JOHNNY HUANG

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

EDUCATION

Washington University in St. Louis Aug 2022 - December 2025

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

PROFESSIONAL EXPERIENCES

Software Engineering Intern *June 2025-Current*

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

- **Scaled Flask**-based dashboard used by all city admins **from 20 to 1000+ concurrent users**; upgraded to distributed server via **Gunicorn** on **Linux**; designed custom RR load balancer; offloading CPU tasks to **Celery** Message queue to ensure **100% server up time;** resulting server is stable for **500+ concurrent Req/sec**; designed robust **python** mutex for STA processes.
- Deployed **Redis** in backend to track stateful variables; **caching** heavy DB queries & **offloading** from large session cookies; configured **HTTP** headers to stash static assets client side; **drastically reducing LCP load time by 90% to 400ms** on page loads; conformed to async model in **JS** frontend to **eliminate all UI freezes.**
- Optimized **SQL** queries & managed **Flask alchemy** scoped sessions to **alleviate pool exhaustion & thread safety issues**; introduced new read-only DB replica for greatly increased **throughput reads & DB availability.**

Machine Learning Intern (Gap Year) *Jan. – Apr. 2025*

Thermo Fisher Scientific *San Francisco, CA*

- **Designed a multi-modal LLM** for generating summaries for complex forensic cases w/ Narratives + QTOF Lab data; 8-bit quantized Qwen-235B & performed LoRA w/ **HF**; optimized training w/ **CUDA**, batching and dataloaders; **reduced training time by 75%** & **RAM memory use by 50%;** project accelerated SF's case **turnaround times by 15% MoM.**
- Built a PnP training repo w/ **HF** & **PyTorch** for **validating key business records** sent to permanent storage; used **OCR** via OpenAI's CLIP; replaced previous manual workflow w/ **fully unsupervised** pipeline; 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** APIs for inference using **FastAPI**; deployed with **Docker** containers & managed on **Kubernetes** pods; achieved real-time **inference speed of only 2000ms per search.**

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 trades & server 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