Zhaofeng Sun

sun-zf22@mails.tsinghua.edu.cn | thuecstasy.github.io | GitHub

Research interests: Efficient language model architectures, ML systems, and model compression.

Education

Tsinghua University Sep 2022 - Jun 2027 (expected)

B.S. in Computer Science | GPA: 3.97/4.00 | Rank: 5/208

Cornell University Jan 2025 - Jun 2025

Exchange Student, Computer Science | GPA: 4.15/4.00

Research Experience

University of California San Diego Jun 2026 - Present

Research with Prof. Hao Zhang, Hao AI Lab

- Developed ForesightLoop, which aggregates completed cross-trajectory states to improve parallel looped language models while preserving token-parallel training and concurrent decoding.
- Trained 0.6B models on 6B tokens and implemented CUDA Graph and Triton decoding. At eight loops, improved average downstream accuracy by 0.84 percentage points over PLT with 0.3% additional decoding time.

Carnegie Mellon University May 2025 - Aug 2025

Research with Prof. Beidi Chen

- Implemented Triton sparse-attention kernels in SGLang and integrated sparse rollouts into LLM reinforcement learning. Studied actor-policy mismatch and stable long-context rollouts; the resulting Sparrow framework achieves up to 2.4× rollout speedup.

Cornell University Jan 2025 - May 2025

Research with Prof. Christopher De Sa

- Co-developed YAQA, an adaptive-rounding method targeting full-model output error. Developed Kronecker-factored Hessian sketches and quantizer-independent rounding, reducing KL divergence by approximately 30% over GPTQ/LDLQ.

Tsinghua University Oct 2023 - Jan 2025

Research with Prof. Jianfei Chen

- Investigated activation outliers and layer-wise reconstruction in Transformers; contributed to hardware-aware V:N:M sparsity research for efficient inference.

Selected Publications

Sparrow: Sparse Rollout for Stable and Efficient Long-context RL of Large Language Models

Yang Zhou, Ranajoy Sadhukhan, **Zhaofeng Sun**, Zhuoming Chen, Souvik Kundu, Saket Dingliwal, Sai Muralidhar Jayanthi, Aram Galstyan, Haizhong Zheng, Beidi Chen. *Preprint, 2026. Earlier version: SPOT Workshop at ICLR 2026.* [Paper]

Model-Preserving Adaptive Rounding

Albert Tseng, **Zhaofeng Sun**, Christopher De Sa. *ICML 2026.* [Paper]

Beyond 2:4: Exploring V:N:M Sparsity for Efficient Transformer Inference on GPUs

Kang Zhao, Tao Yuan, Han Bao, Zhenfeng Su, Chang Gao, **Zhaofeng Sun**, Zichen Liang, Liping Jing, Jianfei Chen. *Preprint, 2024.* [Paper]

Industry Experience

Huawei Noah's Ark Lab | Research Intern Oct 2024 - Dec 2024

- Profiled LLM inference on Ascend NPUs, optimized GEMM kernels, and evaluated mixed-precision quantization.

Honors and Skills

Champion, Tsinghua University Supercomputer Competition (2024). Comprehensive Excellence Scholarship, top 5% (2023-2025).

Programming: Python, C/C++, CUDA, Triton. **Systems:** PyTorch, FSDP, Megatron, vLLM, SGLang, verl.