BOWEN WANG

■ abmfy@icloud.com · • abmfy · ★ abmfy.github.io

EDUCATION

Tsinghua University (THU), Beijing, China

Sep. 2021-Present

B.Eng. in Computer Science and Technology (CST), expected June 2026 GPA: 3.95/4, Rank: 5/162 Minoring in Economics and Finance TOEFL 109/120, R/30, L/28, S/24, W/27

RESEARCH INTERESTS & TECHNICAL SKILLS

Research Interests

- Machine Learning Systems: inference engines, memory management, and large-scale parallelism
- Large Language Models: agents and efficient algorithm optimizations

Technical Skills

- Programming Languages: Python, Rust, C/C++, TypeScript, Scala, Go, SystemVerilog
- GPU Kernel Languages: CUDA, Triton, TileLang
- ML Frameworks: PyTorch, vLLM, SGLang

EXPERIENCE

Sky Computing Lab, University of California, Berkeley

Dec. 2024-Present

Visiting Student Researcher Advisor: Prof. Ion Stoica

Research Topic: LLM Inference Engine Design; Expert Parallelism Optimization

- PrefillOnly: An Inference Engine for Prefill-only Workloads in LLM Applications (SOSP 2025)
 - First to identify and design an inference system specialized for prefill-only workloads, where only a single output token is produced, e.g., recommendation, credit verification, and data labeling
 - Implemented Hybrid Prefilling utilizing torch.compile, where input to linear layers is chunked to reduce memory peaks, achieving up to **7.9x context length** in prefill-only scenarios
 - Precisely estimated job completion time with prefix-cache awareness, enabling effective scheduling, achieving up to **4x higher QPS** without inflating average and P99 latency
- vLLM Expert Parallelism Load Balancer (EPLB) [Merged PR]
 - Profiled and identified the imbalance of the expert usage in sparse Mixture-of-Experts (MoE) inference
 - Implemented a load balancer that dynamically reassigns experts based on observed usage patterns
 - Achieved up to 30% throughput improvement and 25% latency reduction in sparse MoE inference

Stanford Undergraduate Visiting Research (UGVR)

Jul. 2024-Aug. 2024

Research Intern Advisor: Prof. Tsachy Weissman

Research Topic: Information theory perspective on LLM inference

- Explored the potential of utilizing data compression techniques in speeding up LLM inference
- Investigated different tokenization strategies and their impact on compression ratio and performance

Z.ai Beijing, China Jul. 2023–Jun. 2024

Research Intern Research Topic: Agentic LLM training; Inference infrastructure for agentic LLMs

- Data synthesis and post-training for ChatGLM3 and GLM-4, which were the models behind Z.ai
- Designed and developed the GLM-4 All Tools backend agent system [News]

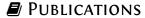
Knowledge Engineering Group (KEG), Tsinghua University

Jul. 2023-Jun. 2024

Research Intern Advisor: Prof. Jie Tang & Prof. Yuxiao Dong

Research Topic: Agentic LLM training; Efficient decoding algorithms for LLMs

- APAR: Auto-Parallel Auto-Regressive Decoding
 - Trained the model to perform fork-based parallel decoding with KV-cache sharing and early release
 - Achieved a speedup of 2x, further up to 4x with speculative decoding
- AgentTuning: Enabling Generalized Agent Abilities for LLMs (ACL 2024 Findings)
 - Discovered a hybrid instruction fine-tuning method to enable generalized agent abilities for LLMs
 - Proposed AgentInstruct, a compact dataset containing ∼1.9k high-quality agent trajectories



GOOGLE SCHOLAR CITATIONS: 1500+

PrefillOnly: An Inference Engine for Prefill-only Workloads in LLM Applications SOSP 2025 *Acceptance Rate: 17.84%*

Kuntai Du, **Bowen Wang**, Chen Zhang, Yiming Cheng, Qing Lan, Hejian Sang, Yihua Cheng, Jiayi Yao, Xiaoxuan Liu, Yifan Qiao, Ion Stoica and Junchen Jiang [**PDF**]

Barbarians at the Gate: How AI is Upending Systems Research

arXiv 2025

Audrey Cheng*, Shu Liu*, Melissa Pan*, Zhifei Li, **Bowen Wang**, Alex Krentsel, Tian Xia, Mert Cemri, Jongseok Park, Shuo Yang, Jeff Chen, Lakshya Agrawal, Aditya Desai, Jiarong Xing, Koushik Sen, Matei Zaharia, Ion Stoica [**PDF**]

AgentTuning: Enabling Generalized Agent Abilities for LLMs

ACL 2024 Findings

Aohan Zeng*, Mingdao Liu*, Rui Lu*, Bowen Wang, Xiao Liu, Yuxiao Dong and Jie Tang [PDF] [Repo]

APAR: LLMs Can Do Auto-Parallel Auto-Regressive Decoding

arXiv 2024

Mingdao Liu*, Aohan Zeng*, Bowen Wang, Peng Zhang, Jie Tang and Yuxiao Dong [PDF]

ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools

Team GLM [PDF] arXiv 2024

Contributions: data synthesis, model training, inference infrastructure, open-source releases

OPEN-SOURCE PROJECTS

vLLM, a high-throughput and memory-efficient inference and serving engine [Repo] ★ 63.4k Stars

- Implemented and am currently in charge of the Expert Parallelism Load Balancer (EPLB), which is used in real-world large-scale MoE deployments, for example by Red Hat
- Authored 5 non-documentation PRs, reviewed 20+ PRs

GLM Series, open multilingual agentic LLMs [GLM-4 Repo] [ChatGLM3 Repo] ★ 20.6k Stars

- Post-training, evaluation and open-source releases of ChatGLM3-6B and GLM-4 9B
- Developed the local agentic demo for GLM series

TEACHING ASSISTANT & SOCIAL WORK

Software Engineering Teaching Assistant, THU CST Spring & Fall 2023, Spring & Fall 2024 *Part of THU CST's curriculum reform* [Course Homepage]

- Designed and implemented the scaffold for the assignment, a Next.js + Django instant messaging app
- Designed the CI/CD workflow of assignment submission to replace manual grading

Digital Logic Experiments Teaching Assistant, THU CST

Spring 2024

Part of THU CST's curriculum reform

• Helped to transition the course from VHDL to SystemVerilog, a more modern and powerful hardware description language

Student Association of Science and Technology, THU CST

Jun. 2022-Jun. 2024

Vice President

- Organized SAST Summer Training 2023, a 4-week training camp for freshmen with 40,000+ replays [Link]
- Initiated Weekly9, a blog sharing platform for students to share their learning experiences and insights

➡ Honors & Awards

Comprehensive Excellence Scholarship (Top 3%), Tsinghua University

Oct. 2024, 2023, 2022

Special Prize (National Top 1), 7th "Loongsun Cup" CPU Design Competition [Repo][News] AEON Scholarship

Aug. 2023 Dec. 2023

First Prize (Provincial Level), Chinese National Olympiad in Informatics in Provinces (NOIP/CSP) Oct. 2019