

罗文杰

AI推理与模型评测工程师 | 2027 届

+86 183 1949 6086 | e1576499@u.nus.edu

github.com/LUOaini1213 | luoini1213.github.io

核心优势

通过同硬件对照、数值精度校验和逐场景统计分析评估推理性能与模型表现。具备SGLang/vLLM部署压测、GPU算子实验和PyTorch训练经验，以日志、源码与干预实验验证机制。

教育背景

新加坡国立大学 | 硕士 · 交通运输工程

2025.08 - 2027.01

研究方向：交通信息工程及控制

长安大学 | 本科 · 交通运输

2021.09 - 2025.06

GPA 3.16/5.0; 专业约前 15%

项目经历

SGLang/vLLM推理服务与性能归因

2026.09

- 在T4、Qwen2.5-0.5B上定位SGLang吞吐骤降：默认解码图仅捕获至batch 8；通过关图和扩大捕获范围验证因果机制。
- 将解码图捕获上限设为32，同轮batch 32吞吐608→2690 tok/s (4.42倍)；比较已捕获与未捕获组的关图代价，保留每档单次测量记录。

fp16x3 Transformer GPU优化

2026.08.23 - 至今

- 编写Triton融合核拆分fp16高低位，以K扩为三倍的一次cuBLAS张量核GEMM完成计算；结合融合注意力与分块降低显存占用。
- T4上13/13形状通过校验，最大绝对误差9.5e-6；每形状独立运行3次，相对FP32参考实现中位加速2.83倍；另以FP16运行100k序列，峰值显存14.2GB。

NAVSIM | 规划模型训练与评测

2026.08 - 2026.09

- 在103,288场navtrain训练MLP、12,146场navtest评测；地图×他车输入做2×2设计，每组3种子，统一cache计算配对95%置信区间。
- 复现恒速20.7 (论文20.6) 与Human 94.6 (论文94.8)；加入真值框后，已有地图的模型3种子PDMS均显著下降，而开环L1近乎不变。

实习经历

新加坡远东幕墙 | 工程部 · 项目工程师

2026.06 - 至今

- 用Python解析发运与测试邮件，提取到货批次及检测结果，支持15种材料的状态跟踪与工程协同。

彩讯科技 | 产品实习生

2025.06 - 2025.08

- 针对抽奖入口位于页面底部、不易被用户发现的问题，将入口调整为页面侧边弹窗，提升入口可见性；完成相关需求梳理与交互原型设计，跟进开发及功能验证。
- 独立撰写3份PRD，使用Axure完成5个核心页面原型，方案获采纳并进入开发排期；跟踪缺陷修复与复验，推动解决10余项问题。

技能

推理与性能：SGLang、vLLM、CUDA C++、Triton、PyTorch、CUDA Graph

评测：受控实验、数值精度校验、配对bootstrap、因子设计、多种子、防泄漏切分