

陈正园

AI大模型应用开发

专业技能

Agent开发:

- Agent框架**: 熟悉 LangChain、LangGraph、DeepAgents框架, 掌握非线性 workflow 编排, 具备问答、生成及对话系统开发能力
- CodingAgent**: 熟悉 CodingAgent 中 AgentLoop、上下文压缩、Skill、MCP、工具调用及多 Agent 等模块的核心机制与设计实现
- LLM**: 了解 GPT、Llama、Qwen 等主流 LLM 架构, 掌握基于 PyTorch 进行 SFT、RLHF/DPO/GRPO 等训练范式及推理优化技术
- AI工具**: 熟练使用 ClaudeCode/Trae/ZCode 等 AI 编程工具进行 VibeCoding 开发, 掌握 Spec-Kit、Superpowers 规范开发使用

基础技能:

- 编程语言**: 熟悉 Java、Python 等多种语言开发, 掌握面向对象编程思想、熟悉常用数据结构、集合框架及多线程并发编程
- Web框架**: 熟悉 SpringBoot、SpringCloud、FastAPI 框架, 理解 IoC、AOP、依赖项, 掌握常用注解及其底层实现机制
- 数据库**: 熟悉 MySQL、Redis、Milvus、Elasticsearch、MongoDB 等数据库的调用与优化, 具备索引设计及性能调优经验
- 操作系统/容器**: 熟悉 Linux 系统及常用 Shell 命令, 掌握 Docker、K8s 编排容器, 能够在云上搭建环境部署推理模型或应用

工作经验

北京幂律科技有限公司 AI应用开发-全栈 2025/01 - 2026/06

- 主导教育问答 RAG 核心模块的工程化落地, 全面涵盖知识库向量索引、检索生成、上下文管理及提示词工程设计
- 参与 Agent 系统架构设计, 构建电商问数及经营分析专家 Agent 并推动落地, 独立完成核心模块功能开发工作
- 主导 PodaAgent 架构设计与全部核心技术决策, 借助 ClaudeCode 基于规范驱动开发完成全部模块编码与测试

广东亿迅科技有限公司武汉分公司 后端Java开发工程师 2021/07 - 2024/12

- 参与计费列账系统开发, 汇聚生产数据并依规则执行列账处理, 实现自动化财务核算, 输出财务报表与可视化展示
- 参与一体化研发管理中台开发, 纵向构建多层级项目管控体系, 横向覆盖研发生命周期, 实现需求、任务与缺陷集中闭环管理

项目经历

知行 AI RAG 系统 AI应用开发 2026/01 - 2026/04

项目背景:

面向教育培训场景的智能问答系统, 基于 LangGraph 编排内容导入与检索问答双 workflow, 融合 Milvus 混合检索、BGE-M3 向量化、RRF 多路召回与重排序、意图分类路由及 qwen 大模型, 实现课程、文档、题库的高精度语义检索问答

技术栈: Python / FastAPI / LangGraph / Milvus / BGE-M3 / MongoDB / MinIO / Docker / MinerU

个人职责:

- 基于 LangGraph 设计导入与查询两条独立状态图 workflow, 导入侧按数据类型动态路由处理分支, 查询侧按请求意图分流检索路径。通过 TypedDict 状态对象统一流转数据, 配合节点级进度追踪与单点异常隔离, 实现流程全链路可观测与容错
- 自研分层分块策略**: 先按结构标题正则初切保留语义边界, 超长章节用递归字符分割器二次切分, 同源短块智能合并避免碎片化。入库以业务主键幂等删旧, 并自研字符串转义防护 filter 注入, 保障重复导入与多集合写入稳定
- 查询重写**: 结合最近多轮对话历史实现请求意图识别与指代消解, 输出结构化意图标签与改写查询。意图命中专用场景时短路直出, 跳过冗余检索与重排, 降低无效计算开销
- 三路混合检索**: 基于 Milvus 构建稠密与稀疏双索引, 一次 BGE-M3 向量化同步产出双向量, 经加权排序器融合 (权重 0.8:0.2)。叠加 HyDE 假设文档检索、MCP 网络检索作为独立召回路, 缓解短查询与长文档的语义鸿沟
- RRF与重排序精排**: 自实现加权 RRF 按文档 ID 去重融合多路召回, 取 Top10 进入 BGE-Reranker 句对精排。重排后采用动态断崖截断 (绝对/相对双阈值) 自适应确定返回数, 避免机械取固定 N 导致噪声混入

项目成果:

- 知识库召回 Recall@10 从 71% 提升至 93%, 问答准确率提升 22 个百分点
- 场景短路路由使 35% 简单问答跳过检索流程, 整体检索算力开销降低 30%

电商问数经营分析 AI Agent 系统 AI应用开发 2025/06 - 2026/10

项目背景:

面向电商场景的对话式经营分析平台, 采用微服务架构, 包含认证中心、NL2SQL 数据 Agent、对话式分析 Agent。用户通过自然语言对话即可完成销售归因、经营诊断与根因定位, 系统自动编排工具链生成可视化分析报告

技术栈: Python / FastAPI / LangChain / LangGraph / DeepAgents / MySQL / Qdrant / Elasticsearch / BGE-M3 / Vue.js

个人职责:

- 星型电商数仓&数据同步**: 构建知识库离线自动化构建脚本, 将元数据同步至 MySQL + Qdrant + ES 三端, 实现字段级语义索引与取值全文索引

- 2. **NL2SQL 全链路 Data Agent**: 融合分词、向量、全文多路召回,LLM 过滤降噪生成 SQL; 内置 EXPLAIN 执行校验 + 自动纠错闭环, 拦截非法查询、语法错误
- 3. **多会话隔离**: 基于 **DeepAgents** 混合路由设计动态工作区, 通过 RunnableConfig 注入独立工作目录, 文件、缓存、会话数据完全隔离, 并发查询互不干扰
- 4. **兜底降级策略**: 简单统计查询使用低成本轻量模型, 复杂经营归因切换高阶大模型, 在保障问题解决率前提下显著削减 LLM API 调用成本
- 5. **会话自动压缩&断点恢复**: 长对话自动摘要入库, 统一压缩边界偏移换算, 任务异常中断可从历史摘要精准重试, 无需重复提问

项目成果:

- 1. 业务人员自助数据查询覆盖率提升 90%, 数据分析重复取数工作量减少 70%
- 2. 分级模型调度使整体 LLM 调用成本下降 42%, 复杂问题解决率维持 98% 以上
- 3. NL2SQL 语法错误、非法查询拦截率 100%, 多会话并发无数据冲突、文件覆盖问题

PodetCode终端 AI 编程助手 - 【个人项目】 AI应用开发 2026/05 - 2026/07

技术栈: Python / MCP / ReAct / Skill / Multi-Agent / JSONL持久化 / OS Sandbox

项目介绍:
PodetCode 基于 Python 实现的终端 AI 编程助手, 支持多协议 LLM、MCP 工具、子代理/团队/worktree 隔离、双层上下文压缩与权限规则引擎, 提供交互式 TUI、非交互式 CLI 与远程浏览器三种形态

- 个人职责:**
- 1. **MCP 延迟加载机制**: 设计并实现 MCP 工具延迟加载机制, 百级工具场景下工具描述 Token 占用减少 85%, 避免上下文窗口被工具定义挤占
 - 2. **双层渐进式上下文压缩**: 设计并实现"工具结果预算 + LLM 摘要"压缩策略, 支持 8 小时以上连续编程会话不丢失关键上下文
 - 3. **五层权限确认机制+OS内核沙箱兜底**: 将单次会话权限弹窗从平均 20 次降至 5 次以内, 避免审批疲劳影响使用体验; 可选启用 OS 级沙箱进一步兜底, 内核级隔离支持 Agent 全自动模式安全执行
 - 4. **多Agent协作**: 实现多 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 65%
 - 5. **实现记忆的分层设计与实现**: 基于 JSONL 实现会话的持久化与会话的恢复, 提取 4 类记忆的分层实现与自动提取以及长期记忆的自动化治理

收入回现稽核系统 Java后端开发 2023/01 - 2023/12

项目描述:
为中国电信国际开发的电信收入对账与审计平台, 实现BSS(业务支撑系统)与SAP(财务系统)之间的收入、成本、支付数据的自动化对账、结算列账与合规稽核。覆盖收入确认、SAP列账、MVNO稽核、多维财务报表等核心业务, 支撑CTG全球业务(含香港、巴西等国际市场)的财务数据闭环。

技术栈: SpringBoot / Dubbo / MyBatis-Plus / Vue.js / Element UI / MySQL / Redis / ZooKeeper / Apache POI / Quartz / FTP / MinIO

- 个人职责:**
- 1. 构建**分布式定时任务调度框架**: 基于Spring TaskScheduler +ZooKeeper实现动态Cron调度系统, 支持任务链顺序执行、前驱任务依赖等待、断点续跑、MD5执行去重
 - 2. 基于**Oauth2授权码模式认证**结合SpringSecurt组件实现包括用户账号密码登录,手机号验证码登录,第三方平台登录及接入子系统的单点登录
 - 3. 实现了**基于RBAC模式的权限校验系统**.将校验权限的颗粒度细化到菜单、按钮、URL,能够在页面上实现对角色权限的配置.后端基于SpringSecurity中的注解快捷、灵活校验当前用户是否有权限