

李浩渊

AI Agent 应用开发工程师 / 大模型应用开发工程师

137 3199 1967 935605409@qq.com 杭州

GitHub: github.com/LHYintheCode 博客: blog.csdn.net/ghostinthecode

求职意向: AI Agent 应用开发工程师 / 大模型应用开发工程师 | 意向城市: 杭州

个人优势

- 6 年软件/算法工程经验, 近 3 年聚焦大模型应用、Agent 平台、RAG 知识库、智能报告生成与企业级 AI 应用工程化, 具备从需求拆解、方案设计、接口开发、前后端联调到私有化部署的完整交付经验, 并有腾讯 WeKnora 官方合并的开源贡献经历。
- 具备 AI Agent 应用平台建设经验, 理解主/子智能体协作、Skill/MCP 工具扩展、ChatRun 流式执行、SSE 实时响应、上下文压缩、分层记忆、沙箱代码执行与自动化任务调度等核心链路。
- 主导/深度参与智能报告、知识库解析、RAG 检索与 Agent 办公平台项目, 覆盖十余个 Skill/MCP 工具接入、数千级文档处理、9 类文档格式解析、4 类报告生成与 Docker Compose 私有化部署。

IT 技能

AI Agent 应用开发: AgentScope、ReAct Agent、Multi-Agent、Tool/Function Calling、Skill 机制、MCP 工具接入 (stdio / Streamable HTTP)、工具动态注册、运行态能力装配

大模型应用工程: Prompt Engineering、RAG、工作流编排、长上下文治理、历史摘要/工具结果压缩、引用溯源、结构化输出、异步任务、流式响应与人机协同确认

RAG 与知识库: 文档解析、Chunking、Embedding、Milvus、混合检索、Reranker、父子 Chunk、权限过滤、引用溯源、GraphRAG / 知识图谱预研

后端与工程化: Python、Go、FastAPI、asyncio、SSE、MySQL、PostgreSQL、Redis、Elasticsearch、Docker Compose、Linux/Shell

前端与联调: React、Ant Design、Zustand、Vite, 可完成大模型应用页面、流式消息、工具调用卡片、后管配置与接口联调

模型与平台: Qwen、DeepSeek、ChatGLM 等模型接入; Dify 工作流/知识库、Milvus 单机部署、MinerU 文档解析模型部署

工作经历

杭州量知科技有限公司 · AI Agent 应用开发工程师 · 杭州

2025.06 - 至今

- 负责公司大模型应用与 Agent 平台研发, 参与 AI Cowork 智能体办公平台的 Agent 执行链路、能力接入、私有化部署与线上问题排查。
- 主导/独立承担智能报告生成、知识库解析与 RAG 检索等大模型应用项目, 覆盖 Prompt 设计、文档解析、向量检索、工具调用、FastAPI 接口、SSE 流式输出与 Word 文档生成。
- 参与 Skill、MCP、知识库接入、分层记忆、沙箱代码执行、Artifact 文件资产和自动化任务等核心能力建设; 完成 Docker Compose 私有化部署与环境排障 (7+ 服务组件)。

北明软件有限公司 · 算法工程师 · 合肥

2020.04 - 2025.05

- 参与电网行业解决方案的系统研发、数据处理与知识库维护, 支持检索、问答、数据分析与报表需求; 负责算法模块、接口模块与技术文档编写。
- 后期持续学习实践大模型应用开发, 为转向 RAG、Agent 与企业 AI 应用方向奠定基础。

项目经历

1. AI Cowork 智能体办公平台 · 杭州量知科技 · 参与并部分主导 · 2025.12 - 至今

以「主智能体推理 + 子智能体协作 + Skill/MCP 扩展 + 知识库检索 + 分层记忆 + 沙箱代码执行 + Artifact 资产 + 自动化任务」为核心架构的企业级 Agent 办公平台。

- 基于 AgentScope ReActAgent 构建主智能体执行链路，按会话动态装配模型、系统 Prompt、Skill、MCP、知识库、子智能体与代码执行工具。
- 建设 ChatRun + Redis Stream + SSE 流式运行链路，支持任务续播、取消、异常恢复、工具调用事件追踪与消息落库。
- 接入 stdio、Streamable HTTP 等 MCP 工具形态，覆盖数据库查询、网页/知识库检索、图表生成、报告导出等办公场景；通过连接池与生命周期管理避免泄漏。
- 建设上下文治理、分层记忆、沙箱代码执行与 cron 定时 Agent 任务能力。

成果：接入十余个内部 Skill/MCP 工具；支撑数千级文档知识库；完成 7+ 组件 Docker Compose 私有化部署。

技术栈：AgentScope、ReAct Agent、MCP、SSE、Redis Stream、mem0、FastAPI、PostgreSQL、Redis、Milvus、Neo4j、Docker Compose、React

2. 企业知识库解析与 RAG 检索服务 · 项目主导 · 2025.09 - 2025.12

面向企业知识库的文档解析、向量化入库与检索增强服务，参考 RAGFlow 文档处理思路，为 Agent 平台与报告系统提供统一知识底座。

- 设计 9 类文档解析链路（PDF/DOCX/WPS/XLSX/CSV/TXT/Markdown 等），老版本 Office 经 LibreOffice 转换兼容。
- 接入并部署 MinerU 解析模型处理复杂 PDF/扫描件，失败自动降级；设计父子 Chunk 切分、混合检索、权限过滤与父块回填。
- 预研 RAGFlow 知识图谱接入方案，以实体/关系抽取增强多跳问答。

成果：支撑数千级文档处理，覆盖 9 类办公文档格式，成为平台统一 RAG 检索底座。

技术栈：Python、FastAPI、MinerU、LibreOffice、Pandoc、Milvus、Reranker、Elasticsearch、Docker

3. 智能报告生成服务 · 项目主导 · 2025.06 - 2025.09

面向产业/招商/技术报告与标准化 Word 报告的智能生成服务，「数据 + 模板 + RAG + 大模型生成」的业务落地。

- 设计「参考大纲解析 + 业务大纲融合 + LLM 大纲生成」流程；基于 FastAPI + asyncio 构建章节级并行生成机制（并发、超时、重试、顺序聚合）。
- 每章节接入网络检索与知识库工具；建设引用溯源流程（来源整理、去重、编号校验、规范引用标记）。
- 标准化报告按模块组织数据查询、图表生成与模板填充（MySQL/Elasticsearch 数据 + docxtpl + ECharts）。

成果：支持 4 类报告生成，将人工数小时产出压缩至分钟级初稿。

技术栈：Python、FastAPI、Milvus、Elasticsearch、MySQL、asyncio、SSE、docxtpl、ECharts、OSS/MinIO

4. 个人 AI 应用与开源工具实践

- AI 合同自动生成应用：大模型抽取合同关键字段自动填充模板，单份合同由约 5 分钟降至 30 秒。
- 播客摘要 MVP：独立完成产品设计与开发，支持摘要、金句抽取、时间戳定位与点击跳转播放。
- AI 编程记录归档 CLI (Go)：合并不同 AI 编程工具聊天记录，支持归档、分类与结构化整理。

开源贡献

Tencent WeKnora • Contributor • 2026.07 —— 修复 TaskInspector 周期性误报惰性 asynq 队列不存在的问题：定位 asynq v0.26.0 Inspector.GetQueueInfo 透传内部 NOT_FOUND 错误的根因，实现兼容判定；补充 sentinel、wrapped error 与 miniredis 回归测试。PR #2218 经维护者直接合并至 main，关联 Issue #2139 解决。

教育背景

新南威尔士大学 (UNSW Sydney) · Data Science and AI 硕士 · GPA 3.2/4.0 2017.09 - 2020.01
沈阳航空航天大学 · 计算机科学与技术 本科 · GPA 3.5/4.0 2013.08 - 2017.06

其他情况

英语：雅思 6.5，具备英文技术文档阅读能力。 | 技术博客：blog.csdn.net/ghostinthecode | GitHub：github.com/LHYintheCode