Hanhui Su

Email: hsu053@ucr.edu Tel: +1 951 334 1937

EDUCATION

University of California, Riverside | Riverside, CA (Sep. 2024 – *Present*)
Master of Science in Computer Science

Beijing Information Science & Technology University | Beijing, China (Sep. 2019 – Jun. 2023)
Bachelor of Engineering in Internet of Things Engineering Ranked in the **top 5%** of the department
Relevant Courses: Large Models & Advances in AI, NLP, Deep Learning, Machine Learning, Trustworthy AI for Autonomy, GPU Architecture & Parallel Programming, High Performance Computing, Operating System, Data Mining Techniques
Awards & Honors: Dean's Honor List Level-I (Top 3%) for 2020 Spring, 2021 Spring, 2021 Fall, and 2022 Fall; Dean's Honor List Level-II (Top 10%) for 2022 Spring

EXPERIENCE

Institute of Software, Chinese Academy of Sciences | Beijing, China *LLM Intern* (May 2023 – Aug. 2023)

- Processed **30 GB** of unstructured Chinese eBook data (MOBI, EPUB) for LLM pretraining, contributing to rule-based noise filtering, hash-based deduplication, and structural normalization.
- Contributed to the ArkPDF **SFT dataset** for complex PDF understanding, generating and reviewing **8k** QA pairs with instruction templates for table extraction and layout-dependent reasoning.
- Added **validation checks** for schema, output format, table representation, and reading order to improve consistency of PDF QA training data.
- Evaluated Llama 2 outputs against ChatPDF and Bing Chat on citation hit rate, hallucination behavior, and retrieval recall.
- Ran **LoRA** fine-tuning experiments on Llama 2 for PDF parsing data under GPU memory constraints.

Beijing Saishu Technology Co., Ltd. | Beijing, China *Embedded Software Engineer Intern* (Jul. 2021 – Sep. 2021)

- Contributed to the development and launch of a Bluetooth debugging tool and optimized networking solutions via firmware updates, reducing device connection issues by **20%**.
- Participated in hardware integration, firmware installation, and stress testing for IoT devices, assisting in the deployment of **50 terminals** across multiple sites for research data collection.

PROJECTS

Evidence-Grounded RAG Agent for Financial Document Analysis

- Built a local-LLM-compatible document-analysis and QA agent with FastAPI, **LangGraph**, ChromaDB, and DuckDB for query understanding, retrieval, tool execution, evidence planning, and question answering over public company filings.
- Indexed 980 SEC filings into 88K+ filing chunks across **80 public companies**, combining ChromaDB vector search with 150K+ DuckDB financial facts and price/event data.
- Implemented an **evidence-planning** layer that maps user questions into target companies, analysis intent, required evidence, and tool calls before answer generation.
- Built deterministic **Python tools** for derived metric computation, dependency tracing, and provenance tracking, keeping numerical reasoning outside free-form LLM generation.
- Added runtime answer checks for **citation validity**, numeric grounding, missing-evidence disclosure, and non-advisory boundaries before returning responses.

Fin-Qwen: Reasoning-Augmented Distillation for Financial Sentiment in Social Media

- Fine-tuned Qwen3-8B with **QLoRA** using teacher-generated reasoning traces for financial social-media sentiment classification; ran local inference on a 12 GB VRAM GPU.
- Built a **distillation** pipeline that trained the student model to generate reasoning traces and sentiment labels for financial slang, such as "rug pull" and "diamond hands".
- Used **Unsloth** and **4-bit quantization** to run **QLoRA** fine-tuning within limited GPU memory.
- Added structured JSON output constraints and measured **99.6%** JSON-valid responses on the project evaluation set for downstream parsing.
- Evaluated against a zero-shot prompting baseline on the project test set; observed **Weighted F1** improvement from 0.38 to 0.89, **Pearson** $r = 0.84$, and **85% lower MAE**.
- Automated preprocessing, teacher-output generation, training-data preparation, and evaluation scripts to compare student outputs against teacher-generated labels and reasoning traces.

SKILLS

- LLM: PyTorch, Hugging Face Transformers/PEFT, LoRA/QLoRA, Unsloth, Distillation, vLLM.
- Agents / RAG: LangGraph, ChromaDB, DuckDB, FastAPI, MCP, tool calling, multi-agent workflows.
- Languages / Tools: Python, SQL, C/C++, Linux/Shell, Git, Docker, Pandas, NumPy, Pytest.