李世政

15216727572 | 740925018@qq.com
https://github.com/shizhengLi | https://lishizheng.blog.csdn.net
求职意向：算法工程师

教育背景

复旦大学　计算机技术　硕士 2023-09 - 2026-06

复旦大学　辅修金融学　本科 2018-09 - 2020-06

上海理工大学　通信工程　本科 2017-09 - 2021-06

项目经历

多智能体系统与大语言模型Agent优化　项目负责人 2025-02 - 2025-07

项目描述：本项目整合了 Qwen-Agent、Void、Multi-Agent和MCP协议等开源框架的学习与实践，专注于多智能体系统的设计、开发与优化，结合大语言模型（LLM）的Agent功能实现高效协同与任务处理。通过研究 Qwen-Agent 的对话与任务规划能力、Void （开源版Cursor）AI辅助编程工具设计逻辑、Multi-Agent的多智能体协作机制以及基于MCP协议的MCP Server开发，探索了多智能体系统在实际业务场景下的应用与落地。项目基于Python和TS，实现了从多智能体协作原理到软件设计，以及开发实现的完整流程。

负责内容：

（1）源码分析并实现

分析 Qwen-Agent 的对话规划与工具调用机制，学习其如何优化在多轮交互场景下的任务分解与执行效率。

研究 Void的源码，学习工业级产品如何设计落地。

深入Multi-Agent （例如Open Deep Research）的多智能体协作策略，学习工程落地经验。

分析学习Context7 MCP源码，并设计4万行代码的MCP server项目（DevInsight AI Platform，开发助手），单元测试、集成测试全部通过。

高效大语言模型推理框架原理剖析　项目负责人 2025-05 - 2025-08

项目描述：本项目整合了 vLLM、verl的学习与实践，专注于高效大语言模型（LLM）的推理和训练优化。通过深入研究 vLLM 的高吞吐量推理引擎、verl 的分布式训练框架，学习大型训练与推理项目的开发与优化经验。

负责内容：

（1）算法研究与工程实践：

深入分析 vLLM 的 PagedAttention 和连续批处理机制，研究其内存高效管理和高吞吐量推理的实现原理，提升推理性能。

探索 verl 的 FSDP（Fully Sharded Data Parallel）训练策略，优化分布式训练中的通信开销和内存分配。

CUDA 与 Triton 高效 GPU 计算优化　项目负责人 2025-06 - 2025-08

项目描述：本项目专注于通过学习和实现 CUDA 与 Triton 编程技术，重点研究 Flash Attention 算法原理与实现。通过深入分析 CUDA 的并行编程模型和 Triton 的高性能算子开发，掌握了 GPU 加速计算的核心技术。项目基于 Python、CUDA 和 Triton 技术栈，实现了从算法研究到高性能算子开发的完整流程。

负责内容：

（1）算法研究：

深入研究 Flash Attention 算法的实现原理，分析其在 Transformer 模型中的内存优化与计算加速机制。

研究 Triton 的算子开发框架，设计高效的 GPU 内核，降低内存访问延迟并提升计算吞吐量。

（2）核心实现：

使用 C++ 和 CUDA以及Triton实现 Flash Attention 的核心计算模块，优化矩阵运算与内存访问模式。

开发高效的 Triton 内核，提升复杂深度学习任务的计算性能。熟悉Adam和Muon优化器原理，使用Triton实现，并通过测试。

研究成果

Paper 1: Searching for Best Practices in Retrieval-Augmented Generation (EMNLP 2024 Main, Citation: 143)
Authors: Xiaohua Wang, Zhenghua Wang, Xuan Gao, Feiran Zhang, Yixin Wu, Zhibo Xu, Tianyuan Shi, Zhengyuan Wang, **Shizheng Li**, Qi Qian, Ruicheng Yin, Changze Lv, Xiaoqing Zheng, Xuanjing Huang
主要内容：(1) 针对检索增强生成（RAG）技术在复杂实现和响应时间较长的问题：分析发现RAG工作流程涉及多步骤处理，每一步存在多种执行方式，影响性能与效率；提出了一种优化RAG部署策略，通过系统性分析现有RAG方法及其组合，兼顾性能与效率；(2) 通过广泛实验验证：所提出的RAG部署策略在性能与效率之间取得平衡，多模态检索技术在视觉输入问答任务中表现优异，且“检索即生成”策略有效缩短响应时间。

专业技能

- 熟悉 Python, C++, TypeScript, Cuda, Triton编程
- 熟悉基本的强化学习原理：PPO，DPO，GRPO, VC-PPO, VAPO等

个人评价

- 思维方面：喜欢数学，物理，逻辑推理；深度思考；喜欢阅读。
- 工作方面：自我驱动，喜欢进入心流状态。
- 身体方面：喜欢跑步，健身。之前半马 PB 1小时45分。
- 生活方面：乐观开朗，善于沟通，为人和善。