Jingyu Zhou

EDUCATION

Artificial Intellegence, SJTU   Sept. 2023 – Present
GPA 4.04/4.30 (92.0/100), Rank 4/95   *Minhang, Shanghai*
Zhiyuan Honor Program, SJTU   Sept. 2023 – Present
Top 10% of SJTU with advanced coursework   *Minhang, Shanghai*

- **Math**: Discrete Mathematics, Mathematical Analysis, Linear Algebra, Convex Optimization, Probability
- **CS**: Program Design, Data Structure and Algorithm, Machine Learning, Deep Learning

EXPERIENCE

IPADS, SJTU   Dec. 2024 - Present
Research Intern   *Minhang, Shanghai*

- Implemented KUNSERVE, a parameter-centric approach to optimize GPU memory usage in large language model (LLM) serving. (Codes coming soon)
- Introduced remote attention mechanism via pipeline parallelism, enabling memory borrowing from remote GPUs during load spikes. Thus tackled challenges such as KVCache exchange, balancing memory requirements, and minimizing execution overhead.
- Achieved up to 27.3x reduction in tail TTFT compared to state-of-the-art methods.

RHOS, SJTU & Qing Yuan Research Institute   Sept. 2024 – Jan. 2025
Research Intern   *Minhang, Shanghai*

- Proposed that the current definition of modalities is inadequate, searching for new definition of them.
- Joined a research team focused on multimodal machine learning, with an emphasis on Large Language Models (LLMs).

X-LANCE, SJTU   Feb. 2024 – Sept. 2024
Research Intern   *Minhang, Shanghai*

- Contributed to the preparation of a tutorial report on *TTS Based on Discrete Representations* for NCMMSC 2024, specifically focusing on ASR probe experiments.
- Conducted a comprehensive comparison of three types of TTS models (reconstruction, classification, and hybrid) against traditional continuous representations, demonstrating that discrete representations are comparable or superior performance to continuous representations under specific conditions. This finding offers valuable insights for future TTS model development.

HONORS AND AWARDS

Shanghai Scholarship, 2024

- Awarded to the **TOP 0.02%** of students in Shanghai.

Zhiyuan Honors Scholarship, SJTU, 2023 & 2024

- Awarded to the **TOP 5%** of students in the Zhiyuan Honor Program.

Merit Scholarship, B level, SJTU, 2024

- Awarded to the **TOP 5%** of students at SJTU.

PROJECTS

KunServe

- Contributed to the framework of KunServe, including LlaMa backend and LLM scheduling part.
- Implemented pipeline and tensor parallelism coordinated with remote attention mechanism.

SKILLS & INTERESTS

Language: Mandarin (native), English (TOEFL under preparation)
Programming: C/Cpp, Python, Pytorch
Interests: Large Language Model, Machine Learning System

jy_zhou@sjtu.edu.cn | Homepage