML: RAG, LangChain, LangGraph, agent orchestration, evals, prompt engineering, vector databases, PyTorch, Hugging Face, XGBoost, scikit-learn