

朱旭

求职方向: AI Agent开发 / Agent后端开发 (2027届)

核心摘要:

- 独立开发可治理Agent: 完成单体ReAct 1.0到AgentRunner Harness控制面2.0升级, 覆盖执行路由、共享预算与审批, 支持Direct / ReAct / Plan-Execute、长期记忆、多租户服务化与可重放SSE
- RAG Benchmark: 30条冻结集Recall@5 80.56%提升至93.33%, 门禁识别劣化Rerank; 600+ Pytest入CI
- 法奥校企项目构建168条/26类RAG Gold Dataset与79个证据块, 落地可观测Plan-Execute与HITL QA闭环

在线演示: <https://yaoanxin.xyz/>

邮箱: zhuxu2825@gmail.com

电话: 13895932825

教育背景

南昌大学

软件学院 | 软件工程 | 211 本科

2023.09 - 2027.07

项目经历

智能硬件客服可治理Agent平台

项目角色: 项目Owner

2025.12- 至今

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

项目描述: 面向智能硬件客服问答与复杂服务任务的可治理 Agent 平台, 覆盖产品咨询、使用指导与故障排查等场景。系统通过 RAG 知识检索、Direct / ReAct / Plan-Execute 多模式路由及工具编排完成问题处理, 并以预算控制、敏感操作审批、答案验证和全链路 Trace 保障执行可控、结果可追溯, 支持多租户服务化部署。

技术栈: Python、FastAPI、LangChain/LangGraph、Chroma、PostgreSQL、Redis、SSE、MCP JSON-RPC、OpenTelemetry

个人职责:

1.Agent运行时架构升级

- 将LangChain ReAct封装为执行后端, AgentRunner统一状态、ToolPolicy、敏感工具审批、Verifier、Artifact与Trace
- BudgetManager默认约束8 steps, 8 tool calls和32k tokens; 调用前reserve, 完成后commit/release, 避免并发超卖

2.执行路由与任务编排

- TaskRouter选择Direct/ReAct/Plan-Execute, 低置信或模型异常时回退确定性规则
- Plan-Execute由Planner生成子任务、工具输入和步骤依赖, Runner校验工具与预算, 支持无依赖并发及动态升降级

3.RAG Benchmark与证据链

- 将纯Dense改为Dense + 中文BM25 + RRF, 并统一Evidence/Citation; 30条冻结集Recall@5 80.56%到93.33%朴素Rerank
- 使Recall@5降至73.89%; 90:10融合恢复基线但CPU均耗约10s、未过1s门禁, 故默认关闭

4.Agent Verify与自动化评测

- 搭建Agent Harness全链路评测, 按执行成功、路由、必需工具、Evidence / Citation一致性及安全性分层评估, 记录Tool轨迹、Token、延迟与失败步骤, 支持版本对比和Bad Case回归
- 设计Verifier与CI门禁: 执行前校验工具参数、权限和预算, 执行后核验工具结果与答案证据; 600+ Pytest覆盖审批、超时重试、熔断降级及SSE时序, 异常触发重试或降级

法奥机器人智能客服平台 | 南昌大学 × 法奥机器人校企合作

项目角色: Agent后端开发

2026.01 - 2026.06

项目描述: 面向协作机器人技术咨询、资料检索与售后服务, 打造企业级智能客服平台; 集成复杂文档解析、RAG问答、Agent任务规划与HITL QA闭环, 支持管理端、用户端及私有化部署, 提升知识获取效率与服务可追溯性。

技术栈: Java、Spring Boot、Agents-Flex、TinyFlow、MyBatis-Flex、Python、MySQL、Redis、Qwen、SSE、Docker

个人职责:

1. 复杂PDF入库与结构保真

面向扫描件/图文混排/跨页表格, 引入版面识别、OCR、表格结构还原与Markdown中间表示, 保留标题、页码和层级

2. RAG Benchmark与Failure Analysis

构建168条Gold Dataset (26类/164可答/4拒答) 与79个页级证据块; 以SSE Trace支撑7层评测与Bad Case归因

3. Plan-Execute与工具治理

Qwen按Bot绑定的Knowledge / Workflow / Plugin / MCP生成计划, 自评修订后校验tool type/id/name; SSE同步步骤与节点状态

4. Human-in-the-loop QA飞轮

按Bot/时间窗抽取User→Assistant对, 以消息ID幂等去重并过滤系统错误/噪声; 人工复核后按标签批量导出Markdown

个人招聘知识Agent (已上线)

基于RAG + DeepSeek + SSE构建面向HR的个人问答系统, 支持流式回答与来源引用, 在线体验: <https://yaoanxin.xyz/>

专业技能

- Python后端: Python / FastAPI / Pydantic / Pytest; RESTful API、SSE、参数校验; PostgreSQL / MySQL / SQLite / Redis
- Agent开发: LangChain / LangGraph、Tool Calling、ReAct、MCP / JSON-RPC; 执行路由、Plan-Execute、ToolPolicy、共享预算、敏感工具审批、长期记忆、多模型路由、答案验证、Trace与异常降级
- RAG工程: 复杂文档解析、Chunk、Embedding、Dense / BM25 / RRF、Rerank、Evidence / Citation; Chroma / BGE
- 工程化: Git / Docker Compose / Nginx / Linux / GitHub Actions; OpenTelemetry、Prometheus与CI回归; Java / Spring Boot / MyBatis-Flex / Agents-Flex / Vue 3