殷晗

求职意向：AI Agent 后端开发 / AI 应用开发工程师

武汉｜9年Java后端 · 2年以上AI应用工程｜长期定居武汉
本科 · 软件工程（2013.09—2017.06）
13135661502｜yinhanxy@aliyun.com
在职交接中，**预计2026年11月上旬到岗**｜期望薪资：面议

专业概述

9年Java后端、4年技术小组管理经验，近年专注私有化AI应用。基于Spring AI、LangChain4j与LangGraph开展Agent工具编排、上下文治理和RAG建设，具备政企信创与气象行业交付经验。

核心技能

分类	技术栈	应用方向
Agent工程	Spring AI、LangChain4j、LangGraph、Function Calling、MCP	工具编排、Token预算、记忆与摘要、SSE
RAG与检索	pgvector、Milvus、Elasticsearch、HNSW、Caffeine、Ragas	结构/父子分块、混合召回、RRF、重排与评估
Java与微服务	Java 8/17、Spring Boot、Spring Cloud Alibaba、MyBatis	并发、JVM、服务拆分、鉴权与限流
数据与集成	PostgreSQL、MySQL、Redis、RabbitMQ、MinIO、OAuth2	缓存、异步解耦、文件治理、业务接口集成
权限与交付	RBAC、Flyway、Docker/K8s、Nginx、Jenkins	部门范围、审核、私有化部署、达梦/东方通适配
前后端协作	Vue 3、TypeScript、Vite、Git、Maven	页面联调、流式交互、工具结果展示

工作经历

深圳市中讯网联科技有限公司｜AI应用开发工程师　2026.02—至今

气象行业私有化AI平台：Synapse提供Agent、RAG与工具能力，AHAI承接业务编排、权限、知识管理及场景交互。
技术栈： Java 17、Spring Boot 3.5/4、Spring AI 2、PostgreSQL/pgvector、MinIO、Vue 3。

一、Agent上下文治理（Synapse）

- 针对约42万字符工具输出，采用主体截断与完整数据服务端旁路，结合独立工具链缓存、上下文快照支持脚本参数注入；配套Token预算、历史摘要与上下文截断。
- 历史单场景截断记录：prompt token由**917,403降至6,945**；该阶段后续工具未续接，数字仅说明上下文体量变化。

二、RAG混合检索（Synapse）

- 结构感知分块，结合pgvector语义检索与PostgreSQL全文检索，通过RRF融合并支持重排；配套HNSW索引配置、Caffeine缓存和监控，以开放接口支撑业务接入。

三、业务集成与权限/存储治理（Synapse / AHAI）

- **Synapse**：接入脚本、Python绘图、MCP及工作流工具，自定义工具保存时执行注册校验；治理资源共享与部门可见性，处理知识删除时数据库、向量索引和文件的级联清理。
- **AHAI**：建设知识文件审核、批量操作权限预检；对接OAuth2令牌缓存与SSE代理，衔接气象图表交付、MinIO文档版本及OnlyOffice保存回调。

四、后端性能与交付（AHAI）

- 将组织子树人数统计改为一次聚合，配合用户部门字段索引，测试环境查询**P50由452ms降至9—10ms**；通过刷新令牌唯一标识与upsert处理同账号并发登录冲突。
- 使用Flyway、健康检查、容器配置与自解压安装脚本组织交付。2026年10月测试环境压测记录：**100并发混合业务持续90秒，接口P95≤106ms**；服务器本机经Nginx测试，不含模型推理与Agent端到端容量。

北京芯盾时代科技有限公司｜核心后端开发　2021.10—2026.01

带领5—8人研发小组，承担需求拆解、技术评审、Code Review与版本迭代。

企业安全AI智能体中台　2024.03—2026.01

技术栈： Java 17、LangChain4j、LangGraph、Milvus、Gateway、Redis、RabbitMQ、Ragas。

- 封装对话记忆、工具调用与提示词管理，接入身份校验、访问控制、工单审批等业务工具，结合LangGraph编排多步骤任务。
- 建设私有化RAG知识库：父子分层分块、向量/BM25混合检索、重排与增量更新；使用Ragas对照评估，定位知识处理、检索和生成环节问题。
- 接入模型网关、Sentinel限流、Token统计与ELK日志，结合CI/CD及达梦、东方通适配完成政企离线环境交付。

SDP零信任业务安全平台　2021.10—2026.01

技术栈： Spring Boot、Spring Cloud Alibaba、MyBatis、Redis、MySQL、ETCD、Docker。

- 主导单体系统微服务拆分，支持独立扩缩容与灰度发布；通过连接池、多级缓存、热点Key治理、慢SQL优化及异步化处理核心接口性能问题。
- 建设CI/CD流水线与统一异常、错误码规范，衔接版本交付和故障定位；完成MySQL向达梦、Tomcat向东方通的迁移适配。

武汉捷讯信息技术有限公司｜中级Java工程师　2019.09—2021.09

武水客户服务系统｜Spring Boot、MyBatis、Redis、MySQL、Nacos、RabbitMQ、Elasticsearch。

- 负责客户信息、工单、通知与第三方接口；通过Redis原子操作实现接口幂等与工单编号生成，RabbitMQ解耦通知和延时推送。
- 使用Spring AOP与自定义注解统一审计日志；以线程池和CountDownLatch并行处理批量导入，配合Elasticsearch实现检索与日志查询。

武汉世纪卓创科技发展有限公司｜Java工程师　2017.06—2019.09

参与销售、小程序及调研系统，承担业务后端模块和第三方接口开发；项目场景见下表。

早期项目补充

项目	场景与职责
慧经销	劲酒销售下单、客户拜访、经营统计；订单、统计接口及第三方对接
中怡宸信小程序	加油、送油、积分兑换；订单支付与核销流程
互助调研系统	自定义问卷、消息推送、统计分析；问卷配置及业务接口

自我评价

关注AI应用中的上下文膨胀、工具失败、权限边界与数据一致性。兼具后端开发和小组协作经验，能够配合业务、前端及现场团队完成联调与私有化交付。

公开作品

在线演示： AI工程实验室｜知识治理、混合检索与上下文治理演示，附版本保护和SQL优化案例（合成数据）。
GitHub： yinhanxy/hybrid-rag-spring-ai｜独立教学样例：H2、哈希嵌入与RRF，不作生产效果证明。
CSDN： Agent长会话上下文治理实践｜主体截断、旁路缓存与参数注入取舍；以上均脱敏编写。