



朱旭

求职意向: Agent 应用开发实习生/AI 应用开发实习生

具备 RAG、Tool Calling 与 Agent 工程化项目经验, 参与企业智能客服平台二次开发, 并独立完成 Agent 项目的开发、评测和服务部署。

在线演示: <https://app8095.acapp.acwing.com.cn/>

邮箱: zhuxu2825@gmail.com

电话: 13895932825

教育背景

南昌大学

软件学院 | 软件工程 | 本科

2023-09 至 2027-07

专业技能

- Python 后端:** 熟悉 Python、FastAPI、Pydantic、SQLAlchemy Async、pytest, 能够完成 RESTful API、SSE 流式接口、参数校验、异常处理及自动化测试; 使用过 PostgreSQL、MySQL、SQLite、Redis。
- RAG 与检索:** 掌握文档加载、Chunk 切分、Embedding、向量检索及引用返回流程; 具备 Dense + BM25 双路召回、RRF 融合、检索评测和 Bad Case 分析实践, 使用过 ChromaDB、pgvector 和 BGE Embedding。
- Agent 应用:** 具备 Tool Calling、ReAct、MCP、JSON-RPC 和 Agent 状态管理实践, 能够实现工具路由、参数校验、运行预算控制、敏感工具审批、执行 Trace 和异常降级。
- 工程化能力:** 熟悉 Git、Docker Compose、Nginx 和 Linux 基础操作, 具备服务容器化、反向代理、SSE 长连接配置及服务器部署经验; 了解 Java、Spring Boot、MyBatis-Flex 和 Vue 3。

项目经历

面向智能硬件客服场景的可治理 Agent 平台

独立开发

2025-10 —至今

项目地址: <https://github.com/sleeping-Zack/agentproject>

项目描述: 面向扫地机器人售后问答、选购咨询及使用报告场景, 构建 RAG + Tool Calling 智能客服 Agent 系统, 集成混合检索、运行预算控制、工具权限治理、HITL 审批、Trace 与 Artifact 持久化等能力, 提升领域知识召回准确性及 Agent 工具调用过程的安全性、可控性和可追溯性。

技术栈: Python、FastAPI、Pydantic、LangChain、Chroma、SQLite、pytest、SSE、MCP JSON-RPC、RAG、Tool Calling

个人职责:

- 针对纯向量检索对部件名和故障关键词召回不足的问题, 实现Dense向量检索与中文BM25双路召回, 并通过RRF进行排名融合; 建立30条独立检索样本, Recall@5由80.56%提升至93.33%, MRR由0.8778提升至0.9333, nDCG@5由0.7683提升至0.9035, P95延迟约313ms。
- 对RRF融合后的Top 20候选使用BAAI/bge-reranker-v2-m3精排并返回Top 5; 测试发现Recall@5由93.3%下降至73.9%, 本地CPU环境平均单查询延迟约29.4s, 因此将Cross-Encoder设计为可选能力而非默认链路。
- 针对并发模型及工具调用“先执行、后记账”导致的预算穿透问题, 实现基于锁保护的请求级BudgetManager; 调用前原子执行reserve, 成功后按实际用量commit, 异常或取消时release, 统一约束步骤数、工具次数、Token、成本和Deadline。
- 构建Claim-Evidence结构化校验、引用合法性检查及选择性LLM-as-Judge, 对高风险或低置信回答进行二次复核; 30条检索样本、12条生成样本和62条Agent Harness固定回归案例, 覆盖工具路由、参数传递、预算超限、Prompt Injection、HITL审批、租户隔离及异常降级, 累计255项自动化测试纳入CI门禁。

项目角色：Agent后端开发

技术栈：Java、Spring Boot、MyBatis-Flex、Vue 3、MySQL、Redis、Qwen、SSE、Docker

项目描述：面向机器人技术咨询、资料检索与售后服务场景，基于AIFlowy开展企业级二次开发，构建包含智能问答、Agent任务规划、QA数据沉淀和售后工单管理的客服工作台，支持管理端、用户端及私有化部署。

- 个人职责：
- 1. 针对简单问题也会进入完整Agent流程、增加响应开销的问题，设计Direct/Agentic两级路由；结合规则判断与Qwen意图识别，将普通寒暄直接交由对话模型，复杂请求再进入知识库、Workflow、插件及MCP工具编排链路。
 - 2. 实现Agent动态规划模块，根据用户问题和Bot已绑定能力生成3~5步执行计划，并增加工具类型、工具ID、参数合法性校验及异常降级；通过SSE向前端同步计划生成、步骤执行和任务完成状态。
 - 3. 开发QA数据闭环模块，从历史会话中抽取User-Assistant问答对，完成消息配对、低质量内容过滤、消息ID幂等去重、人工审核及Markdown批量导出，支持优质问答持续回流知识库。
 - 4. 开发用户端与管理端售后工单流程，支持问题类型、优先级及待处理、处理中、待补充、已解决、已关闭5类状态流转，并依据用户身份实现工单数据隔离。
 - 5. 使用Docker Compose编排后端、前端、MySQL和Redis等服务，通过Nginx完成API反向代理、静态资源访问及SSE长连接配置，支持平台在企业内网环境私有化部署。

个人招聘知识 Agent

2026.06—2026.07

项目地址：<https://github.com/sleeping-Zack/myagent>

项目描述：

面向 HR 和技术面试官构建个人招聘知识问答网站，将教育背景、实习经历、项目职责和技术能力整理为结构化知识库，通过 RAG 检索、证据判断、引用展示和流式生成回答用户问题。

技术栈：Python、FastAPI、PostgreSQL、pgvector、SQLAlchemy Async、BGE Embedding、DeepSeek、SSE、Docker Compose、Nginx

- 个人职责：
- 1. 独立完成个人招聘知识 Agent 的产品设计与系统架构，基于 FastAPI、SQLAlchemy Async、PostgreSQL 与 pgvector 搭建后端，设计项目、知识分块、匿名会话、消息、引用和反馈等数据模型。
 - 2. 负责完整 RAG 链路开发，接入 BGE Embedding 与 pgvector 向量检索，融合标题、标签和项目相关性进行排序，并通过证据充分性判断与保守拒答机制降低无依据生成。
 - 3. 实现 DeepSeek SSE 流式问答与多轮会话记忆，支持历史会话恢复、滚动摘要、上下文 Token 预算、消息幂等和同一会话并发生成控制。
 - 4. 负责系统安全与数据隔离，完成匿名令牌哈希存储、会话所有权校验、接口限流、管理端防爆破、Host 校验、敏感信息脱敏及 Prompt Injection 防护。
 - 5. 负责项目上线与质量保障，使用 Docker Compose 编排 Web 和 PostgreSQL，通过宝塔 Nginx 配置 HTTPS、反向代理与 SSE 长连接，并建立健康检查、数据库迁移、自动清理及测试回归流程。

竞赛与荣誉

- 2024 年第十五届蓝桥杯全国软件和信息技术专业人才大赛大学 A 组省二等奖，具备一定算法基础与编程实现能力。
- 大学生创新创业大赛校级银奖，作为组长参与项目方案设计、原型验证与展示答辩。