# ZELEI SHAO

+1 2177661047 | zelei2@illinois.edu | https://github.com/zelci-hub

## EDUCATION

**University of Illinois Urbana-Champaign** College of Liberal Arts & Sciences 08/2024 - now
*Master of Science in Statistics*

**Peking University** School of Mathematical Sciences 09/2019 - 07/2023
*Major in Computational Mathematics*

## PUBLICATION

**Beat the long tail: Distribution-Aware Speculative Decoding for RL Training** MLSys 2026
**Zelei Shao***, Vikranth Srivatsa*, Sanjana Srivastava, Qingyang Wu, Alpay Ariyak, Xiaoxia Wu, Ameen Patel, Jue Wang, Percy Liang, Tri Dao, Ce Zhang, Yiying Zhang, Ben Athiwaratkun, Chenfeng Xu, Junxiong Wang

**MiLo: Effective Quantized MoE Inference with Mixture of Low-Rank Compensators** MLSys 2025
Beichen Huang*, Yueming Yuan*, **Zelei Shao***, Minjia Zhang

**Scaling Speculative Decoding with Lookahead Reasoning** NeurIPS2025
Yichao Fu, Rui Ge, **Zelei Shao**, Zhijie Deng, Hao Zhang

## INTERNSHIP

**Together AI** | *Research Intern in Turbo team* 06/2025- Now

- Designed and implemented **DAS** (Distribution-Aware Speculative Decoding), a systems-ML co-designed framework that **accelerates RL rollout** by **30–50%** while preserving identical training curves.
- Built a self-evolving suffix-tree-based drafter with **incremental online updates**, enabling high-acceptance draft tokens aligned with evolving RL policies.
- Developed a **length-aware draft budget allocator**, optimizing compute utilization for long-tail trajectories and reducing straggler makespan during RL rollouts.
- Integrated DAS into **VeRL and vLLM**, scaling experiments to up to six 8×H100 nodes.
- Co-authored a paper accepted to MLSys 2026, presenting a unified analytical model for optimal speculative budget assignment and suffix-tree-based speculation.

## RESEARCH

**Efficient Low-Bit Quantization and Inference System for MoE** 07/2024 - 01/2025
*Group Research, Supervised by Prof.Minjia Zhang, University of Illinois Urbana-Champaign*

- Implemented a highly efficient **W3A16 mixed-precision GeMM CUDA kernel** based on Marlin, achieving a **4.7x speedup** v.s. PyTorch on A100
- Designed a zero-bit-waste 3-bit weights packing technique, maximizing memory utilization
- Developed an efficient I2F (INT3-to-FP16) dequantization algorithm leveraging **register-level parallelism**
- Integrated optimizations into MiLo inference backend, achieving **3×** speedup v.s. Pytorch

**Scaling Speculative Decoding with Lookahead Reasoning** 03/2025 - 05/2025
*Group Research, Supervised by Prof.Hao Zhang, University of California San Diego*

- Co-designed **Lookahead Reasoning**, a novel step-level speculation method that scales speculative decoding (SD) along a new dimension orthogonal to token-level approaches
- Provided a **theoretical analysis** of the technique both in isolation and in combination with token-level SD, showing that under finite concurrency, peak performance is achieved only when both are used together
- Designed and implemented ablation studies on the impact of draft tree width and depth
- Integrated Lookahead Reasoning into **vLLM** and benchmarked the system, demonstrating up to **1.5× additional speedup** over token-level SD alone

## SKILL

**Programming Language**: Python, CUDA, C++
**Tools&Framework**: Pytorch, vLLM, VeRL

## AWARDS AND HONORS

- Merit Student, Peking University 2021
- Half-Dream Scholarship, School of Mathematical Sciences, Peking University 2021
- **1st Prize** in the National High School Mathematics Competition 2018
- **2nd Prize** in Chinese Mathematical Olympiad 2018