JOHNNY HUANG

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

EDUCATION

Washington University in St. Louis (*St. Louis, MO*) 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+ daily users**; coordinated workers & threads via **Gunicorn** on **Linux** w/ load balancing; resulting server is non-blocking & **stable for 500+ concurrent RPM**; designed robust **python** GIL locks for STA processes; offloading CPU tasks to a **Celery** Message queue ensuring **100% server up time.**
- Deployed a **Redis** cache in backend for heavy DB queries; configured HTTP headers to stash static assets client side; drastically **reducing LCP time by 70% to 400ms** for page loads; upgraded to async model in **JS** frontend that **eliminated all UI freezes.**
- Optimized **SQL** database queries & managed scoped sessions to **alleviate pool exhaustion + thread safety issues**; established new read-only DB replica for higher throughput reads & **100% 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 SF'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 on **AWS** lambda instances; achieved real-time **inference speed of 2 s per transaction.**

PERSONAL PROJECTS (See More!)

- **Jarvis:** **Personal Genie** linked to your Google account; processes your Drive+Calendar data to **answer detailed personal queries**; **SpringBoot** server; Google API for OAuth + auto 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 for a fully concurrent model; **JS** websockets for real-time syncing up to **50+ users;** MRU **Redis** cache in RAM for instant access.
- **HalluAgent:** Developed a proprietary framework for **detecting hallucinations in GPT-3.5 utilizing SLMs**; trained SLMs as agents to evaluate LLM response via functional API's (ie. Calculator, google maps, etc); tuned w/ **HF Trainer** library.

TECHNICAL SKILLS

- **ML:** PyTorch, TensorFlow, HuggingFace, LangChain, PySpark, Pinecone, Qdrant, CUDA, Pandas, Pydantic
- **Backend:** Java, Python, Rust, Node.js, 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