Last updated in April 2025

Zian(Andy) Zheng

Ontario, Canada Email Phone Linkedin Github Personal Homepage

Education

MMath **University of Waterloo**, Computer Science, Ontario, Canada — Sept 2024 – Now

- Advised by Prof. Jian Zhao (Director of the WVisdom lab)

MComp **National University of Singapore**, Artificial Intelligence, Singapore — Sept 2022 – May 2024

- Advised by Prof. Yang You (Director of the HPC-AI lab)

BEng **Lanzhou University**, Data Science, Lanzhou, China — Sept 2018 – May 2022

- GPA: 92.8/100 (Ranking: 1/192)

Research Projects

OpenMoE: An Early Effort on Open Mixture-of-Experts Language Models — Github Repo

- Worked on the Pytorch implementation of **the first open-source, decoder-only MoE language model, OpenMoE**, providing insights about the routing mechanism to the open-source community [model checkpoint].
- Prepared the training dataset, tokenizer and conducted the **model evaluation**.

Matrix: Infinite-Horizon World Generation with Real-Time Interaction — Github Repo

- Built a **data collection pipeline** for Cyberpunk 2077, recording per-frame video data and corresponding control signals to support model training
- Served as a **core contributor to a multi-GPU inference framework**, implementing Ray-based workers (DiT, VAE, post-processing) and building an interactive frontend/backend system. Delivered the **first real-time game generation demo with user-controllable** inference [demo].

AdaVocab: Boosting SLM Inference with Sparse Vocabulary Activation — Github Repo

- Identified the growing **vocabulary size** as a major bottleneck for the Small Language Model (SLM) **inference efficiency**.
- Proposed and implemented a **sparsely active vocabulary** method; prepared training data, modeled Trainer, and completed evaluation with teammates.
- Achieved over **20% computation reduction** and **10% inference speedup** for SLMs in **CPU settings**.

Work Experience

HPC-AI Tech, Artificial Intelligence Engineer Intern — Beijing, China, July 2023 – Nov 2023

- **Extended** LLaMA's **vocabulary for Chinese** and **contributed** to **data preparation** in the **Colossal-LLaMA-2** project, selected as an official base model in the **2023 NeurIPS LLM Efficiency Challenge**.
- Investigated common **context length extrapolation** methods (e.g. PI, NTK, LongLoRA), and implemented corresponding **training and evaluation** pipelines to extrapolate Colossal-LLaMA-2 with **multi-GPU training**.
- Working on the ColossalQA project, a **RAG framework** based on Langchain.

Publications

[ICML 2024] OpenMoE: An Early Effort on Open Mixture-of-Experts Language Models — Jan 2024

Fuzhao Xue, ***Zian Zheng***, Yao Fu, Jinjie Ni, Zangwei Zheng, Wangchunshu Zhou, Yang You

[ICLR 2025] MixEval-X: Any-to-Any Evaluations from Real-World Data Mixtures Oct 2024

Jinjie Ni, Yifan Song, Deepanway Ghosal, Bo Li, David Junhao Zhang, Xiang Yue, Fuzhao Xue, ***Zian Zheng***, Kaichen Zhang Mahir Shah, Kabir Jain, Yang You, Michael Shieh

[Arxiv] The Matrix: Infinite-Horizon World Generation with Real-Time Moving Control Dec 2024

Ruili Feng, Han Zhang, Zhantao Yang, Jie Xiao, Zhilei Shu, Zhiheng Liu, ***Andy Zheng***, Yukun Huang, Yu Liu, Hongyang Zhang

Teaching & Service

Teaching Assistant: CS135, CS479/679 at UWaterloo

Reviewers: ICLR 2025, AISTATS 2025

Honors & Awards

China National Scholarship (Top 0.1% across nation), 12/2019 & 12/2021

Merit Student in Colleges and Universities in Gansu Province (Top 1% across province), 05/2021

Dr. Derick Wood Graduate Scholarship, 12/2024

Computer Skills

Programming & Software: Python, Java, C, SQL, Tableau, Echarts, Linux, Hadoop

Libraries: PyTorch, Ray, Pandas, NumPy, Scikit-Learn, PyQt5