Jiaxuan Huang

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

EDUCATION

New York University Sep 2025 – Present
- M.S. Data Science (Center for Data Science)
University of California, Los Angeles Jun 2023 – Dec 2024
- B.S. Mathematics/Economics, Computer Science
Specialization GPA:3.91/4.0

TECHNICAL SKILLS

ML Platforms/Serving: PyTorch; TensorFlow; Hugging Face; Numpy; Scikit - learn; Kubernetes; A/B testing;

Backend/Infra: Pybind 11; Docker; MySQL; FastAPI; CI/CD; CRON; WebSocket; ONNX runtime; CTranslate2

Domains: NLP; Machine Learning (Bayesian, [Full/Self/Semi/Weak] supervised, unsupervised); RAG; Computer Vision

Proficient Languages: Python; C++ 17; SQL; Bash; R

EXPERIENCE

Machine Learning Engineering Intern Feb 2025 – Sep 2025
Ricoh Research Center (Natural Language Processing Team)

- Improved Speech Recognition (ASR) model Character Error Rate (CER) of Cantonese ~38% to ~15% (no multilingual/time-alignment forgetting) by building a config-driven training pipeline on AWS with Hugging Face (Transformers/Datasets) and LoRA finetuning with experimented data mixture/augmentation.
- Addressed under-represented language detection problem of Whisper (LLM by OpenAI), achieved 99.4% accuracy with < 50 ms inference across Mandarin/English/Cantonese/other by customizing and training a compatible language router with tunable uncertainty threshold fallback to default detector.
- Deployed speech recognition model with P95 ~800 ms latency and reduced VRAM ~7 GB vs. parallel deployment by designing an encoder-integrated, Language-ID routing architecture (major languages to Kimi[7B, multimodal LLM], others to Whisper [1.5B, CTranslate2 accelerated]), and exporting PyTorch models to ONNX Runtime/CTranslate2, tuning quantization, beam search, and batching (Open source assets replication available on Github).
- Shipped semi-streaming ASR service (FastAPI + WebSocket API, Built Docker; CI/CD on AWS); For team collaboration, defined SLOs/SLA, set up Prometheus/Grafana dashboards to streamline integration and support.

Data Scientist Intern Apr 2023 – Jul 2023
Uber (Hong Kong)

- Quantified loss-savings impact of long-distance business-trip fraud on 2M+ trips/day by building historical backtests and factor benchmarks in Jupyter Notebook.
- Enabled offline evaluation/monitoring of incident rate, adjustments, and loss-savings pressure by defining metrics, engineering model-ready datasets with prototype ETL (SQL/Hive), and shipping Looker dashboards.
- Launched risk strategies across 2 products in collaboration with product team, preparing canary and A/B rollouts with production monitoring.

PROJECTS

Qwen VL 7B Multimodal Finetuning (video frames + text) July 2025 – present

- Built a reproducible video-text pipeline (8–12 frames/clip @ 1–2 fps, transcript alignment) packaged to HF Datasets/Parquet with lazy ffmpeg decode, length bucketing, and streaming loaders.
- Improved Qwen2-VL 7B fine-tuning efficiency/stability with QLoRA; ran sweeps (lr, effective batch size, LoRA rank) and ablations (frames/clip, transcripts on/off) while preserving fluency.
- Demonstrated serving readiness (FastAPI) with an eval harness (EM for Q&A/summaries)

Kaggle ARC featured Competition (Silver Prize, Top 2%, Solution on Github) June 2024 – November 2024

- Leading a team of 5, implemented and open-sourced a multi-strategy Directed Acyclic Graph (DAG) Aggregated Solver (9 sub-solvers, 5,000+ lines) with C++17ckends via pybind11; Achieved 2–46× faster than Python.