

冯嘉俊

电话: 19849869176

政治面貌: 共青团员

现所在地: 广州

邮箱: 19849869176@163.com



求职意向

意向岗位: 大模型应用开发/AI 全栈开发

意向城市: 广州

当前状态: 随时到岗

期望薪资: 8k

教育经历

2022.9-2026.7

广东工业大学

计算机科学与技术 | 本科

- 核心课程覆盖人工智能原理、分布式系统、计算机网络、数据结构与算法等, 扎实掌握 AI 应用开发底层理论基础。
- 曾获蓝桥杯 C++ B 组省级赛三等奖, 具备扎实的编程能力与问题解决能力。

实习经历

2025.8-2026.3

广州智创未来科技有限公司

AI 开发实习生

- 参与公司多模态内容生成平台的开发工作, 负责 RAG 检索模块的功能迭代与性能优化, 基于 LangChain 实现了多源文档的结构化解析与向量入库。
- 协助搭建大模型应用测试框架, 完成 12 类业务场景的 prompt 调优, 将模型输出准确率从 82% 提升至 93%, 满足业务上线标准。
- 参与基于 FastAPI 的后端接口开发, 完成 3 个核心接口的设计与编码, 配合前端团队完成联调, 接口平均响应时间控制在 200ms 以内。
- 编写项目技术文档与操作手册, 累计输出文档超过 1.5 万字, 为后续团队维护与功能迭代提供了清晰的参考依据。

项目经历

2026.4-2026.9

FinAnalyst-RAG | 金融年报 RAG 知识中台

核心开发者/全栈开发

项目描述: 面向金融年报阅读与财务分析, 构建覆盖文档解析、混合检索、工具计算、问答生成及质量评测的完整流程, 支持财务指标查询与原文证据追溯; 配套开发知识资产库、检索沙盘和运行监控工作台, 实现从财报入库到回答核验的可视化操作。

- 财报解析与结构化入库:** 设计“书签目录→正文标题扫描→分页兜底”的章节定位策略, 基于文字坐标聚类重建跨页财务报表, 结合结构、双期会计关系和披露数据交叉校验; 在 3 份年报、822 页 PDF 上形成 884 个切片、707 条财务事实记录, 保留单位、期间及原页来源。
- 混合检索与证据追溯:** 实现关键词与向量双路召回、RRF 融合和模型重排, 结合公司、年份、文档范围及上下文预算筛选证据; 支持服务异常时显式降级, 通过检索沙盘展示四阶段候选, 并定位 PDF 原页。
- Agent 编排与财务计算:** 自研异步状态图, 编排路由、检索、工具计算、生成、校验 5 阶段流程, 使用 Decimal 支持金额换算与 7 类财务比率; 设置默认 90 秒执行预算、最多 3 次生成模型请求, 对证据不足和执行异常返回明确状态。
- 权限控制与交互隔离:** 基于认证身份构建 Actor 上下文, 通过角色、租户、用户归属及三级 ACL 控制数据访问, 在

召回与计算前校验权限；使用 SSE 推送进度与结果，角色切换时取消在途请求并清理敏感状态。

- **可观测性与质量验证：**记录 Trace、节点耗时、Token 用量、模型成本及 P50/P95/P99 延迟，集成 Ragas 忠实度、答案相关度评分与 Recall@K 检索诊断；针对来源绑定、科目冲突、期间识别及权限问题完成 58 项后端回归测试及 8 个年报离线验收场景。

2026.9-2026.10

ClipRAG-Omni——多模态素材检索与成片 AI 中台

核心开发者

- **多模态视频解析与切片资产管理：**基于 FFmpeg 设定 0.4 灵敏度将视频切分为 1.5 到 6 秒独立镜头，调用 Qwen-VL 抽取场景与动作语义标签；自研基于 Float32Array 的 1024 维向量点积加速引擎，将检索延迟压缩到 20 毫秒以内，台词匹配准确率达 92%以上，支持成片流式点播。
- **大模型分镜策划与双智能体质量审核：**基于 Qwen-Plus 与 JSON Schema 约束自适应生成 6 类营销脚本，内置 40 多个行业词库实现关键词自动替换；构建创作与质检双智能体闭环，设定 85 分质量评分阈值控制内容反写，落地两阶段卡点内容审核机制，支持渲染前在线微调台词，降低无效转码算力消耗 60%以上。
- **底层扩展渲染管线与平台去重工程：**基于 FFmpeg FilterGraph 设计双路渲染管线，适配背景高斯模糊铺底逻辑消除黑边；运用 sidechaincompress 滤镜实现口播时背景音乐 15dB 动态闪避与 200 毫秒平滑回弹；注入 0.95 到 1.05 倍动态微变速与感知哈希扰动，单段 1080P 渲染仅需 8 到 15 秒，平台成片通过率提升至 95%以上。

相关技能

- **智能体架构与工作流编排：**熟悉 LangGraph 与 LangChain 状态图编排，熟悉 ReAct 范式与多智能体协同；熟练运用 Tool Calling 挂载外部工具，能够结合 Dify、Coze 与 LangSmith 高效搭建具备全链路追踪高确定性业务智能体。
- **多模态感知与高阶 RAG 工程：**熟练运用 Qwen-VL 进行画面语义解析与特征打标；深入掌握基于 pgvector 的向量库构建、多路混合检索与重排序流程，具备高维特征向量相似度计算优化经验，可有效抑制模型幻觉。
- **企业级前后端与高并发开发：**具备 Java 与 Python 双栈基础，深入掌握 Spring Boot 3、FastAPI、MySQL 与 Redis 高性能服务开发；熟练结合 WebSocket 与 SSE 实现即时通讯与模型流式响应，依托 Pydantic 实现严格结构化输出。
- **云原生运维与端到端商业交付：**深入掌握 Linux 运维、Docker 容器编排、原生 FFmpeg 音视频编解码与 Nginx 网关流式调优；熟练接入 MCP 协议，具备从需求拆解、全栈架构搭建到快速上线实现商业变现的完整闭环能力。

荣誉证书

蓝桥杯 C++ B 组省级赛三等奖（2024 年）、大学英语六级（567 分）

自我评价

- 在校期间深耕 AI 应用开发领域，拥有 3 个从 0 到 1 的全栈项目落地经验，覆盖多模态 RAG、智能体开发、高并发系统搭建等核心场景，具备扎实的技术功底与独立解决复杂问题的能力。
- 熟悉大模型应用全链路开发流程，能够独立完成需求分析、架构设计、编码实现、上线运维全流程工作，所开发项目均已上线并具备稳定的服务能力。
- 对 AI 技术落地保持高度热情，持续跟进大模型应用领域前沿技术，具备良好的学习能力与技术迁移能力，能够快速适配不同业务场景的开发需求。
- 性格严谨并善于社交合作，注重代码质量与系统稳定性，具备良好团队协作意识与沟通能力，能够快速融入团队并承担核心开发任务。