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*

- *B.S/M.S:* **Mathematics + Computer Science** *Aug. 2022 – May 2026*
- Honors and Activities: Taylor Scholar (full tuition), Chancellor’s Fellow, 3x WashU Hackathon (2023 Organizer) *GPA:* ***3.9***
- Relevant Coursework: Machine Learning, Object Oriented Design, Advanced DSA, System Design, Bayesian Statistics, Optimization

PROFESSIONAL EXPERIENCES & INVOLVEMENT

Software Engineering Intern *June – Sep. 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 light-weight **Redis** cache in backend for heavy DB queries; configured HTTP headers to stash static assets client side; drastically reducing Largest Content Paint by **70%** to **<400ms** for page loads; introduced async model in **Javascript** frontend that eliminated all UI freezes.
- Optimized **SQL** database queries & managed DB sessions to alleviate existing pool exhaustion issues; established a new read-only DB replica for higher throughput reads & **100%** DB availability.

Machine Learning Intern *Jan. – Apr. 2025*

Office Of The Chief Medical Examiner *San Francisco, CA*

- Designed a LLM framework for generating concise impressions for complex forensic cases; applied 8-bit quantization to Qwen-235B & performed PEFT w/ **HF’s** LoRA; optimized GPU memory of pipeline w/ **CUDA**; achieved training speedup of **400%** and memory reduction of **50%**; 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 lightning-speed async **Microservices** for inference using **FastAPI**; deployed through **Docker** containers on **AWS** lambda instances; achieved real-time inference speed of **<2000ms** per transaction.

PERSONAL PROJECTS (More on my Portfolio!)

- **Social Link:** **Golang** webserver for hosting any shared content; instantly accessible by all users; setup go routines & channels for a fully concurrency-safe 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, maps, etc); tuning done w/ **HF Trainer** library
- **Reflective:** Full-stack desktop app for writing diaries; built w/ **.NET** MVC framework in **C#**; setup Rest APIs and a dynamic UI w/ **React** that automatically reminds users daily; stored data in a **NoSQL** JSON document store; deployed on **AWS** ec2**.**

TECHNICAL SKILLS

- **ML:** PyTorch, TensorFlow, HuggingFace, Pydantic, Langchain, PySpark, Pinecone, CUDA, Pandas
- **Backend:** Java, Python, Rust, NodeJS, C++, Socket.io, SQL, NoSQL, C#, .NET, Golang
- **Frameworks:** Springboot, FastAPI, Flask, Celery, AWS, MongoDB, Docker, Kubernetes, Git, Bash, Redis, Microservices