



个人优势

- 2027届211计算机硕士(保研), 拥有3段大厂(包括字节)AI应用开发实习经验, 主导多个从0到1工业级大模型项目落地。
- 精通Python、RAG、多Agent编排, 熟悉LangGraph、LangChain、Dify、Claude Code等主流技术栈。
- 擅长从业务痛点拆解方案, 落地项目均实现显著的效率提升与成本优化。
- 可独立承担完整业务模块, 实习期间全权负责整条业务线开发。

实习经历

北京字节跳动科技有限公司-火山引擎-混合云-AI应用开发(暑期实习)

2026年07月 - 至今

博世(中国)投资有限公司-AI智行学院-AI应用开发

2026年05月 - 2026年07月

沪士电子股份有限公司-资讯部-AI应用开发

2025年07月 - 2026年04月

项目经历

字节跳动-ArkClaw产品DeepResearch专家模版建设

2026年07月 - 至今

- 项目描述:** 负责ArkClaw产品DeepResearch专家模版的设计与开发, 基于多轮检索与信息聚合策略构建自动化深度研究链路, 覆盖选题拆解、多源信息抓取、内容结构化整合与报告生成全流程, 显著降低用户信息检索与报告撰写成本, 提升研究效率。
- 关键业务指标提升:** 研报生成由2-4h降至5min内, 检索耗时降低35%; 上线1个月覆盖15人, 完成80+次研究、生成70+份研报、节省150h+; 200+条Benchmark上幻觉率25%→6%、引用准确率达94%+; 高质量信源35%→78%、低质量信源45%→12%; 意图路由准确率96%, 无效研究调用降低60%; 任务成功率97%+。
- 核心技术:** Python、FastAPI、LangGraph、LangChain、RAG、MCP、Milvus、PostgreSQL、Redis、Langfuse、Grafana、Ragas、Benchmark
- 主要成果与贡献:**
 - 多Agent研究链路编排建设, 解决流程复杂、串行执行耗时高问题: 基于LangGraph编排8类Agent, 采用ResearchState共享上下文、Pydantic规范IO, 并行双源检索使检索耗时降低35%。
 - 意图路由降本优化, 解决简单请求误触发深度研究链路问题: 设计“规则优先+LLM兜底”双模态路由, 准确率96%, 无效研究调用降低60%, Token消耗下降55%。
 - 双源RAG与证据审计建设, 解决信息覆盖不足、信源质量不稳定问题:
 - 双源检索融合: 打通网络搜索与Milvus知识库, 证据覆盖率较单源检索提升40%;
 - 证据质量管控: 实现证据去重、信源评分与冲突检测, 高质量信源占比由35%提升至78%。
 - 反思补搜闭环建设, 解决一次检索信息缺失、研究完备性不足问题:
 - 动态补搜机制: 分析Agent识别证据缺口, 反思Agent生成补充查询并递归触发检索链路;
 - 迭代安全管控: 设置轮数与Token预算双重护栏, 研究完备性满意度由62%提升至89%。
 - 引用溯源与幻觉防控建设, 解决研报结论不可验证、引用失真问题: 通过合法引用约束、规则校验与自动重写实现答案可追溯, 幻觉率由25%降至6%, 引用准确率达94%+。
 - 记忆与评测闭环建设, 解决历史研究难复用、模型迭代无标准问题:
 - 三层记忆搭建: 基于Redis、PostgreSQL、Milvus构建短期、长期与语义记忆, 支持跨会话结论复用;
 - 评测监控落地: 基于Ragas搭建200+条Benchmark, 结合Langfuse Trace定位Bad Case与Agent链路问题。

博世中国(Bosch)-智驾代码及根因定位平台建设

2026年05月 - 2026年07月

- 项目描述:** 面向智驾数据分析误/漏识别痛点, 原有人工排查单例耗时3-5h、高度依赖资深工程师经验; 主导落地多Agent智能根因定位系统, 打通全流程端到端AI辅助链路, 大幅提升排查效率与经验复用率。
- 关键业务指标提升:** 单问题平均处理时长从3.5h降至1min以内, 覆盖部门80%以上高频问题类型; 代码模块定位Top-5召回率90%+, 根因Top-1命中率85%+; 上线2个月累计处理工单600+, 节省人工工时约2000h, 月活工程师40+; 沉淀结构化历史Case库300+条。
- 核心技术:** Python、FastAPI、LangGraph、LangChain、RAG、Neo4j、MCP、Milvus、Redis、Langfuse、Grafana、Ragas、Benchmark
- 主要成果与贡献:**
 - 根因分析评测体系搭建, 解决人工排查耗时长、经验难沉淀、模型迭代无量化标准问题:
 - 流程标准化拆解: 访谈8名资深工程师, 梳理5类标准化排查原子任务;
 - 评测数据集搭建: 标注200+工单构建Benchmark, 设计多维评测指标;
 - 自动化回归落地: 搭建评测链路, 量化验证模型、RAG、Prompt迭代效果。
 - 代码知识库RAG分层优化, 解决百万行代码检索准确率低问题:
 - 向量层优化: 采用jina-code-embeddings代码专用向量模型, 召回率较通用模型提升31%;
 - 图谱增强检索: Tree-sitter构建Neo4j函数调用图谱, 向量+图谱双路检索, 沉淀200+业务-代码映射。
 - 多Agent协作编排搭建, 解决串行分析时延高、模块难迭代问题:

- 多角色链路编排：基于LangGraph编排Triage/Data/Code/RCA/Solution Agent，统一结构化IO，支持独立评测替换；
- 并行调度优化：并发运行Data Agent、Code Agent，端到端延迟由110s降至30s。
- 4. 根因推理可信度优化，解决智能体输出存在引用幻觉、结论可信度不足问题：
 - 引用校验增强：约束RCA Agent绑定代码行号与Bag帧，新增Verifier Agent核验引用，错误内容自动重生成；
 - 置信度分级管控：增设Agent置信度打分与低置信人工复核机制，抽样幻觉率由23%降至6%。
- 5. 历史Case复用闭环建设，解决历史经验无法复用、修复方案准确率低问题：
 - 案例库结构化沉淀：规整结案工单多维字段，搭建300+条历史Case库；
 - 案例检索能力优化：Solution Agent基于Case-based RAG召回相似案例，方案有用率由51%提升至78%。
- 6. 灰度上线与持续迭代机制建设，解决模型上线效果不可控、迭代无闭环问题：
 - 双盲对比验收：对接内部工单系统，落地AI与人工双盲对比，常态化开展每周Bad Case Review；
 - 迭代闭环保障：依托用户打分与Langfuse Trace完成问题归因，结合Benchmark回归测试，保障版本迭代稳定。

沪士电子-云平台智能客服平台建设

2025年09月 - 2026年04月

- 项目描述：面向云平台客服场景中高频标准问题占用大模型算力、复杂云产品参数查询易幻觉、跨会话缺少记忆继承等问题，负责基于多Agent与RAG搭建云客服，实现精准问答、毫秒响应与Token稳定降本。
- 关键业务指标提升：L1语义缓存命中率38%，命中场景首字延迟从3.2s降至80ms，月度LLM Token成本节省约40%；云产品复杂参数召回准确率从60%提升至92%；复杂场景意图路由准确率95%；用户重复陈述轮次减少45%，满意度提升18%；降本建议生成从人工45min缩短至15s，采纳率60%+。
- 主要成果与贡献：
 1. 多Agent协作编排建设，解决业务咨询场景单一能力适配弱、上下文割裂问题：
 - 智能体拆分编排：基于LangGraph状态机搭建4类专家Agent，通过Orchestrator实现业务意图统一路由；
 - 全局上下文联动：依托AgentState共享对话上下文，复杂场景意图路由准确率达95%。
 2. L1/L2双层语义缓存优化，解决高频问答推理延迟高、Token消耗大问题：
 - L1语义缓存部署：网关层对高频标准问题计算Query向量，接入Milvus语义缓存，高置信度直接返回标准答案；
 - L2链路降级策略：未命中/低置信请求转入Agent推理链路，首字延迟由3.2s降至80ms，月度Token开销节省40%。
 3. 图向量混合检索Hybrid RAG建设，解决复杂云产品架构、参数查询召回不准的问题：
 - 向量检索能力搭建：依托Milvus覆盖云产品文档、长文本及模糊语义查询，强化咨询类语义召回效果；
 - 知识图谱融合检索：构建Neo4j云产品知识图谱，落地向量+图谱双路检索，复杂场景召回准确率由60%提升至92%。
 4. MCP工具化封装与业务闭环建设，解决工具能力耦合、跨Agent复用难、迭代周期长问题：
 - 标准化工具封装：基于FastMCP封装MySQL查询、云资源API、监控指标接口为MCP Server，实现能力解耦与跨Agent复用；
 - 研发效能提优：工具开发、注册、联调周期由天级压缩至小时级，整体研发效率提升60%。
 5. 长短期记忆系统建设，解决多轮对话Token冗余、用户重复陈述背景问题：
 - 短期记忆优化：基于Redis维护会话级记忆，窗口压缩多轮上下文，降低对话Token消耗；
 - 长期记忆沉淀：基于Milvus存储用户长期偏好，自动注入System Prompt、异步更新用户偏好，重复陈述轮次减少45%，用户满意度提升18%。
 6. 评测与可观测性闭环建设，解决模型迭代无量化标准、线上问题难归因问题：
 - 标准化离线评测：基于Ragas搭建多维度评测体系，核心变更需回归验证后方可上线，保障迭代质量；
 - 全链路可观测：依托Langfuse Trace复盘Bad Case，精准定位Agent、缓存、工具调用链路问题，支撑持续迭代优化。

教育经历

河海大学(保研)- 计算机类 硕士 计算机与软件学院 全日制

2024年09月 - 2027年07月

211 双一流

- 荣誉奖项：学业一等奖学金、CPEC2025论文一等奖、发表学术论文1篇、获得发明专利2项
- 校园经历：担任研究生生活委员兼心理委员

湖南中医药大学-信息管理与信息系统 本科 信息科学与工程学院 全日制

2020年09月 - 2024年07月

- 荣誉奖项：国家励志奖学金、校级二等奖学金、中国大学生服务外包大赛全国三等奖、中国大学生计算机设计大赛全国三等奖、湖南省数学竞赛省赛三等奖、湖南省大学生计算机程序设计竞赛省赛三等奖、软件著作权6项
- 校园经历：担任本科班长