

张嘉

18292060445 · plus_zzz@163.com · 2002-09
个人主页: pluszzz.top
GitHub: github.com/Pluszzz



教育经历

2024 - 至今	西北大学 211 & 双一流	软件工程 / 硕士
2020 - 2024	西安邮电大学	信息与计算科学 / 本科
"华为杯"中国研究生数学建模竞赛 — 国家二等奖 - 全国大学生数学建模竞赛 — 省级二等奖 - CET-4		

项目经历

RegentAI	github.com/Pluszzz/RegentAI	2025-09 ~ 2026-02
----------	-----------------------------	-------------------

项目介绍: 企业级 RAG 智能体平台, 核心实现从文档入库到智能问答的全链路 RAG 系统, 完整覆盖意图识别、问题重写、多路检索、会话记忆、模型路由容错等企业生产环节。采用分布式队列限流控制并发, 通过熔断器和优先级降级保障多模型场景下的高可用性

技术架构: SpringBoot + MyBatis Plus + Milvus + Redis + Redisson + Apache Tika + Sa-Token

职责描述:

- 多路检索架构设计:** 设计多路检索引擎 (意图定向 + 全局向量检索 + 后处理流水线), 在保证召回率的同时提升检索精准度
- 问题理解优化:** 基于 LLM 实现问题重写与拆分机制, 自动补全上下文并拆解复杂问题
- 会话记忆管理:** 设计会话记忆压缩策略 (滑动窗口 + 自动摘要), 降低 token 消耗, 解决长上下文 Token 超限问题
- 模型调度与容错:** 实现模型路由与三态熔断机制, 支持多模型优先级调度与故障自动切换
- 意图识别体系:** 构建树形多级意图识别结构, 置信度不足时系统主动引导用户澄清需求
- 工具调用框架:** 实现 MCP 工具集成框架, 通过注册表自动发现, 实现知识检索与工具调用融合
- 可编排入库流程:** 设计文档入库 Pipeline, 新增处理节点只需实现统一接口即可自动接入流程
- 分布式限流控制:** 实现 Redis ZSET + Lua 滑动窗口限流方案, 支持全局与用户级并发控制
- 上下文遗传机制:** 基于 TransmittableThreadLocal 实现用户上下文与 Trace 在异步线程池中的透传, 保障多线程场景下的全链路追踪能力

短链接平台	github.com/Pluszzz/ShortLinkPlatform	2025-05 ~ 2025-08
-------	--------------------------------------	-------------------

项目介绍: SaaS 短链接系统, 为企业和个人用户提供高效、安全和可靠的短链接管理平台, 提供深入的分析和跟踪功能, 用户可以灵活地管理和优化其链接

技术架构: Spring Boot + Spring Cloud Alibaba + MySQL + Redis + RocketMQ + ShardingSphere + Sentinel

职责描述:

- 消息幂等控制:** 基于 Redis 实现 MQ 消费幂等性, 确保消息在指定时间窗口内仅消费一次
- 缓存穿透防护:** 封装缓存穿透防护方案, 通过双重判定锁与空值缓存优化大量不存在数据的查询问题
- 布隆过滤优化:** 使用 Redisson 布隆过滤器判断短链接存在性, 提高判断效率, 有效减少无效查询请求
- 流量治理与限流保护:** 整合 Sentinel 实现 QPS 级别限流, 触发限流时进行降级处理保障系统稳定性
- 异步削峰架构:** 采用 RocketMQ 消息队列异步消费模式, 削峰处理高并发访问下的监控数据写入需求
- 缓存一致性设计:** 设计"更新 DB 删除缓存"策略, 保障短链接缓存与数据库之间的数据一致性
- 分布式并发控制:** 使用 Redisson 分布式读写锁, 支撑短链接高并发场景下的数据修改操作
- 分库分表扩展:** 基于 ShardingSphere 实现分片路由, 支撑跳转与分页场景下的定向查询

专业技能

AI Coding

熟悉 AI Agent 开发 workflow, 能够使用 Codex、Claude Code 等工具完成代码生成、调试与自动化开发

Java & JVM

理解 Java 并发模型 (JUC、AQS、CAS)、线程池调优; JVM 内存模型、GC (G1/CMS)、OOM 排查与性能调优

后端框架

Spring Boot / Spring Cloud Alibaba 微服务体系; Nacos、OpenFeign、Sentinel 注册发现与流量治理

数据库

MySQL 索引设计、执行计划分析、MVCC 机制与事务隔离级别; 慢查询优化与读写分离经验

缓存与中间件

Redis 数据结构与持久化; 缓存穿透/击穿/雪崩方案; 分布式锁、滑动窗口限流、布隆过滤器

大模型应用 (RAG)

企业级 RAG 全链路: Milvus 向量检索、多路召回、问题重写、会话记忆压缩、模型路由与熔断

工程能力

Git 版本管理、Maven 项目构建与模块化开发