Jiaqi Yang

Ottawa, ON | (343) 462-2832 | jyang297@uottawa.ca | LinkedIn | GitHub | Portfolio

Summary

AI/ML Engineer with experience in LLM routing, RAG systems, and recommendation pipelines, with additional background in deep learning research and cloud deployment.

Education

University of Ottawa — Ottawa, ON
Master of Applied Science in Electrical and Computer Engineering — *Sep. 2022 – Apr. 2025*

Hefei University of Technology — Hefei, China
Bachelor of Engineering in Optical and Electrical Engineering — *Sep. 2018 – Jun. 2022*

Technical Skills

ML/LLM: Transformers (HF), LLaMA, FLAN-T5, BERT, DSPy, vLLM, SGLang, RAG, LightGBM, Two-Tower RecSys
Infra & MLOps: AWS (EC2, S3), GCP, Docker, Terraform, Kafka, ClickHouse, Redis, Qdrant, CI/CD, Prometheus, Grafana
Programming: Python (PyTorch, Scikit-learn, SpaCy), Go, TypeScript, C/C++, SQL, MATLAB

Professional Experience

AI Engineer | Staff Rodeo — *Sep. 2025 – Oct. 2025*
- Designed and implemented a **two-tower recommendation model** for candidate–job matching, using semantic embeddings for more accurate similarity scoring.
- Built a **low-latency retrieval pipeline** with Qdrant supporting sub-50ms search across thousands of profiles.
- Developed a **reranking agent** to refine candidate relevance, improving top-k match quality and strengthening overall recommendation reliability.

Machine Learning Engineer | RemoBytes — *Jun. 2025 – Sep. 2025*
- **LLM Routing Platform:** Built an intelligent routing system across OpenAI, Anthropic, and vLLM, achieving a **40% cost reduction** through ML-based query classification and 3:1 context compression.
- **Real-Time Data Pipeline:** Engineered a Kafka + ClickHouse streaming pipeline processing **10k+ queries/day**, with sub-100ms response time enabled by Redis caching (65% hit rate).
- **Recommendation System:** Developed a YouTube Shorts–style engagement prediction model using LightGBM and temporal feature engineering, reaching **0.85 AUC**.
- **Production Deployment:** Shipped Dockerized microservices with CI/CD automation and end-to-end monitoring via Prometheus and Grafana for 1000+ enterprise users.

AI Lead (Part-Time) | Health STORIA — *May 2024 – Present*
- Led the development of a physician-facing AI agent, translating complex clinical use cases into precise LLM requirements using **Chain-of-Thought (CoT)** prompting.
- Implemented a **Retrieval-Augmented Generation (RAG)** pipeline over medical textbook embeddings, significantly reducing hallucinations and enhancing response accuracy.
- Orchestrated end-to-end cloud deployment on AWS (EC2, S3), delivering a fully functional MVP for beta testing within 4 weeks.

Data Scientist Intern | EchoPlus AI LLC — *Apr. 2025 – May 2025*
- Fine-tuned LLMs (LLaMA, FLAN-T5, BERT) for summarization and QA tasks, utilizing **RLHF** pipelines with Proximal Policy Optimization (PPO) and LoRA parameters.
- Constructed end-to-end training workflows using the **Hugging Face Trainer API**, implementing dynamic padding and efficient batching to optimize GPU utilization.

Academic Projects & Research

Video Frame Interpolation Research | *Research Assistant, uOttawa* — May 2023 – Dec. 2024
- Proposed a novel temporal pyramid structure with LSTM modules for Video Frame Interpolation, integrating a modified U-Net for feature fusion.
- Implemented **Knowledge Distillation** from a ground-truth teacher model to a student model, significantly improving motion continuity and object structural integrity.
- Achieved a **PSNR of 35.78** on the Vimeo90K dataset, outperforming standard baseline models.
- Collaborated with **Ross Video** R&D team via weekly technical seminars to present research progress and align model architecture with industrial requirements.