Wenjie Du

duwenjie24@gmail.com • Linkedin • Google Scholar • Hangzhou, China

Education

Sichuan University, Chengdu, China — Sept. 2020 – June 2024

B.Eng. in Computer Science and Technology, **Average Scores:** 89.4/100, **GPA:** 3.74/4.00

Research Interests

Large Language Models, Machine Learning System

Experience

Westlake University — Hangzhou, China

Research Intern in Huan Wang's Team — July 2025 – Present

- Integrated full and local attention with gating values into SGLang and AReaL, enabling RL-based identification of critical attention heads in reasoning LLMs.
- Identified a set of reasoning heads distinct from retrieval heads, achieving 20-50% KV cache reduction with near lossless performance compared to the uncompressed baseline across 4 benchmarks.

Hong Kong University of Science and Technology — Hong Kong SAR

Research Assistant in Xiaomin Ouyang's Team — Sept. 2024 – Mar. 2025

- Applied text semantic embeddings to guide model training and inference processes, improving cross-domain adaptation accuracy for Human Activity Recognition (HAR) problems through contrastive pre-training.
- Attempted to fine-tune LLMs to align with IMU time-series sensor data and textual labels using LLaVA and LLaMA-AdapterV2 architecture.

Institute for AI Industry Research, Tsinghua University — Beijing, China

Research Intern in Yuanchun Li's Team — Dec. 2023 – Aug. 2024

- As a core developer, designed and implemented AutoDroid-V2, a script-based GUI Agent, which pre-constructed APP API documentation, saving 90% token consumption and reducing latency by 93.1% compared to traditional methods.
- Designed a UI reassembling framework based on LLM, identifying important UI components and providing better visualization solutions as low-code schemes.
- Developed a low-cost, high-efficiency App exploration Agent LLM-Explorer, using LLM to memorize and explore APPs and improving coverage by 12% compared to traditional solutions.

ByteDance Ltd. — Shanghai, China

Golang Software Engineer Intern — May 2023 – Nov. 2023

- Developed features for a multi-source alert auto-processing platform using the Go language.
- Refactored system using Domain-Driven Design and Chain of Responsibility to enhance the code quality.

Selected Awards

- National Scholarship (Top 3/318) 2023
- Sichuan University Comprehensive First Class Scholarship(Top 4/318) 2023
- Sichuan University Excellent Student 2023

Publications

- **Wenjie Du**, Li Jiang, Keda Tao, Xue Liu, Huan Wang, "Which Heads Matter for Reasoning? RL-Guided KV Cache Compression" in **arXiv**, [page], [pdf]
- Shanhui Zhao, Hao Wen, **Wenjie Du**, Cheng Liang, Yunxin Liu, Xiaozhou Ye, Ye Ouyang, Yuanchun Li, "LLM-Explorer: Towards Efficient and Affordable LLM-based Exploration for Mobile Apps" in **MobiCom'25**, [pdf]
- Hao Wen, Shizuo Tian, Borislav Pavlov, **Wenjie Du**, Yixuan Li, Ge Chang, Shanhui Zhao, Jiacheng Liu, Yunxin Liu, Ya-Qin Zhang, Yuanchun Li, "AutoDroid-V2: Boosting SLM-based GUI Agents via Code Generation" in **MobiSys'25**, **Best Artifact Award**, [pdf]

Skills

Python, C, C++, Go, JavaScript, Linux, Git, LaTeX.