Yuzheng Cong

+ 1 (607) 327 1776 | yc2838@cornell.edu | linkedin.com/in/yuzheng-cong

EDUCATION

Cornell University
Master of Science in Computer and Information Science — Aug 2024-Dec 2025

The Chinese University of Hong Kong
Bachelor of Engineer in Computer Engineering — Sep 2020-Jun 2024

TECHNICAL SKILLS

Machine Learning: LLM inference, SGLang, vllm, NLP, LLMs, RAG, Model evaluation, fine-tuning (QLoRA),
Programming: Python, TypeScript, Go, C++, SQL, R, CUDA
ML & Data Tools: PyTorch, FAISS, Neo4j, scikit-learn, Pipecat, Distributed Systems, Real-time & Event-Driven Architecture
Miscellaneous: Git, Docker AWS EC2, CI/CD workflows, RESTful API, Google Colab, Pytorch, Jira, Hugging Face

WORK EXPERIENCE

MEITUAN — Shanghai,China
Machine Learning Engineer Intern — May 2025 – Aug 2025

- Built and deployed a **production-grade** voice ordering SaaS system used by **10+ live restaurants**, owning backend service logic, LLM orchestration, and cloud deployment; achieved 75% successful order completion during peak hours.
- Deployed and optimized a 70B-parameter **LLM inference service** on AWS EC2, using **SGLang** for **prompt preloading and controlled decoding,** reducing first-token latency by 200ms and improving production responsiveness.
- Revolutionized order processing through a multi-agent LLM-based order system integrating LLM function calling coupled with advanced **prompt and context engineering**, resulting in a 10% increase in successful complex order completion.
- Pioneered an **allergy-aware recommendation** engine leveraging a **Neo4j knowledge graph,** integrating hybrid search with dish decomposition logic; improved menu accuracy by 12% and slashed query latency by 210ms.

COTIVITI.INC — Atlanta, GA (Hybrid)
Machine Learning Engineer Intern — Jan 2025 – Apr 2025

- Orchestrated graph-based **retrieval-augmented generation**(RAG) framework utilizing semantic chunking and knowledge graph, enhancing AI-driven medical fraud detection capabilities while reducing LLM hallucinations by 18%.
- Fine-tuned a **Llama-3.1-8B language model** using **QLoRA** for medical QA, improving classification accuracy to 79.4% on MMLU (medical subset) and increasing symptom recognition accuracy by 15%.
- Developed an **interactive evaluation dashboard** to analyze embedding strategies and **chunking configurations**, accelerated identification of optimal configurations, enabled rapid iteration/analysis to boost model tuning efficiency.

Shanghai Yijingjie Information Technology Co., Ltd. — Shanghai,China
Database System Engineer Intern — May 2023 – Aug 2023

- **Contributed to Greenplum (Open-source contribution)**: extended ORCA optimizer type support by implementing constraint propagation logic from the intermediate representation (DXL) to the internal expression tree for custom Domain types, resolving issues where NOT NULL and other constraints were ignored during query optimization.

PROJECTS

AI & Law: Madison Bookshelf | ***Cornell University***

- Developed an **AI-powered semantic search** and QA system over James Madison's personal library, enabling **context-aware** question answering and citation generation for legal and historical research.
- Implemented s**emantic retrieval pipelines** using OpenAI embeddings over **large-scale digitized text corpora**, and optimized retrieval for historical terminology alignment, improving query accuracy by 40%.

AI Agent for Commerce Website | ***Independent Project***

- Designed and deployed an **LLM-based commerce agent** supporting conversational assistance, product recommendation, and multimodal (text + image) product search within a unified architecture.
- Implemented text-based retrieval via embedding similarity over a **FAISS vector** store with dynamic thresholding and top-k ranking, and image-based retrieval through GPT-generated textual descriptions.
- Configured **CI/CD pipeline** using **GitHub Actions** and **Docker** for automated testing and deployment to **Render**.