邵思洋

siyang.shao@outlook.com | +86-15021988618 | github.com/SiyangShao

教育背景

佐治亚理工学院 2025 年 8 月 – 现在
计算机科学博士生 美国乔治亚州亚特兰大
- 研究领域：大语言模型推理，分布式系统

南洋理工大学 2021 年 8 月 – 2025 年 6 月
工学学士（计算机工程） 新加坡，新加坡
- 一等荣誉学位; GPA: 4.63 / 5.0

开源项目

ServerlessLLM https://github.com/ServerlessLLM/ServerlessLLM
核心开发者 2024 年 3 月 – 2025 年 5 月
- 支持 `sllm-store` 的 ROCm 版本开发与维护，`sllm-store` 是 ServerlessLLM 的内部库，提供高性能模型加载
- 集成 vLLM 后端，使 ServerlessLLM 项目能够通过 vLLM 进行推理
- 探索使 vLLM 后端受益于通过 `sllm-store` 进行高性能模型加载的方法
- 维护 ServerlessLLM 项目的控制器，管理推理后端的生命周期

工作经历

简街资本 中国香港
软件工程师实习 2025 年 5 月 – 2025 年 7 月
- 设计了一个支持 JSON-RPC 被调用方版本转换的库，实现了 JSON-RPC 与 async-RPC 的无缝对齐。该库支持以统一的声明方式注册两种 RPC 类型，并自动完成版本转换，简化了集成和维护工作。
- 设计并实现了一个数据库原型，能够将内部数据库（由自定义 DSL 定义）与新 SQL 数据库进行增量同步。该方案简化了用户的使用流程，并使查询性能提升约 5 倍。

抖音集团 新加坡
后端工程师实习（视频架构） 2024 年 1 月 – 2024 年 5 月
- 参与设计并实现了指标元数据的发现与管理系统，成功弥合了开发团队与 SRE 团队在监控指标方面的协作鸿沟。
- 实现了全局持久化 SLI (服务级别指标) 监控与管理系统, 有效提升了全链路的稳定性。

课外活动

南洋理工大学算法竞赛团队 https://icpc.global/ICPCID/B15T259WIX3C
团队成员 Dec 2021 – Mar 2025
- 代表学校参加国际大学生程序设计竞赛(ICPC), 并成功解决多项复杂算法问题

奖项

- 2022 年 ICPC 亚洲马尼拉赛区银牌 (第 2 名) 2022 年 12 月
- 2023 年 ICPC 亚洲雅加达赛区第 13 名 2023 年 12 月
- 2024 年 ICPC 亚太总决赛第 22 名 2024 年 3 月
- 2025 年 ICPC 亚洲雅加达赛区第 11 名 2024 年 11 月
- 2025 年 ICPC 亚太总决赛第 24 名 2025 年 3 月
- 2022-23 学年院长名单 (前 5%) 2023 年 8 月
- 2023-24 学年南洋理工大学校长研究学者 2024 年 8 月