RUIFENG (ELLEN) LI

• Phone #: (213) 378 4611 • Email: ruifengl@usc.edu • Linkedin: https://www.linkedin.com/in/ellen-lee-285a7a373/ GitHub: https://github.com/Lrf8182

SUMMARY

Results-driven Computer Science Master's student with extensive experience in optimizing and deploying production-level AI systems, specializing in LLM pipeline engineering. Versatile experience across LLM pipeline engineering, Retrieval-Augmented Generation (RAG), computer vision research (VSR), and distributed training, supported by two second-author publications.

EDUCATION

Undergraduate: Shanghai University **Major:** *Computer Science* *09/2021-06/2025*
Graduate: University of Southern California **Major:** *Computer Science* *08/2025-05/2027*

PROFESSIONAL EXPERIENCE

OpenJobs AI, ML *Intern* San Jose *08/2024-03/2025*

- Developed an AI-powered system to generate personalized resumes and cover letters, improving text relevance and professionalism. Leveraging fine-tuned LLMs (DeepSeek, GPT-4) to power semantic search across 40+ dimensions. This system now serves as the core engine for personalized job recommendations. https://www.openjobs-ai.com/
- Optimized the generation pipeline by leveraging asynchronous processing for parallel LLM API calls, reducing end-to-end latency.
- Engineered a high-performance resume generation pipeline, migrating from a GPT-4o API prototype to a self-hosted, fine-tuned open-source model (Llama 3) served via vLLM. This strategic shift, combined with a 5-step prompt chain, incorporating Chain-of-Thought prompting in key analysis stages to improve rewriting accuracy, achieved a 67% reduction in generation latency (from 15s to 5s) and slashed operational costs by 30% by eliminating API token fees.

Yum China, IT Intern Shanghai 01/2024-05/2024

- Gained foundational experience in enterprise LLM applications by supporting the team on a Retrieval-Augmented Generation (RAG) system, with exposure to the full pipeline from data scraping and embedding to model evaluation.
- Developed a Proof-of-Concept (PoC) using Stable Diffusion and ComfyUI to explore automated marketing asset generation. Generated over 100 sample images from text prompts, demonstrating the technology's potential to the team for future application.
- Provided operational support for a customer satisfaction platform serving thousands of daily users.Utilized Docker within a Linux environment to monitor service status and assist in regular deployment checks, contributing to the platform's 99.8% uptime

RESEARCH EXPERIENCE

East China Air Traffic Management Bureau Meteorological Research Group *06/2023-07/2024*

- Proposed a model for meteorological super-resolution by fusing a VSR model's temporal attention with a UNet's spatial hierarchy via a custom route-selection module. Achieved a significant performance uplift, improving the PSNR by 1.5 dB and SSIM by 0.05 over established SOTA models (HAT).
- Conducted a benchmark of SOTA VSR models and successfully replicated a baseline paper to ground the research.

Cybersecurity Lab Vehicle Communication System Modeling and Simulation Group *03/2023-04/2025*

- Designed a Sequence Mapping Block (SMB) with One Hot encoding to map each frame's bitstream from vehicle ECU communication messages into a high-dimensional space, significantly improving model convergence.
- Provided distributed data-parallel training solutions, developed a compatible training pipeline for various anomaly detection architectures, and employed Ring-All Reduce, leading to a ~30% reduction in training time.
- Use deep learnin to detect anomalies. Enhance security analysis through a LSTM-based recurrent deep learning networks. Achieved a 5% higher F1-score in anomaly detection compared to the baseline model.

1. Q. Liu†, Y. Liu†, **R. Li**, C. Cao, Y. Li*, X. Li*, P. Wang, R. Feng, "MDHP-Net: Detecting an Emerging Time-Exciting Threat in IVN." 2025. (Preprint, Second Author) [Paper][Code]

2. Li, Xingyu, **Ruifeng Li**, and Yanchen Liu. "HP-LSTM: Hawkes Process–LSTM-Based Detection of DDoS Attack for In-Vehicle Network." Future Internet, vol. 16, no. 6, 2024, p. 185. (Second Author) [Paper][Code]

SKILLS

- **Language Skills:** English (TOFEL:105) , Chinese (Mother language) **Fundamental Languages**: Python, JavaScript, C++, SQL
- **ML & AI:** RAG, Transformers, Anomaly Detection, Fine-Tuning, VLLM, Distributed Data-Parallel Training, Ring-All Reduce,
- **Stack:** React, Node.js, Django, MySQL, PostgreSQL**,** Elastic Search, Docker, K8s, Github, Linux, Agent Management

HONOR&AWARDS

- 2024 National First Prize - National University Student Computer Application Competition 07/2024