Zhou Zijian

(+65) 94476084 | zhou_zijian@u.nus.edu | https://www.linkedin.com/in/zijian-zhou-95abb7138/

EDUCATION

National University of Singapore Singapore

PhD in Computer Science Aug 2023 – Present

- **Research Focus:** data-efficient machine learning; data valuation, large language model post training, AI agents

B.S in Computer Science, 2nd Major in Mathematics Aug 2019 – May 2023

- **GPA:** 4.92 / 5.0

ACHIEVEMENTS

- NUS Science & Technology Scholarship (cover full undergraduate tuition fee with annual living allowance)
- School of Computing Outstanding Computing Project Prize 2022 (S$500 prize awarded to 6 students out of the entire School of Computing)
- School of Computing Dean's List (for all semesters up to now where the award is available)
- NUS Top 1% Performing Student for (Programming Methodology I (cohort size: 529); Machine Learning (cohort size: 264); Algorithmic Mechanism Design (cohort size: 73))

WORK EXPERIENCE

Singapore-MIT Alliance for Research and Technology Centre Singapore

Research Engineer Feb 2024 – Present

- Conducted research on efficiency in large language models under prof. Daniela Rus and prof. Low Kian Hsiang.
- Developed a method to actively select the most helpful task demonstrations for in-context learning, which allows the LLM performance to be maintained with only a half of the prompt tokens.
- Developed a method to improve the efficiency of speculative decoding by carefully selecting which tokens to proposal for verification.

ByteDance (TikTok Pte. Ltd.) Singapore

Machine Learning Engineer Intern Mar 2021 – Mar 2022

- Applied contemporary state-of-the-art algorithms to update several machine learning models to improve online performance; recall rate increased by an average of 5%; model executing speed (measured in max queries per second) doubled on average.
- Developed a set of automation tools to streamline the process of machine learning model iteration, including training sample selection using active learning and automated hyper-parameter searching tools. With the deployment of the automation tool, average model iteration time is reduced by 40%, with less human intervention.

SELECTED PUBLICATIONS

TETRIS: Optimal Draft Token Selection for Batch Speculative Decoding (ACL-25 main accepted)

Co-First Author Jan 2025

- This work proposed Tetris, an optimized speculative decoding method that improves inference efficiency in large language models (LLMs) by selectively generating draft tokens to maximize throughput and reduce computational overhead. Tetris achieves consistent improvement in performance by around 5% across various models and datasets.

DETAIL: Task Demonstration Attribution for Interpretable In-context Learning (NeurIPS-24 accepted; acceptance rate: 25.76%)

First Author May 2024

- This work proposed a novel method to efficiently attribute task demonstrations for in-context learning. By viewing in-context learning as performing implicit gradient descent on an internal optimizer, this work computes the attribution score as the influence function on the optimizer. The score, termed DETAIL score, can be used to effectively detect corrupt task demonstrations, and help curate and trim redundant demonstrations.

Probably Approximate Shapley Fairness with Applications in Machine Learning (AAAI-23 Oral accepted; acceptance rate: 19.6%)

Co-First Author Jan 2023

- This work identified a critical yet overlooked problem regarding the fairness guarantee of Shapley value in Machine learning. On this issue, this work proposed a novel *fidelity score* to measure the extent to which the fairness guarantee of Shapley value is preserved in practice. Under this framework, this work designed a simple yet efficient algorithm to estimate Shapley values with probabilistic fairness guarantee which prior works have not considered before.

SKILLS & INTERESTS

Languages: Fluent in English and Mandarin
Programming: Python, Go, Docker