



教育背景

2023-2027

石家庄铁道大学

软件工程(本科)

荣誉奖项: 蓝桥杯省级三等奖; 计算机设计大赛省级二等奖; 大学生创新创业大赛省级奖项; 优秀团员。

专业技能

- 编程语言:** 掌握 JAVA/Python 基本语法, 熟悉 SpringBoot、FastAPI、Flask 等, 以及 Pandas、NumPy 进行数据处理与分析。
- 数据库:** 熟练使用 MySQL、Redis、Neo4j, 能够完成数据建模、日常存取、查询优化及业务数据处理工作。
- 算法基础:** 掌握常见机器学习算法, 熟悉模型评估指标(Accuracy、Recall、F1、AUC)及常见损失函数(BCE、Focal Loss)。
- 模型应用:** 掌握 LangChain、GraphRAG、向量库(Milvus/Chroma)、MCP、Function Calling、Prompt (Few-shot、CoT)。
- 模型微调/部署:** 熟悉 PyTorch 及 LoRA/QLoRA 微调, 掌握超参数调优, 以及 vLLM、Ollama 等推理与部署工具。
- Agent 能力:** 熟悉 Dify、Coze、Flowise、n8n、LangGraph 等, 具备多 Agent 架构设计能力, 能够基于 ReAct、Reflection、Plan-Execute-RePlan 实现任务拆解与协同执行; 深入理解 Harness 的工程化理念, 掌握 AgentOps 及可观测性体系建设。
- AI 工具:** 熟悉 CodeX、Claude Code、Cursor、Trae、codebuddy 等 AI 工具, 擅长 Vibe Coding, 做过多个项目。

实习经历

2025.12-2026.02

杭州米里云软件有限公司

智能电商风控与客服 Agent 平台

Agent 开发实习

- 项目背景:** 为杭州优勤网络科技有限公司定制开发, 整合订单、用户、售后与风控全量业务数据, 实现交易风险自动拦截与用户咨询智能应答。日均处理高频交易与客服交互, 降低人工审核漏判率, 提升风控与客服响应效率。
- 技术栈:** LangChain、LangGraph、BERT、BM25、BGE-M3、RRF、BGE-Reranker、Milvus、Qwen3-14B、LoRA 微调
- 1. Agent 设计:** 参与 Agent 设计, 按意图划分交易风控/账号安全/售后问答三类 Agent, 结合 ReAct 实现工具调用与执行闭环。
- 2. RAG 优化:** 结合混合检索、RRF 融合与 Rerank 精排, 基于 Milvus 支持向量增量更新, 召回率提升 15%+、命中率提升 10%+。
- 3. NL2SQL 能力:** 将自然语言多意图解析为 SQL, 提升电商订单、用户行为及风控数据的复杂查询与工具调用成功率。
- 4. 多轮记忆:** 设计分层记忆机制(短期对话缓存+长期向量记忆), 结合滑动窗口压缩与关键信息抽取, 将用户历史行为、关键意图与上下文状态动态注入 Prompt, 有效缓解长对话信息遗忘问题, 提升多轮风控与客服场景下的对话一致性与决策稳定性。
- 5. 风控策略:** 设计可配置风控规则引擎, 支持实时策略调整与灰度发布, 实现交易异常自动拦截, 人工审核漏判率降低约 30%。
- 6. 模型微调:** 构建 5000 条电商风控与客服高质量指令数据集, 基于 Qwen3-14B 进行 LoRA 微调, 针对多意图识别与工具调用场景进行优化, 使工具调用成功率由 87.3%提升至 95.2%, 意图识别准确率由 91.2%提升至 93.5%;
- 7. 评测优化:** 参与检索评估与调优, 基于 HitRate、MRR 等指标进行离线评估与对比实验, 持续优化召回效果与排序质量。

项目经历

2026.06-2026.07

智伴助手 Harness 智能体框架

Agent 开发

- 项目背景:** 智伴助手是面向高消费退休老人, 集成银发社交、高端文旅、家属关怀搭载 Harness 语音 AI 智能体的多端 App+Web 平台。
- 技术栈:** FastAPI、Redis、RabbitMQ、Celery、WebSocket、LLM、Harness、Qdrant、ASR/TTS、CDN、Docker、Prometheus
- 1. 框架搭建:** 开发 Harness 智能体运行中心, 实现意图解析/任务规划/多层记忆/循环调度/全链路审计, 打通语音/文字/图片多模态输入输出通道。
- 2. 工具管控:** 设计四角色标准化工具集, 配置 Function Calling 调用规则, 划分 L0-L4 多级操作权限, 新增参数校验、脱敏、防越权验证器。
- 3. 知识工程:** 搭建银发文旅专属知识库(30W 数据), 完成文档清洗、切块、向量入库、RAG 检索链路开发, 区分实时业务数据与静态知识查询逻辑。
- 4. 安全风控:** 构建隐私防护机制, 拦截 Prompt 注入、身份 / 位置 / 健康信息越权访问, 配置高风险场景人工接管流程, 全操作留痕审计日志。
- 5. 评测迭代:** 构建全场景 Agent 测试数据集, 量化意图准确率、幻觉率、任务完成率指标, 完成模型不可用时功能降级方案开发。
- 6. 业务集成:** 对接老人/家属/文旅/运营四端业务 API, 实现 AI 找同伴、行程查询、打卡提醒、线路草稿生成等适老智能场景落地。
- 7. 运维监控:** 集成 Prometheus+LangSmith+OpenTelemetry, 统计接口延迟/调用成本/异常失败工单, 优化向量检索、模型限流、会话缓存性能。

2026.04-2026.06

智旅云图

Agent 开发

- 项目背景:** 该旅游助手系统, 解决旅游信息零散、行程规划繁琐问题, 支持个性化智能行程生成与文旅信息问答服务。
- 技术栈:** LangChain、FastAPI、LangGraph、Chroma、MySQL、Redis、BGE-M3、Harness、Qwen3-Max、MCP
- 1. 系统架构:** 基于通义千问与高德 MCP 自研 MCP-Service 工具基座, 参与设计多角色协同的旅游规划多 Agent 系统。
- 2. Agent 编排:** 基于 LangGraph 搭建主从式多 Agent 工作流, 实现景点、天气、酒店等 Agent 的协同调度与智能行程生成。
- 3. Agent 优化:** 基于 Harness 完善 Agent 评测流程, 设计检索纠错工具重试机制, 结合 RAG 结果校验与补全, 提升任务完成率。
- 4. 数据处理:** 采用 SimHash + MinHash + TF-IDF 组合去重策略处理 10 万+旅游知识库文本数据, 数据处理效率提升 73%。
- 5. RAG 搜索:** 结合 BM25 + BGE-M3 混合检索、RRF 融合与 Rerank 精排, 配合向量热更新机制, 问答匹配效率提升 20%。
- 6. 指标评估:** 完善 RAG 评估体系, 引入 MRR、Noise Rate、Latency 等指标, 优化检索准确率与响应性能。
- 7. 上下文工程:** 设计动态上下文裁剪与历史记忆压缩策略, 基于 Redis 优化 Token 窗口, 降低 LLM 调用成本 35%。