JOHNNY HUANG

San Mateo, CA, 94403 | (650)-278-6570 | h.johnny@wustl.edu | Portfolio | Linkedin | GitHub

EDUCATION

Washington University in St. Louis — December 2025

- *B.S/M.S:* **Mathematics + Computer Science** — GPA: *3.9*

PROFESSIONAL EXPERIENCES

Software Engineering Intern — *June – Nov. 2025*

City and County of San Francisco — *San Francisco, CA*

- **Scaled Flask**-based webserver dashboard used by city admins **from 20 to 1000+ concurrent users**; upgraded to a distributed workflow via **Gunicorn** on **Linux** w/ load balancing; offloading CPU-heavy tasks to a **Celery** Message queue to ensure **100% server up time;** resulting server is stable for **500+ concurrent RPM**; designed robust **python** mutex locks for STA processes.
- Deployed a **Redis** cache in backend for heavy DB queries; configured HTTP headers to stash static assets client side; **reducing LCP time by 70% to 400ms** on page loads; upgraded to an async model in **JS** frontend to **eliminate all UI freezes.**
- Optimized **SQL** database queries & managed scoped sessions to **alleviate pool exhaustion + thread safety issues**; established new read-only DB replica for optimized throughput reads & DB availability.

Machine Learning Intern — *Jan. – Apr. 2025*

Thermo Fisher Scientific — *San Francisco, CA*

- **Designed a LLM framework** for generating concise impressions for complex forensic cases; quantized Qwen-235B to 8-bits & performed PEFT w/ **HF's** LoRA; optimized GPU memory of pipeline w/ **CUDA**; achieved **training speedup of 400%** and **memory reduction of 50%;** impressions helped accelerated San Francisco's forensic case **turnaround time by 15% MoM.**
- Built a reusable training repo w/ **HF** & **PyTorch** for **validating key business records** sent to permanent storage; image-based (OCR) validation using OpenAI's CLIP; resulting workflow is **100% unsupervised** & filters **99% of false negatives**.

Large Language Models Intern — *May – Aug. 2024*

Rad AI — San Francisco, CA

- Led developments of a full-stack **search feature** using a **RAG framework on 300k+ internal patient records**; stored embedded documents in a **Pinecone** vector database w/ **Langchain** pipeline for optimized retrieval; **Pydantic** for API validation.
- Established robust **Microservices** for inference using **FastAPI**; deployed through **Docker** containers & managed on **Kubernetes** pods; achieved real-time **inference speed of only 2000ms per request.**

PERSONAL PROJECTS (See More!)

- **FTX: A secure HFT platform to trade** Johnny and Tofu coins; designed my take of a binary b-tree storage engine in **Rust** + **C**; coded a raw 4kb-aligned **pager** that bypasses the OS; maps bytes directly from SSD to program RAM; my pager is **10,000x faster** in data transfers v.s. standard library; WAL streamed to **Kafka** for instant returns & crash resistance.
- **Jarvis: Personal Genie** linked to your Google account; processes your Drive+Calendar data to **answer detailed personal queries**; **SpringBoot** server; Google API for OAuth + file crawling across docs, pdfs, even videos; RAG pipeline w/ **LangChain**, **Qdrant** vector DB, OpenAI API for function calling + multimodal embedding + response; frontend in **React/Typescript/Vite.**
- **Watchparty 2.0:** **Golang** webserver for **hosting shared content**; setup go routines & channels in a **MPSC** model to ensure data integrity on writes; **JS** websockets for real-time syncing up to **50+ users;** MRU **Redis** cache in RAM for access speed.

TECHNICAL SKILLS

- **ML:** PyTorch, TensorFlow, HuggingFace, LangChain, PySpark, Pinecone, Qdrant, CUDA, Pandas, Pydantic
- **Backend:** Java, Python, Rust, Node.js, C++, C#, C, .NET, Golang, SQL/NoSQL
- **Frontend:** React, TypeScript, Vite, JavaScript, HTML/CSS, Vue
- **Frameworks:** Socket.io, Spring Boot, FastAPI, Flask, Celery, AWS, Docker, Kubernetes, MongoDB, Redis, Bash, Git