Xinzhou Zheng

Name:郑心舟 | Email: zheng_xz@mail.ustc.edu.cn | (86) 198-5710-0843
Address: No.96, JinZhai Road, Baohe District, Hefei, Anhui, 230026, P.R. China

Education

University of Science and Technology of China (USTC) Sept 2023 - Jul 2027 (Expected)

School of Gifted Young

B.E. in Computer Science and Technology (expected in Jul 2027)

GPA:**4.02/4.3**, ranking: **3/236** (Major in Computer Science and Technology)

Publications

- Yuetao Chen, Xuliang Wang, **Xinzhou Zheng,** Ming Li, Peng Wang, Hong Xu**.**
  Optimizing Speculative Decoding by Reusing Discarded Draft via Hidden-State-Level Autoregression
  Submitted to OSDI 2026, under review.
- Xuliang Wang, Yuetao Chen, Maochan Zhen, Fang Liu, **Xinzhou Zheng**, Xingwu Liu, Hong Xu, Ming Li.
  PRISM: Parametrically Refactor Inference for Speculative Sampling Draft Models
  MLSys 2026

Research Experiences

Intern of USTC ADSL Lab, University of Science and Technology of China Sep 2025 - Present
Optimization of Long-Tail Effects in the Generation Phase of RLHF Training
Advisor: Professor **Cheng Li** (Associate Professor of School of Computer Science and Technology, USTC)

- Experimentally validated that beam search in rollout reduces the number of training steps required for convergence.
- Designed and implemented an early stopping mechanism to reduce runtime of beam search rollout.
- Developed entropy-based dynamic selection algorithm to refill batch slots after some sequences stopped.
- Reduced beam search runtime by 21% - 38.5% by enabling multi-token generation per step.
- Integrated beam search into Verl, a commonly used training framework, and supported tracking rollout FLOPs.

Intern of CUHK NetX Lab, Chinese University of Hong Kong Aug 2025-Dec 2025
Optimizing Speculative Decoding by Reusing Discarded Draft via Hidden-State-Level Autoregression
Advisor: Professor **Hong Xu** (Associate Professor of Department of Computer Science and Engineering, CUHK)

- Integrated hidden-state-level autoregressive draft generation (hidden algorithm) into SGLang framework.
- Implemented CUDA graph and target-to-draft vocabulary conversion logic for hidden algorithm.
- Integrated resample algorithm into hidden speculative decoding to reuse draft model generation results.
- Split draft batch into two parts, applied resample algorithm to one part only, then merged the verification results.
- Organized the two micro batches to enable CPU-GPU overlap and hide resample overhead.
- Supported manuscript development, currently under review for OSDI 2026.

Summer Intern of HKU Systems Software Lab, Hong Kong University Jul 2025-Aug 2025
Optimization of Communication Overlap in MoE Model Long Sequences Inference
Advisor: Professor **Heming Cui** (Associate Professor of Department of Computer Science, Hong Kong University)

- Explored mixed sequence-batch dimension splitting for comm-compute overlap in vLLM framework.
- Proved and experimentally validated that mixed splitting makes no difference under vLLM's implementation.
- Modeled complex compute demands, e.g. DeepseekV3 Multi-Head Latent Attention's prefill and decode stages.
- Implemented batch partitioning for balanced computation and chunked-attention-friendly adjustments.
- Implemented comm-compute overlap between microbatches for MoE models in vLLM.
- Achieved 33% speedup on a dual-node, four-GPU cluster compared to the baseline vLLM implementation.

Competition: 2025 ASC Student Supercomputer Challenge Nov 2024-May 2025
Optimization of DeepSeek V3 and Alphafold3 Inference on CPU and GPU
Advisor: Professor **Junshi Chen** (Research Associate Professor (Special Appointment), USTC)

- Optimized AlphaFold 3 algorithm to reduce time complexity and enable parallelization.
- Leveraged CPU core binding, Triton kernels, speculative decoding, and multi-node parallelization.
- Participated in hardware selection, server assembly, and power consumption management.
- Placed 6th Place in the Preliminary Round and won first prize in the Final Round.

Competitions & Awards

First Prize in the Final Round, 6th Place in the Preliminary Round, ASC25 Challenge	2025
Silver Medal, The 2025 CCPC Jinan Regional Contest	2025
Bronze Medal, The 2024 ICPC Asia Shanghai Regional Contest	2024
Bronze Medal, The 2023 ICPC Asia Nanjing Regional Contest	2023
Silver Medal, Finals, 20th BaiduStar Programming Contest.	2024
Individual First Prize, National Finals, China Collegiate Computer Contest	2024
Second Prize, Undergraduate Track, Anhui Provincial Robotics Competition	2024
Rose Fund Ambition Scholarship in School of Gifted Young	2024
Excellent Student Scholarship – Gold (top 3%)	2025

Skills

Experience with the vLLM and SGLang LLM inference frameworks and Verl RL training framework.
Familiar with C++, Python and Verilog programming languages and have some understanding of CUDA programming.
Application Development: Hands-on experience in game development with Unity.