Siyang SHAO

sshao@gatech.edu | (470) 378-9780 | github.com/SiyangShao

Education

Georgia Institute of Technology Aug 2025 – Current
Graduate student in Computer Science Atlanta, Georgia, USA

- Research Area: Large Language Model Inference, Distributed System

Nanyang Technological University Aug 2021 – Jun 2025
Bachelor of Engineering (Computer Engineering) Singapore, Singapore

- Honours (Highest Distinction); GPA: 4.63 / 5.0

Work Experience

Jane Street Asia Limited Hong Kong SAR
Software Engineer Intern May 2025 - Jul 2025

- Built a version-conversion library for JSON-RPC that aligns JSON-RPC and async-RPC under a unified declaration, enabling seamless backward/forward compatibility and reducing integration overhead across services
- Designed an incremental synchronization prototype to mirror an internally defined DSL-based database into a SQL backend, supporting schema evolution with idempotent upserts and conflict-safe application of changes
- Benchmarked the new path and observed 5× faster queries versus the legacy approach under representative workloads; added end-to-end tests and tooling to validate correctness and performance

TikTok Pte. Ltd. Singapore
Backend Engineer Intern, Video Infrastructure Jan 2024 - May 2024

- Co-designed and implemented a metrics metadata management service that standardizes metric naming, ownership, and label conventions, bridging development and SRE practices across large-scale services
- Built a persistent global SLI framework that defines, computes, and monitors service-level indicators consistently across regions and tiers, enabling uniform dashboards and alerting
- Automated discovery and governance for new metrics/SLIs with validation and documentation hooks, improving observability hygiene and reducing onboarding friction
- Partnered with SREs to roll out the framework to high-priority services, improving visibility and reducing manual intervention during incidents

Open Source Projects

ServerlessLLM https://github.com/ServerlessLLM/ServerlessLLM Jun 2024 - May 2025
Core Contributor Jun 2024 – May 2025

- Added ROCm support to enable high-throughput model loading on AMD GPUs and reducing cold-start latency
- Maintained the system controller coordinating lifecycle management of inference backends (init, health, scale in/out), improving reliability under multi-tenant workloads
- Collaborated with contributors by reviewing PRs, triaging issues, and updating docs to ensure release quality and reproducibility

Awards

• 2022 ICPC Asia Manila Regional Ranked 2	Dec 2022
• 2023 ICPC Asia Jakarta Regional Ranked 13	Dec 2023
• 2024 ICPC Asia Pacific Championship Ranked 22	Mar 2024
• 2025 ICPC Asia Jakarta Regional Ranked 11	Dec 2024
• 2025 ICPC Asia Pacific Championship Ranked 24	Mar 2025
• Dean’s List in Academic Year 2022-23 (Top 5% of cohort)	Aug 2023
• NTU President Research Scholar in Academic Year 2023-24	Aug 2024