Boning Huangfu

margin689@sjtu.edu.cn | smiling689.com | GitHub | Google Scholar

EDUCATION

Shanghai Jiao Tong University (SJTU) *Sep. 2024 – Jun. 2028 (expected)*
B.S. in Computer Science and Technology, Zhiyuan College, John Honors Class
GPA: 4.00/4.30 **Academic Credit Score:** 92.7/100 **Rank:** 2/20

RESEARCH EXPERIENCE

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

- **GoodServe** *Co-first Author, arXiv Preprint*
  Towards High-Goodput Serving of Agentic LLM Inferences over Heterogeneous Resources
  * Co-designed GoodServe for end-to-end SLO-aware agentic LLM serving on heterogeneous GPUs, combining lightweight output-length prediction, GPU serving-status estimation, and just-enough instance selection.
  * Developed periodic SLO-risk checks and lightweight token-ID-based request migration; GoodServe achieves up to 31.0% higher goodput than baseline routing methods.
- **HarmonyRoute** *First Author*
  Maximize Compute-Budget Utility with Coordinated Multi-Model Routing
  * Formulated coordinated routing for LLM requests sharing a compute budget (an API spending cap or a private GPU cluster), maximizing the budget's utility (total satisfaction) rather than each request's own quality–cost trade-off.
  * Designed a predict–plan–rectify router: relative demand prediction, a lookahead solver that prices compute for unknown later arrivals, and budget-aware runtime correction of in-flight requests.
  * Outperforms eight existing routers at all 21 budgets, with 18.72% higher total satisfaction than the strongest.
- **CoServe: Fine-Grained Resource Management for Online–Offline Co-location** *Collaborative Project*
  * Co-developed dynamic SM-level GPU resource partitioning with NVIDIA CUDA Green Contexts, reallocating SM resources between co-located online and offline workloads in response to online request bursts.
  * Designed layer-wise KV-cache offloading and prefetching on SGLang, ordering loading, computation, and write-back across CUDA streams with events so that only one layer of offline KV cache stays on the GPU.

Microsoft Research Asia (MSRA) *Jul. 2026 – Present*
Research Intern, Microsoft Research Asia – Vancouver Lab
Mentors: Jiamin Li and Peng Cheng

- **ShrinkWrap: AI Infrastructure World Modeling for Inference Optimization** *Independent, Ongoing*
  * Building an agent that reconstructs the design rationale behind inference techniques and builds a causal world model of how design and configuration choices affect system behavior, resource costs, and performance under different conditions.
  * Completed an end-to-end proof of concept on DistServe (design rationale $\rightarrow$ reasoning graph $\rightarrow$ causal knowledge graph); now evaluating whether the graph's predicted effects match measurements.

HONORS & AWARDS

National Scholarship *(Awarded to top 0.2% of undergraduates nationwide)* *2025*
Outstanding Undergraduate Scholarship (A-Level) *(Top 1% at SJTU)* *2025*
Hanying Juhua Outstanding Student Scholarship *(15 recipients in Zhiyuan Honors Program)* *2025*
Zhiyuan Honors Scholarship *2024, 2025, 2026*
Second Prize, Huawei Hackathon Software Challenge *Oct. 2025*
Third Prize, Shanghai Division, CUMCM *Sep. 2025*

TECHNICAL SKILLS

Programming & Tools: C++, Python; Linux, Git, Docker, Shell scripting; Claude Code, Codex.
ML Systems: SGLang (proficient), vLLM, PyTorch; P/D disaggregation, TP/PP/EP, MoE inference, serving optimization.
English: TOEFL iBT 90/120 (4.5/6); CET-6 614.

STUDENT LEADERSHIP

Zhiyuan College Student Union, Academic Affairs Department *Feb. 2025 – Sep. 2026*
Member (Feb.–Jun. 2025); Deputy Minister (Sep. 2025–Sep. 2026)