施奇

AI应用开发工程师

男 | 36岁 | 14年经验 | 合肥
15375337717 | 591990368@qq.com
Gitee：gitee.com/shiqi_2
项目预览：企业级安全自动化运维 Agent
期望薪资：20-30K

AI应用开发工程师方向，具备 RAG、Agent、Tool Calling、Text-to-SQL 、LoRA 轻量微调等企业级 AI 应用落地实践。拥有14年后端开发与分布式系统经验，主语言 Java，具备 Java / Go / Python 跨语言工程交付能力，擅长将大模型能力接入企业业务系统、数据平台和自动化流程。

LLM / RAG / Agent | Tool Calling | Text-to-SQL | 模型服务接入 | 数据处理 / 链路追踪 | Java / Go / Python | Spring Cloud / Dubbo / FastAPI / Flask | MySQL / PostgreSQL | Redis / MQ / ES

个人优势

- **AI应用开发与落地：**具备 LLM、RAG、Agent、Tool Calling、Text-to-SQL 的应用实践经验，能够基于业务场景设计并实现企业级AI系统。
- **RAG与数据智能：**具备“知识检索 + 接口调用 + SQL生成”的混合方案设计能力，提升回答准确性，同时兼顾权限控制与业务可解释性
- **Agent与工程控制：**具备Agent执行流程控制经验，能够实现工具调用约束、审计日志、异常处理及系统稳定性设计。
- **工程与系统能力：**14年Java与分布式系统经验，熟悉Spring Cloud、Dubbo、Redis、MQ等技术，能够支撑AI应用工程化落地。
- **数据与检索能力：**具备大规模数据处理经验，熟悉Elasticsearch、向量检索（Qdrant）、数据建模与查询优化。
- **跨语言交付能力：**具备 Java / Go / Python 跨语言工程交付能力。
- **项目推进与协作：**具备跨团队协作与项目推进经验，能够推动AI应用从方案到落地闭环。

项目经历

企业级运维自动化 AI Agent（Go实现 | RAG + Tool Calling + 审批控制） AI应用开发 2025.06 - 至今

（核心模块实现）

项目描述：面向企业运维与安全场景，基于 LLM 构建自动化运维 Agent，实现告警分析、知识检索、处置建议、审批控制与执行审计的闭环流程，替代部分人工运维处理，提高系统处置效率与安全性。

- 基于 Go 实现运维自动化 Agent，设计“感知-决策-审批-执行-审计”闭环流程。
- 构建 RAG 运维知识库，接入 SOP、故障手册与历史处置记录，提升告警语义理解与上下文补全能力
- 基于 Tool Calling 封装监控、日志、工单等系统接口，将 LLM 决策转化为可控工具调用
- 设计审批与安全控制机制：高风险操作需经过规则引擎、权限校验与审批流程，避免模型直接执行生产操作
- 实现任务执行控制机制，支持重试、补偿、失败回滚及人工接管
- 构建 Agent 可观测体系，沉淀 Audit Log、Metrics 与 TraceID，实现执行链路追踪与行为审计

项目成果：完成可演示、可部署的运维 Agent 原型系统，覆盖告警分析、知识检索、工具调用、审批控制与执行审计等核心流程；在模拟告警与常见运维场景中，可减少人工查询 SOP、定位日志和整理处置建议的时间，为后续生产环境接入提供基础能力。

技术栈：Go / LLM（Ollama / Qwen / DeepSeek） / Qdrant / RAG / Tool Calling / Redis / SQLite / REST API

企业数据智能问答助手（Java实现 | RAG + Text-to-SQL + 权限控制） **AI应用开发（核心模块实现）** 2025.08 - 至今

项目描述：针对企业知识碎片化与结构化数据孤岛问题，设计并开发基于 RAG + Agent 的智能决策助手，通过自然语言统一接入文档检索、指标查询与数据分析能力，已完成核心能力构建，部分模块持续优化中。

- 设计 Hybrid Query Engine，将请求路由至 RAG 检索、指标接口或 Text-to-SQL 分析，实现多数据源统一访问。
- 构建 RAG 知识检索模块，接入 SOP 文档与业务资料，支持语义检索与上下文增强。
- 实现 Text-to-SQL 能力，基于 Schema 信息生成查询语句，并设计 SQL 安全审计与权限控制机制。
- 构建分层记忆机制：Redis 管理短期对话上下文，Qdrant 存储语义记忆，用于历史信息召回。
- 基于异步非阻塞链路处理 LLM 与 Embedding 请求，提升系统吞吐能力。
- 引入 OpenTelemetry，实现 AI 推理流程、工具调用及查询执行的链路追踪（持续优化中）。

项目成果：

- 数据取数效率提升：面向运营、财务等非技术部门提供自然语言取数能力，覆盖大量高频临时查询场景，将部分取数需求从“人工排期/天级响应”缩短至“自助查询/秒级返回”
- 查询准确率提升：针对核心经营指标采用“预埋接口 + 指标口径固化 + SQL 自我修复”机制，保障高频指标查询结果稳定可靠；针对复杂长尾问题，通过 SQL 生成、校验与重试机制提升查询成功率。
- 安全合规落地：采用私有化本地部署方案，敏感经营数据不出局域网；通过只读从库、权限控制、SQL 审计与高危语句拦截，降低对线上生产库的影响和数据泄露风险。

技术栈：Java / Spring Boot / LLM / RAG / Text-to-SQL / Redis / Qdrant / OpenTelemetry

---

企业级运维大模型轻量微调管理平台（Python实现 | Python + FastAPI + LoRA） **AI应用开发（核心模块实现）** 2025.10 - 2025.11

项目描述：基于 Python + FastAPI + Transformers + PEFT + LoRA 构建运维告警场景下的大模型轻量微调验证项目，支持 Alpaca 格式训练数据上传、LoRA 微调任务创建、训练日志查看、Adapter 输出管理和基础模型/微调模型推理测试。项目不从零训练大模型，重点验证大模型轻量微调、任务管理和推理服务封装的工程化链路。

- 基于 FastAPI 封装数据集管理、训练任务管理、日志查询和模型推理等 HTTP 接口，并提供 Swagger 文档。
- 设计 Alpaca 格式训练数据上传与校验流程，支持样本预览和训练数据管理。
- 基于 Transformers、PEFT、LoRA 实现 Qwen2.5-0.5B-Instruct 的轻量 SFT 微调流程。
- 使用 JSON 文件维护任务状态，通过 subprocess 后台执行训练脚本，支持训练日志查看和 Adapter 输出管理。
- 实现 LoRA Adapter 输出目录管理，支持基础模型和 LoRA 微调模型的推理测试，用于对比微调前后的输出效果。
- 设计 Mock Demo 模式，支持无 GPU、无完整模型依赖环境下演示上传、训练任务、日志和推理结果流程。
- 提供简单 Web 操作页面，覆盖数据集上传、训练任务创建、任务状态查看、日志查看和告警分析推理测试。

项目成果：

- 完成从数据上传、格式校验、微调任务创建、训练日志跟踪、Adapter 输出到模型推理测试的完整大模型轻量微调工程链路。
- 将 LoRA 微调能力封装为可视化页面和标准 HTTP 接口，降低模型微调与推理测试的使用门槛。
- 支持 Mock Demo 模式，适合在普通服务器环境下进行项目展示，避免模型下载、GPU 环境和推理依赖对演示的影响。
- 在运维告警分析场景中，通过同一组告警输入对比基础模型与 LoRA 模型输出，验证微调后在输出格式、运维术语理解、排查步骤完整性和处理建议可执行性上的提升。

技术栈：Python / FastAPI / Transformers / PEFT / LoRA / Qwen2.5 / torch / Docker Compose

亿级分布式电商全链路中控与业务治理平台 系统架构设计 / 核心开发 2025.06 - 至今

项目描述：针对业务增长带来的数据割裂与单体架构瓶颈，主导将核心 ERP 重构为微服务架构，并构建统一数据中控平台，实现订单、库存、售后等多源数据治理与链路稳定性提升。

- 基于 Spring Cloud + Dubbo + Nacos 拆分核心模块，设计 MQ + Redis 异步处理链路，支撑千万级日增数据实时分发。
- 设计“实时消费 + 离线补偿 + 人工回放”闭环机制，结合 XXL-Job、分布式锁与状态机保障最终一致性。
- 基于 PostgreSQL 分区与 Elasticsearch 索引优化，实现亿级数据毫秒级查询。
- 基于 Flowable 实现复杂业务流程编排，支持订单等场景标准化审批与数据留痕。

项目成果：支撑 order_action_logs 单表 7.32 亿数据规模稳定运行，核心链路可用性提升至 99.99%，解决订单同步丢失与高延迟问题。

技术栈：Java / Spring Cloud / Dubbo / MQ / Redis / PostgreSQL / Elasticsearch

分布式跨境电商 ERP 综合管理平台 系统架构设计 / 核心开发 2023.03 - 2025.05

项目描述：面向 Shopee、TikTok、速卖通等跨境平台，构建订单、商品、物流、库存、财务等一体化 ERP 系统，支撑多租户、多平台、高频 API 同步场景。

- 基于 Spring Cloud Alibaba + Dubbo 构建微服务架构，支持多平台订单、库存与物流统一管理。
- 设计多租户权限体系与高并发接口链路，提升系统稳定性与扩展能力。
- 通过 Redis 缓存与异步处理机制优化接口性能与限流控制。

项目成果：核心接口响应降低40%，并发处理能力提升3倍以上。

亿级知识产权大数据检索平台 技术负责人 2017.03 - 2022.12

- 设计 MySQL + MongoDB + Elasticsearch 多存储架构，实现亿级数据高效检索。
- 优化分词、路由与索引策略，支持大规模数据秒级查询。
- 负责系统架构演进与团队技术规范建设。

项目成果：在数据增长10倍情况下，检索性能保持毫秒级响应。

工作经历

安徽俏美味食品有限公司 Java / 架构方向 2025.05 - 至今

- 在系统中引入 LLM 能力，主导AI Agent 与数据智能模块开发，推动 AI 在业务场景中的落地应用。
- 主导高并发数据中控平台研发，统一治理订单、售后、库存等多源数据，构建 MQ + Redis 异步处理链路。
- 建立基于 XXL-Job、分布式锁与状态机的数据补偿体系，解决亿级数据规模下的数据一致性问题。

合肥谦博贸易有限公司 架构 / 全栈开发 2023.03 - 2025.05

- 基于 Spring Cloud Alibaba + Dubbo 3.0 构建跨境电商 ERP 微服务底座，集成 Nacos 实现配置与服务治理。
- 设计 Sa-Token 多租户鉴权体系，实现租户逻辑隔离、动态权限控制与组织架构授权。
- 负责 Shopee、TikTok等平台 API 对接与高并发适配。

龙图腾网科技（合肥）股份有限公司 技术部主管 / Java 2014.08 - 2022.12

- 负责知识产权系统架构演进，设计 MySQL + ES + MongoDB 多存储架构。

- 优化检索与数据处理链路，实现亿级数据高效查询。

早期经历：北京中科汇联科技股份有限公司、上海智青科技有限公司

技术栈

AI 应用	LLM应用开发（RAG / Agent / Tool Calling / Text-to-SQL）、Ollama、Qdrant、OpenTelemetry
后端架构	Java、Go、Python、Spring Boot、Spring Cloud Alibaba、Dubbo、MyBatis、XXL-Job
数据与存储	PostgreSQL、MySQL、Elasticsearch、Redis、MongoDB、宽表建模
工程与治理	微服务、MQ异步链路、分布式锁、状态机、幂等与补偿机制、限流熔断、Docker

教育经历 / 证书

安庆师范大学 本科 信息管理与信息系统 2008 - 2012 **证书：**大学英语四级