



教育背景

中南大学 (985, 推免) | 人工智能 硕士 | GPA: 3.77/4.0 | 2024.09–2027.06

竞赛荣誉：一等奖学金；华为杯研究生数学建模竞赛二等奖、三等奖；亚太杯数学建模竞赛一等奖

武汉理工大学 (211) | 自动化 本科 | GPA: 4.18/5.0 | 2020.09–2024.06

竞赛荣誉：西门子中国智能制造挑战赛国家特等奖；计算机设计大赛省二等奖；二等奖学金；校三好学生；优秀毕业生；CET6

实习经历

蚂蚁集团 | 信贷事业群 – 花呗贷前团队 | 智能体与大模型应用工程实习生 | 2026.05 – 至今

1. 花呗用户正负向冲突诊断智能体

项目介绍：面向花呗用户在签约/提额/解冻后短期遭遇冻结、降额、清退等正负向冲突导致优质用户流失的场景，基于Claude Code构建冲突诊断Agent，辅助产品/风控/消金团队定位异常链路。Agent上线后发现**35W冲突用户**并推动闭环复核，守护签约账户底盘的稳定。

- **Skill分层与 workflow编排：**设计主skill意图路由+子skill原子能力的分层架构，规避单体Skill上下文膨胀导致的任务边界混淆与幻觉，以此打通 数据分析 → 冲突识别 → 用户诊断 → 交互呈现 → 工单处置 → 知识沉淀 的端到端闭环。
- **Skill上下文治理：**通过渐进式披露组织Skill和确定性逻辑脚本化，单次任务上下文从100k+ **压缩至数十k级**，降低输出幻觉。
- **推理与交互延迟优化：**设计三张ODPS中间表承接预算事实，典型诊断链路**压缩至分钟级**；设计**Hook机制**在LLM前置拦截交互请求，将交互渲染响应从5-6秒**降至毫秒级**，同时避免百KB级AixUI卡片输出污染200k上下文，增强Agent运行稳定性。
- **知识闭环与自进化：**设计工单状态管理体系，构建“反馈事实沉淀-知识提炼-跨轮次/跨会话复用”的**知识闭环机制**，Skill保持执行逻辑稳定，仅推动知识层随反馈持续增量演进，实现执行器与知识迭代解耦；最终沉淀为**可复用、可迁移的Agent能力资产**。

2. 花呗签约域知识库建设

- **知识更新策略：**针对知识问答场景下只消费、无反馈导致的知识缺失、知识冲突及知识过期等问题，主导设计由真实问答反馈驱动的知识更新策略，推动知识库在实际使用中持续纠错与演进，并确保知识变更可追溯、更新效果可验证。
- **闭环机制设计：**搭建 会话审核-知识提炼-人工审核-写入评测-知识生效 的闭环，设计hook自动捕获来源会话，由Sub-Agent隔离执行写入前后对照评测，仅生效验证通过的知识；将流程抽象为可配置、可复用机制，已在团队内两个业务域知识库上完成复用验证。
- **检索策略优化：**设计融合知识图谱、全文检索与语义匹配的分层检索策略以及动态排序规则，提升知识问答的召回完整性与答案可信度。

中国铁建重工集团公司 | 电气设计研究院-AI 所 | 大模型应用开发实习生 | 2026.02 - 2026.05

1. 智能问数系统 (Text-to-SQL & ChatBI 数据中台)

项目介绍：面向企业内部MySQL多数据库查询门槛高、跨业务取数效率低的问题，构建基于LLM的自然语言问数与可视化分析系统。

- **Workflow工作流：**将查询任务拆解为“Schema检索 → 意图解析 → SQL生成 → 校验执行 → 错误修正”的工作流执行链路；结合M_schema动态转换、CoT Prompt及执行反馈驱动的**SQL自修正机制**，使端到端查询执行**准确率提升至92%**，较基线**提高15%**。
- **Schema检索架构：**基于RAG实现语义检索、BM25关键词匹配、库内值召回的并行混合检索方案；离线索引构建Table/Column检索文档与元数据存入**Chroma**向量数据库；在220条真实业务查询评测中实现列检索Top-3命中率100%。
- **检索链路优化：**针对宽表列数多、全量列向量检索耗时高的问题，设计**表级粗召回+列级精召回的分层检索机制**，检索时延降至毫秒级。
- **ChatBI：**基于FastAPI搭建后端服务接入本地LLM，通过**SSE**协议流式返回工作流状态与查询结果，并集成ECharts实现数据可视化。

项目经历

1. 面向水利领域的 Text-to-SQL 模型微调

- 面向水利领域自然语言问数场景，基于**Qwen2.5-Coder-7B**开展Text-to-SQL模型微调，完成自然语言问题到SQL语句的结构化转换。
- 基于真实数据库PostgreSQL的Schema信息，采用规则模板映射、大模型扩写、人工校对的方式构建**14000条**Question-SQL样本对。
- 基于LlaMA-Factory完成LoRA微调，并在训练输入中显式注入相关Schema信息，测试集执行准确率由基座模型的**65%**提升至**85%**。

2. 基于大语言模型的短期电力负荷预测方法研究

Large Language Models for Short-Term Electricity Demand Forecasting under Abrupt Changes

一作 | JCR Q1 | 在投

- 面向重大节假日、极端天气等引发的负荷突变与非平稳下的浙江省真实电网预测场景，基于Qwen3开展**短期电力负荷预测**研究。
- 设计**外生变量融合与突变感知增强框架**，全局建模天气、日期等外生信号与负荷趋势的耦合关系，局部结合Prophet与稀疏交叉注意力强化突变时段关键模式；通过多头交叉注意力实现数值时序特征与LLM表征空间的**跨模态对齐**；结合领域提示增强预测任务理解。
- 在12个真实电网数据集上与18个基线模型对比，整体效果最优；预测精度较最优对比基线**提升3%**，较已有LLM时序预测方法**提升5%**。

3. 自主式 AI 编程助手 (Coding Agent)

- 独立设计实现轻量级**Coding Agent**，基于**ReAct**架构构建交互-引擎-工具-记忆-安全**五层**模块，支持多模型接入及自主编程任务执行。
 - 实现**跨会话记忆与安全执行体系**，支持JSONL持久化、LLM自动记忆提取和项目指令分级加载；通过分层权限确认将单次会话弹窗由平均 30 次降至 5 次以内，并支持OS级沙箱隔离。
 - 设计**多 Agent 并行协作与文件级隔离**机制，突破单 Agent 上下文窗口瓶颈，跨文件重构任务相比单 Agent 耗时降低约 60%。
-

专业技能

- **Agent 工程**：具备生产级 Agent 设计与落地经验，掌握 Workflow、Tool Calling、Memory、Skills、Hooks、MCP 及 Multi-Agent 协作；具备 Claude Code、Codex、OpenClaw 等主流 Agent 的使用与实战经验，熟悉上下文治理、工具编排等。
- **RAG 与知识工程**：掌握 RAG 的混合检索方案，具备离线索引构建、召回优化、问答链路搭建以及反馈驱动的知识更新闭环实践。
- **技术栈**：Python | FastAPI | LangChain | LangGraph | Chroma，具备 Qwen 等 LLM 部署调用与微调经验，了解 MLLM/RLHF 原理。