

EDUCATION

Shanghai Jiao Tong University (SJTU)
B.S. Candidate in Computer Science and Technology, Zhiyuan College, John Honors Class
Sep. 2024 – Present

ACADEMIC PERFORMANCE

Academic Credit Score: 92.7 / 100
Rank: 2 / 19
GPA: 4.00 / 4.3
Rank: 2 / 19

RESEARCH EXPERIENCE

Undergraduate Research Assistant
Research Area: Machine Learning Systems
Advisor: Chen Chen, EPCC Lab
Sep. 2025 – Present

- GoodServe** Second Author; Preprint, Under Review
Towards High-Goodput Serving of Agentic LLM Inferences over Heterogeneous Resources
 - * Co-designed and implemented a predict-and-rectify routing framework for agentic LLM inference over heterogeneous GPUs, combining output-length prediction, serving-status estimation, and just-enough instance selection.
 - * Developed periodic SLO-risk re-evaluation and runtime request migration to handle unexpected workload dynamics, improving goodput by up to 27.4% over existing routing methods.
- Co-location of Online and Offline LLM Inference with GreenCTX** Collaborative, Ongoing
 - * Developed dynamic KV-cache scaling with proactive offloading and prefetching to improve memory efficiency and enable elastic co-location of online and offline LLM inference.
- MoE Expert Load Balancing System** Independent, Ongoing
 - * Formulated expert placement and scheduling as an ILP problem and implemented expert loading and routing mechanisms; explored hybrid TP/EP and layer-wise deployment to mitigate expert hotspots.

Microsoft Research Asia (MSRA)
Research Intern, Microsoft Research Asia – Vancouver
Mentor: Tao Zhang and Jiamin Li, Senior Researcher
Jul. 2026 – Present

SCHOLARSHIPS & HONORS

National Scholarship 2025
Awarded to the top 0.2% of students nationwide; the highest national-level scholarship in China.

Outstanding Undergraduate Scholarship (A-Level) 2025
University-wide scholarship awarded to the top 1% of undergraduates for outstanding overall performance.

Hanying Juhua Outstanding Student Scholarship 2025
Awarded to 15 outstanding students in the Zhiyuan Honors Program for academic excellence.

Zhiyuan Honors Scholarship 2024, 2025
Awarded to top-performing students majoring in science.

COMPETITION AWARDS

Second Prize, Huawei Hackathon Software Challenge Oct. 2025
Third Prize, Shanghai Division, China Undergraduate Mathematical Contest in Modeling (CUMCM) Sep. 2025

PROFESSIONAL SKILLS

- Programming Languages:** Proficient in C++ and Python; familiar with CUDA and PyTorch.
- Frameworks:** Proficient in SGLang; familiar with vLLM, including customization and secondary development.
- Tools:** Linux, Git, Docker, and Shell scripting.
- LLM Inference Optimization:** Continuous Batching, PagedAttention, Chunked Prefill, PD Disaggregation, TP, PP, EP, and DP parallelism, hierarchical KV-cache management, and scheduling policies.
- English:** TOEFL iBT: 90/120 (4.5/6); CET-6: 614.

STUDENT LEADERSHIP

Zhiyuan College Student Union, Academic Affairs Department Member Feb. 2025 – Jun. 2025
Zhiyuan College Student Union, Deputy Minister, Academic Affairs Department Sep. 2025 – Present