

# 成炫宏

大模型应用开发

27届大三

2877228695@qq.com

广州

2004/09

17389894637

## 教育背景

沈阳理工大学

计算机科学与技术 · 本科

2023/09 - 2027/07

- 主修课程：数据结构、软件工程、操作系统、计算机网络、Web开发、机器学习、深度学习
- 专业排名前 5%，获奖情况：2025年全国大学生智能技术应用大赛大模型应用赛道国家一等奖、2025年全国大学生数学建模竞赛辽宁赛区省级一等奖、2025年辽宁省大学生微信小程序应用开发大赛省级一等奖等共计十余项

## 项目经历

智能旅游规划 Agent 助手

核心开发者

2025/12 - 2026/02

### 项目背景

传统旅行规划方式效率低、体验差，用户需要在多个平台查询景点、攻略、酒店、交通。初版Agent系统采用串行调度和关键词匹配，意图识别准确率仅65%，响应时间达30秒，无法支持用户偏好记忆和企业知识查询。

### 解决方案

- 两层记忆系统**：设计短期记忆+长期记忆+异步LLM总结机制，实现PreferenceAgent智能识别追加/覆盖偏好动作；引入Redis缓存层（用户偏好热数据、LLM总结结果），缓存命中率85%；
- RAG知识库**：基于Milvus向量数据库+BGE-m3 Embedding模型+滑动窗口Chunking策略，构建差旅政策、城市攻略、FAQ知识库，采用混合检索+文档溯源保证可信度；集成Google Search API实现实时信息查询（天气、限行等）；
- 多Agent协作优化**：重构IntentionAgent为基于LLM语义理解的多意图识别系统（支持6大类意图），实现“优先级+并行”混合调度模式，响应时间从30秒降至15秒；
- Skill Plugins插件化架构**：实现LazyAgentRegistry动态发现机制+懒加载（系统启动时仅扫描不加载，首次调用时按需加载并缓存），支持灵活增减Agent无需修改核心代码；采用Progressive Disclosure渐进式暴露机制（意图识别阶段仅加载Skill元数据，执行阶段按需加载详细指令），系统启动速度优化至3秒。

### 项目成果

- 核心指标提升：意图识别准确率从65%提升至90%+，知识库问答准确率95%，用户偏好记忆准确率95%，系统响应时间（ttft）从30秒降至15秒；

医疗助手Agent系统

核心开发者

2026/04 - 至今

### 项目背景

传统单 Agent 医疗问答系统存在响应效率低、复杂问题分析不全面、多轮对话上下文理解能力弱等问题。面对多症状、多维度的医疗咨询时，回答准确性与稳定性不足；同时缺乏长效记忆机制，无法连贯理解用户意图，难以满足实际医疗场景下高效、可靠、安全的智能问答需求。

### 解决方案

- Skills-Agent两层架构**：底层构建7个原子Skills（知识检索、风险评估、症状分析、生活方式建议、ICD-10编码、临床指南、深度研究）供Agent使用，上层3个专业Agent（健康咨询、症状诊断、医学研究）根据任务需求自主选择调用，Skills和Agent完全解耦；
- Agent Loop执行机制**：基于ReAct框架实现Think-Act-Observe循环，LLM自主决策工具调用并观察结果，设置硬性限制最多调用2次工具防止过度调用，单Agent平均响应时间控制在15秒以内；
- Agent Swarm协作系统**：智能路由判断问题复杂度（70%单Agent快速通道/30%Swarm协作模式），复杂问题自动分解任务并分配给多个Agent并行执行，Agent间通过共享上下文间接共享任务状态和执行结果；
- 双层记忆机制**：短期记忆管理会话级对话历史（单例模式确保所有Agent共享会话上下文，集成熵管理器自动去重压缩消息35%，加载最近5轮对话），长期记忆通过Mem0云服务向量化存储会话总结支持跨会话相似案例检索；
- Harness Engineering约束系统**：通过YAML文件显式定义Agent能力边界并运行时自动验证，实现输出自动修复机制（缺少免责声明自动添加、检测到高危症状自动添加就医警告）；

### 项目成果

- 系统性能**：智能路由准确率从88%提升至95%，知识库检索准确率87%，单Agent响应时间5-15秒，Swarm响应时间20-30秒，多轮对话上下文理解准确率从60%提升至92%；
- 回答质量**：医学专业人员三维度盲评打分，系统回答质量4.5分 vs doubao-seed-2.0-pro 3.9分；

## 专业技能

- 编程**：Python（熟练）、Java（掌握）；熟练使用 Cursor、Claude Code 等 AI 编程工具；
- Agent/RAG**：精通 LangGraph/LangChain 多 Agent 编排；掌握意图路由、长短记忆、工具调用、混合检索与重排；
- 工程部署**：熟练 Docker 容器化；精通 MySQL 设计与优化；熟练 Redis 缓存、会话存储与高性能应用；
- 前沿技术**：了解 LoRA/Q-LoRA 微调；关注 Context Engineering、Skills、Harness Engineering 等方向。