Qi(Muqi) Li

✉ muqi1029@gmail.com · ☎(+86)18756917985 · Personal Website · Muqi1029

INTRODUCTION

I am passionate about Large Language Model (LLM) training and serving infrastructure, with hands-on experience in building scalable systems. I am also an active contributor to the open-source community, with contributions to widely used projects such as Transformers, LLaMA-Factory, and SGLang.

EDUCATION

East China Normal University *Sept 2021 - Now*

Bachelor of Engineering in Data Science and Big Data Technology

Relevant Courses: Statistics & Machine Learning(A), Computer Vision(98, A), Distributed System(93, A)

Visiting student at New York University Shanghai *Aug 2023 - Dec 2023*

PROJECTS

SGLang

Fixed bugs related to *fork* semantics (PR #2835), Added Triton kernel optimization benchmark (PR #6897), Support pipeline parallelism performance (PR #6908), Integrated support for new models (PR #6883, PR # 5485)

Uncertainty Quantification in LLMs

Deployed large language models using SGLang with integrated confidence estimation during response generation, leveraging conformal prediction to deliver statistically rigorous guarantees. Improve AUROC metric by 10% in Qwen series models.

RESEARCH EXPERIENCE

East China Normal University Oct 2023 - Dec 2023

Undergraduate Researcher in Decision Intelligence Lab, advised by Prof. Yang Shu *Shanghai, China*

TEXT2SQL: Benchmarked the code generation capabilities of state-of-the-art large language models (e.g., LLaMA, Qwen, GPT-4) on mainstream Text-to-SQL datasets, including BIRD.

Chinese University of Hong Kong, Shenzhen Jul 2024 - Dec 2024

Research Assistant in FreedomAI Lab, advised by Prof. Benyou Wang *Shenzhen, China*

LLM Inference & Prompt Engineering: Identified key prompt patterns, clarified their roles, and systematized prompt engineering. Furthermore, we proposed a prompt ensemble framework to assess the consistency of LLM inference.

PUBLICATION

Composition Pattern of Prompting Engineering, an Empirical Study

Qi Li, Nuo Chen, Chenyu Wang, Wanlong Liu, Zhongxiang Dai, Benyou Wang *Under review*

HAWKEYE: Efficient Reasoning with Model Collaboration

Jianshu She, Zhuohao Li, Zhemin Huang, Qi Li, Peiran Xu, Haonan Li, Qirong Ho *Under review*

WORK EXPERIENCE

China Foreign Exchange Trade System Aug 2023 - Sept 2023

Summer Internship *Shanghai, China*

1. Research and comparison of Mathematical Optimization tools for Business: Gurobi, CPLEX, COPT.
2. Utilized Numpy, Plotly, Pandas, Tkinter to develop a statistical tool for managing data.

HONORS AND AWARDS

1. National Encouragement Scholarship Sept 2023
2. The Chinese Mathematics Competitions (Shanghai Region). Frist Prize. Jan 2023
3. National College Student Mathematical Modeling Contest (Shanghai Region). Second Prize. Dec 2022

SKILLS

Programming Languages	C, C++, Python, Triton
AI Infra	vLLM, DeepSpeed, Docker