Zhenyu (Curtis) Lin

Software Development Engineering
175 Frankfort St Daly City, CA 94014 | (415)-794-5746 | US Citizen
zhenyulin.cs@gmail.com | www.linkedin.com/in/zhenyu-lin/ | www.zhenyulincs.com

Technical Skills

Languages: Java, Python, TypeScript, JavaScript, MySQL
Web Techstack: ReactJS, NextJS, Spring, Django, Flask, Tailwind CSS
AI Frameworks: Langchain, HuggingFace, PyTorch, TensorFlow, Keras, Sklearn, OpenCV, SpaCy, NumPy, Pandas
Developer Tools: Docker, AWS, Git, GitHub, Linux, Makefile, Clang

Experience

Software Engineer — Jun. 2024 – Aug. 2024

The Mobile and Intelligent Computing Lab (Supported by National Science Foundation) — *San Francisco, CA*

- Built a **ReactJS frontend** to display real-time outputs from the fine-tuned LLAMA3 model. Utilized **WebSockets** for continuous communication and **localStorage** to persist query history, enabling **context-aware responses** and reducing repetitive queries
- Developed **Restful APIs** to serve outputs from the fine-tuned **LLAMA3 model**, using **asyncio** to handle multiple requests simultaneously and **ThreadPoolExecutor** to process model inference in parallel, reducing model inference latency by 25%
- Cached frequently retrieved documents using **Redis**, applying an **LRU (Least Recently Used)** eviction strategy, removes the least accessed data to free up cache space, speeding up the retrieval step in the **Retrieval-Augmented Generation (RAG)** process, minimizing latency from data fetches by 40%

Machine Learning Engineer — Jun. 2023 – Aug. 2023

The Mobile and Intelligent Computing Lab (Supported by Sony) — *Hybrid*

- Coordinated a team to organize research findings and develop data visualizations, draft technical writing, which resulted in a paper published at the IEEE conference
- Compressed a DL Convolutional neural network (CNN) algorithm by 85%, shrinking the baseline model size from 463kB to 73kB with less than 1.5% accuracy drop through **8-bit Quantization**
- Implemented a CNN-based Bionic Arm control on a resource constrained Sony IoT edge device with 1.5MB sRAM, achieved 85% accuracy and 160ms clinical-grade control latency through **system memory management** and **multi-core parallel processing**
- Accelerated sampling rates of async myo-electric signal data streams by over 200% on an Android device by conducting rigorous, iterative **runtime profiling** and **data structures optimization** in Java

Backend Developer — Sep. 2022 – Dec. 2022

Senior Design Project — *San Francisco, CA*

- Led a team of 5 developers, improving efficiency through **task decomposition** and **continuous feedback loops** and winning 'Best Project' in the class competition against 6 other teams
- Configured CI/CD pipelines using GitHub Actions with multi-step workflows for build, test, and deployment stages. Integrated **JUnit** for Java unit testing with **annotations** and **test lifecycle management**, using **parallel job execution** and **matrix builds** to test across different environments
- Deployed the application using Docker on **Amazon Web Service** (AWS), configuring **AWS Elastic Load Balancer** with a **round-robin strategy** to distribute traffic across containers, reducing response times by 20%
- Implemented database indexing techniques using SQL, including **B-tree** and **Full-text indexing**, on key tables, improving query performance by 40% and reducing data retrieval latency for high-traffic operations

Education

San Francisco State University — *San Francisco, CA*
Master of Science in Electrical and Computer Engineering — Aug. 2023 – Dec. 2025
Bachelor of Science in Computer Science — Aug. 2019 – Jun. 2023