Cheng Jiang

(+86)18128690749 | 2209878352@qq.com | chengz2003.github.io | github.com/ChengZ2003

Education

Sun Yat-sen University |***M.S. in Software Engineering***, *School of Software Engineering* Sep. 2026 – Jun. 2029 (Expected)

Wuhan Textile University |***B.Eng. in Data Science and Big Data Technology***, *School of Computer and Artificial Intelligence* Sep. 2021 – Jun. 2025

GPA: 3.89/4.0 (Rank: 1/67)

Selected Coursework: Deep Learning (95), Spark In-Memory Computing (90), Hadoop System Development (94), Python Data Analysis and Mining (96), Linux System Programming (99), Machine Learning (90), Database Principles and Applications (91)

Technical Skills

- **Programming Languages**: Python, Golang, Scala, Java, C/C++, SQL, Shell
- **Frameworks & Libraries**: FastAPI, Flask, LlamaIndex, LangChain, Apache Spark, Apache Hadoop, PyTorch
- **Tools**: Docker, Git, Google Colab, Visual Studio Code, PyCharm

Research Experience

Multimodal Retrieval-Augmented Generation for Financial Document Analysis May 2024 – Aug. 2024

- Proposed a **multimodal retrieval-augmented generation framework** for financial documents, integrating vision-language models (VLMs) for structured understanding of charts and tables.
- Designed a hybrid indexing mechanism combining **vector retrieval and knowledge graphs** to improve retrieval recall and answer generation accuracy for complex financial question-answering tasks.
- Constructed a **multimodal financial document dataset** containing text, tables, and charts, and performed annotation with large language models and manual verification.
- **First-author** paper: *Multimodal Retrieval-Augmented Generation for Financial Documents: Image-Centric Analysis of Charts and Tables with Large Language Models*, accepted by **The Visual Computer (SCI, JCR Q2, CCF-C)**.

Work Experience

Guangdong Singularity.ai |*Applied AI Engineer, Full-time* Dec. 2024 – May 2025

- Designed speech data processing workflows and training corpus construction pipelines for livestreaming scenarios, including ASR transcript cleaning and structured conversational data generation.
- Fine-tuned Qwen series models using LoRA to improve dialogue generation quality for e-commerce livestreaming applications.
- Participated in the deployment of multimodal speech generation systems, building ASR/TTS pipelines based on SenseVoice and CosyVoice.
- Designed and implemented offline deployment and scheduling mechanisms for large language model inference services, benchmarking vLLM and Ollama in terms of inference performance and GPU memory efficiency.

LinkTime Cloud |*Research Intern* Jan. 2024 – May 2024

- Participated in the design and implementation of an enterprise knowledge question-answering system based on LLM + RAG architecture.
- Evaluated multiple domestic LLM APIs in terms of response latency, context window size, and generation quality to support model selection.
- Investigated mainstream LLM application architectures and summarized RAG design patterns for enterprise knowledge QA systems.

Projects

LLM-based E-commerce Livestreaming System |*Guangdong Singularity.ai* Dec. 2024 – Mar. 2025

- Built a livestream data processing pipeline for transcription and text cleaning of over 1000 hours of speech data.
- Designed an LLM-based annotation workflow to automatically extract product selling points and promotional information.

- Fine-tuned Qwen2.5 (7B / 72B) models using LoRA for livestream script generation and style transfer.

Speech Processing and Speech Generation System | *Guangdong Singularity.ai* Feb. 2025 – May 2025

- Developed a speech processing pipeline based on SenseVoice and CosyVoice, supporting ASR transcription, voice cloning, and TTS generation.
- Designed timestamp alignment and audio segmentation algorithms to improve speech generation stability.
- Implemented multi-process inference deployment and optimized GPU memory usage and inference throughput.

Enterprise Retrieval-Augmented Generation System | *LinkTime Cloud* Feb. 2024 – Apr. 2024

- Developed an enterprise knowledge question-answering system based on LLM + RAG architecture.
- Designed a hybrid retrieval mechanism combining sparse and dense retrieval to improve retrieval recall.
- Built retrieval pipelines using LlamaIndex and implemented a web interface with Streamlit.

RAG Retrieval Optimization and Reranking Research | *LinkTime Cloud* Apr. 2024 – May 2024

- Introduced cross-encoder rerankers to address the limited semantic matching capability of bi-encoder retrieval models.
- Evaluated reranker performance using information retrieval metrics such as NDCG and MAP.
- Built reranking model services using Transformers and FastAPI.

Honors & Awards

First-Class University Scholarship Dec. 2022

Third Prize, Huawei ICT Competition Hubei Cloud Track Dec. 2022

National Scholarship, President Scholarship, Outstanding Student Award Dec. 2023

Third Prize, Huawei ICT Competition Hubei Cloud Track Mar. 2024

National First Prize, 12th Teddy Cup Data Mining Challenge (Undergraduate Division) Jun. 2024

Third Prize, FinTech Development Award (People's Bank of China) Nov. 2024

Outstanding Undergraduate Thesis Award Jun. 2025

CET-4, CET-6, IELTS 7.0 Mar. 2023, Jun. 2023, Feb. 2026