John Zhang

217-200-3540 | johnzhan@andrew.cmu.edu | linkedin.com/in/haojiong-zhang | github.com/HaojiongZhang

Education

Carnegie Mellon University — Pittsburgh, PA
Master of Science in Electrical and Computer Engineering — *Dec 2026*

University of Illinois at Urbana-Champaign — Champaign, IL
Bachelor of Science in Electrical and Computer Engineering — High Honors — *May 2024*

Experience

Machine Learning Intern — Sep 2024 – Dec 2024
Hangzhou Lianhui Numeral Technology — *Hangzhou, CN*

- Accelerated inference of multimodal LLMs (real-time text/audio to audio conversion) by 75% using sglang for enhanced processing speed in industrial applications
- Developed a hybrid datastore architecture for efficient data organization and retrieval for LLM long term conversation memory, employing techniques such as BM25, vector similarity, graph, and key-value store
- Developed a knowledge base querying system for various factory plant documents (text, Excel, pictures) using LLM, RAG, and OCR:
- Deployed and enhanced multi-modal Retrieval-Augmented Generation techniques on top of RAPTOR and GraphRAG to achieve top1 recall accuracy of 95%
- Created a natural language Excel query system by fine-tuning and benchmarking text-to-SQL Large Language Models, achieving SoTA performance on BIRD and SPIDER datasets

Software Engineer Intern — Jun. 2024 – Sep. 2024
Yuansuan Technology — *Hangzhou, CN*

- Deployed a cloud compute platform using k8s, Golang, and Rancher for machine learning training and big data processing to run workload for AI4Science
- Implemented multi-job scheduling framework for machine learning based on Huawei Volcano and MySQL for failure recovery and reduce resource waste
- Developed REST APIs for permission verification, file transfers, CRUD and multipart uploads to NFS with Redis for efficient progress tracking and recovery
- Built a KServe-based ML serving runtime for PyTorch, TensorFlow, scikit-learn, and ONNX model types
- Created a docker image management system for uploading, viewing, and downloading images using k8s daemonset

Machine Learning Co-op Intern — Aug 2023 – Dec. 2023
Advanced Micro Devices — *Champaign, IL*

- Recreated and benchmarked state-of-the-art voice source isolation and de-noising autoencoder models using PyTorch and HuggingFace for AMD's client product architecture team
- Constructed data mixing pipeline for different noise levels and multiple source separation tasks
- Built audio denoising models from DeeplabV3 architecture with high scale-invariant signal-to-distortion ratio on compatible with AMD AI engine

Quantitative Developer Intern — Jun. 2023 – Aug. 2023
Deep Data Investments — *Shenzhen, CN*

- Implemented a parallel back-testing environment with scalable visualization tool, exceeding performance of the external option by 20%, saving $180k per year
- Performed data analysis, and feature extraction on 40 GB of intra-minute options data using Pandas and NumPy
- Engineered mean reversion and momentum trading strategies with high Sharpe ratio
- Optimized XGBoost models with hand-crafted features using Scikit-learn and PyTorch, and tuned indicators for better performing trading signals, outputting high-return correlation with 1/5 of original model size

Skills

Programming Languages: Golang, Python, C/C++, SQL
Developer Tools: Git, Docker, GDB, Kubernetes, Postman
Frameworks/Libraries: Pandas, NumPy, Matplotlib, Pytorch, Huggingface, Kubernetes, Kserve, MySQL
Coursework: Computer Systems, Networks, Distributed Systems, Deep Learning, Signal Processing, Generative Models