Jiaxuan Huang

Jersey City, NJ jiaxuanhuang2003@gmail.com github.com/jiaxuan030331

Education

New York University *Sep 2025 – Present*
M.S. in Data Science (Center for Data Science)

University of California, Los Angeles *GPA: 3.91/4.0* *Jun 2023 – Dec 2024*
B.S. Mathematics/Economics, Computing Specialization

CS Core: Data Structures, Algorithms (C++/Python), OOP, Unix/Linux etc.

Technical Skills

ML/Serving:	NLP, PyTorch, Hugging Face, CTranslate2, LoRA fine-tuning, (semi-)streaming ASR, evaluation (WER/CER), profiling (torch.profiler), CUDA, RAG
Backend/Infra:	FastAPI, WebSocket, gRPC, Redis, Docker, Git, CI/CD, CRON, MySQL
Languages:	Python, C++, SQL, Java, R, Bash
Theory & Methods:	NLP, Bayesian Probabilistic modeling, Time series, Stochastic Optimization

Experience

NLP Development Intern *Ricoh Research Institute*
Mar 2025 – Aug 2025

- Developed and deployed a **low-latency ASR service** using **FastAPI + WebSocket semi-streaming**, achieving **30% CER reduction**, **P95 latency ¡500ms**, and **lower GPU memory usage** via CTranslate2 optimization.
- Fine-tuned multilingual ASR models with **LoRA** and integrated a custom **language identification head**; enabled robust Cantonese/code-switching support and reduced Cantonese CER from ˜40% to ˜15%.
- Modularized recognizers and deployed via **Tornado** microservices with Docker; enabled dynamic model switching and gray rollout for scalable production deployment.
- Built an **ASR–TTS closed-loop** evaluation across 5+ languages; automated WER/CER benchmarking and monitoring with Prometheus/Grafana to guide iteration and deployment.

Data Analyst Intern (Remote) *Uber (Hong Kong)*

- Delivered analytics on 100M+ orders (clustering, anomaly handling) and shipped edge deployments for a MobileNet-based mask model.

Asset Management *Haitong Securities*

- Built multi-factor portfolio analyses and industry research reports; automated data workflows using Python/SQL

Projects

Multilingual ASR Router https://github.com/jiaxuan030331/China-Multi-Lingual-ASR-System

- Designed and open-sourced a **router architecture that unifies heterogeneous ASR models (Whisper & Kimi) at the encoder level**, enabling synchronized decoding across models with minimal GPU memory usage.
- Built a **custom language identification (LID) module with confidence thresholds**, dynamically routing Mandarin/English to Kimi and dialects to fine-tuned Whisper; implemented fallback logic for robustness in mixed/uncertain scenarios.
- Optimized inference with **CTranslate2 encoder compression and semi-streaming WebSocket pipeline**, consolidating GPU usage from **23GB + 6GB across models to ˜24GB total**, with slightly lower latency vs. baseline and stable throughput under high concurrency.
- Open-sourced core components (modular decoders, LoRA fine-tuning scripts, **deployed backend REST + WebSocket**); in-progress **benchmark pipelines** and **full Docker packaging** documented as roadmap.

ARC Solution (Kaggle Silver Prize) https://github.com/jiaxuan030331/China-Multi-Lingual-ASR-System

- Built a multi-strategy ARC solver (40+ specialized algorithms) with intelligent solver selection and C++17 backends via pybind11; placed 25th/1431, delivered ˜1s avg/task and 4–46× speedups over Python.
- Implemented a DAG-based transformation engine plus tiling/symmetry/chess solvers; benchmarked at 92% on complex DAG tasks and 94% combined across tasks; shipped a production-grade Python/C++ package (CMake, tests, examples).