MRAG-知识库问答系统

从“能问答”到“懂图文”的企业级知识中枢

哈库纳玛塔塔

2025.08

企业知识管理的现状与挑战

01

知识管理现状

企业 80% 知识以 PDF 图文混排形式存在 → 传统纯文本 RAG 漏掉 30% 关键信息

02

多模态大模型痛点

多模态大模型在线理解 + 在线检索 → 成本高、延迟高、上下文限制

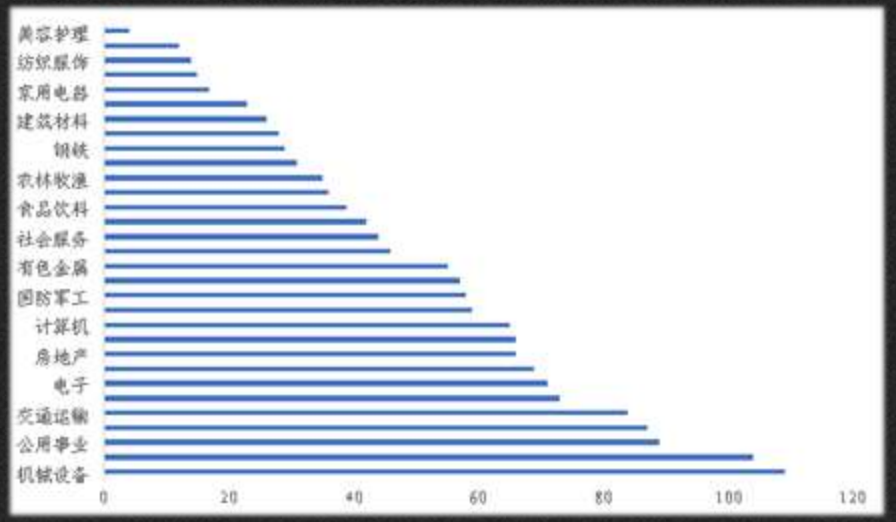
03

项目带来的机遇

真正落地的企业级图文混合 RAG 闭环

”

示例：央国企（A股）上市公司市值战略研究报告片段



切片内容：

图1 2024年央国企（A股）上市公司行业分布美容护理纺织服饰家用电器建筑材料钢铁农林牧渔食品饮料社会服务有色金属国防军工计算机房地产电子交通运输公用事业机械设备0 20 40 60 80 100120

存在问题：无法正确解析图表数据，导致问答时漏掉关键趋势分析

系统架构图详解

分层架构

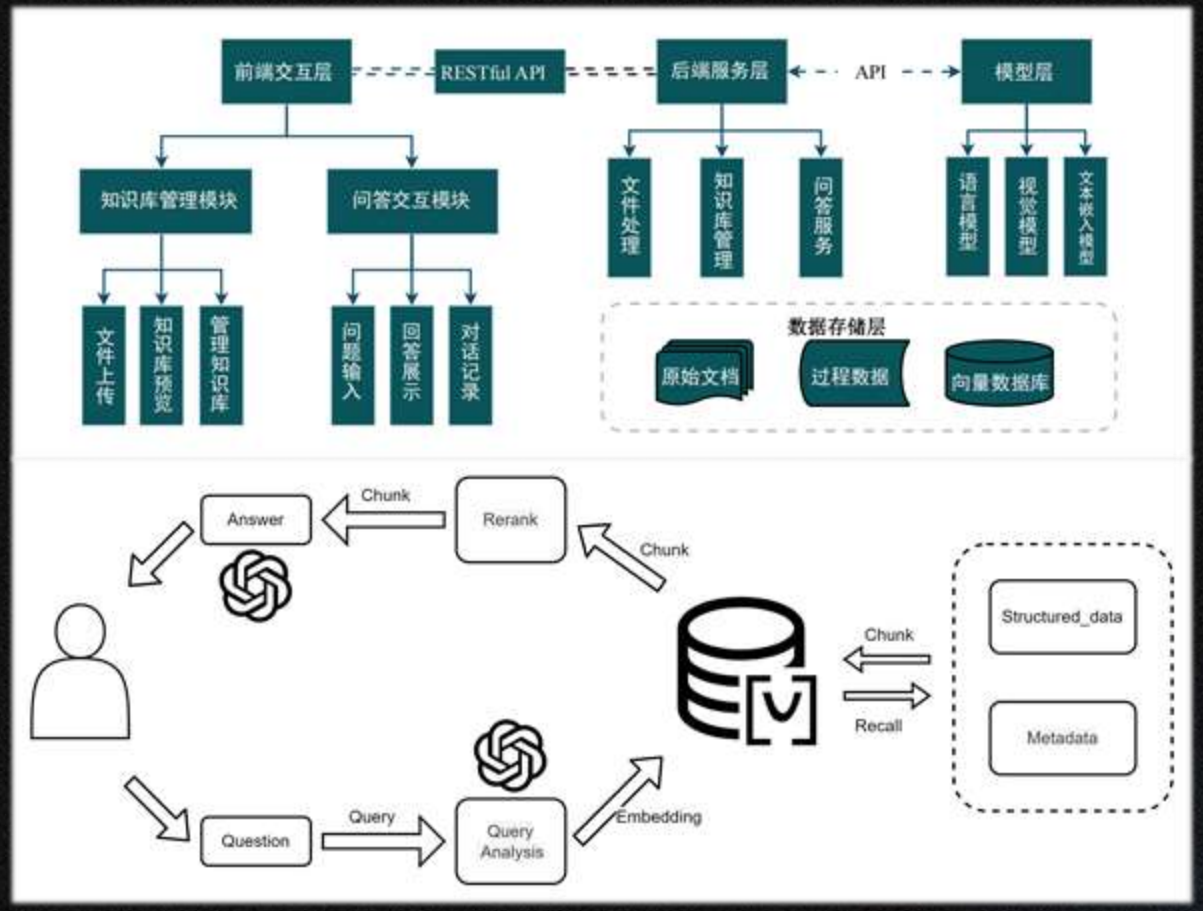
前端交互层、后端服务层、模型层和数据层，实现模块解耦。

模块划分

文件处理、模型、数据库和应用服务模块，明确功能定位。

数据流程

知识库构建和问答交互流程，保障数据处理和问答服务。



创新点1： 图文分离存储 & 并行流水线

利用地址解耦

文本地址：.../texts/{doc_id}/chunk_{n}.txt

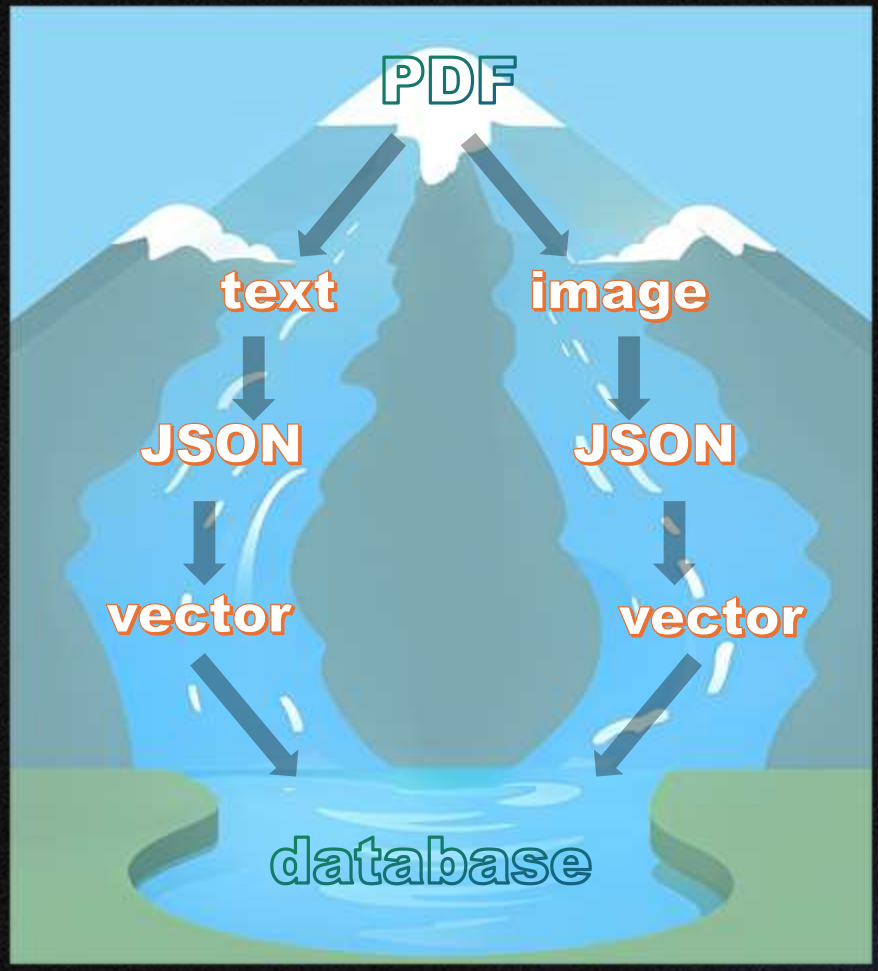
图像地址：.../images/{doc_id}/img_{m}.png

优势

- 文本分块与图像描述两个环节可水平扩展，互不阻塞
- 支持增量更新：仅追加新图片即可

工程亮点

- 自动检测重复文件，节省 18% 存储
- 统一 trace-id，日志可追踪到单张图/段文本”



创新点2：结构化知识片段

结构化JSON对象 01

将文本分块和图像的描述性文本重构为结构化JSON对象，包含以下信息：主题、关键词、摘要、主要对象等。

优势 02

-支持下游检索：可直接命中 topic / keyword实现匹配。

-搭配查询重构功能，提高召回成功率。

Query：

original_text: 「原始问题」
enhanced_query: 「增强表述」
important_terms: 「重要术语列表(3-5个)」
intent_category: 「问题意图分类(如"信息查询"、"概念解释"、"对比分析"等)」
core_information_need: 「用户核心信息需求」

Retrieval

ID：868b77fd-2522-47c7-bc6f-7a198b906a1b
7a198b906a1b
topic: 医院急救训练
keywords: 医护人员，医院环境，救护模拟器，培课程，急救技能
summary: 两名穿着蓝色护士服的人在进行紧急救救护培训，其中一人正在操作救护车内的氧气面罩装置，并与受训者的模特互动练习。背景中有悬挂输液袋和其他医疗用品。
objects: 医疗设备，护士制服，受训者模型
text: GRAYS Pocket NURSE

创新点3：统一路由&实时分流

01

统一路由&实时分流机制

系统能够根据用户上传文件的类型、大小、复杂程度是否需要工具调用执行增强上传处理：

简单上传：用于单一数据类型文件的上传；

增强上传：用于复杂文件处理任务，工具包含文件解析、数据分块、图片分析理解等。

02

工程优化亮点

将工具调用由人工配置变成了后台毫秒级就能完成的自动决策，实现一次性工程封装。既让用户无感，又能实现省钱、可扩展、可维护。

数据血缘与质量治理

01

数据转化流程

从PDF文档开始，通过chunk_id和img_id将数据转化为JSON格式，并进一步生成embedding_id，实现了数据的标准化和结构化处理。

无效数据检测与处理

02

建立了脏数据自动检测机制，在语料进入索引之前对“脏数据 / 无效数据”进行清洗；在检索阶段，对“看似相关、实则无用”的文档以及段落进行过滤。

03

用户反馈闭环

系统前端设“👍”&“👎”按钮，将用户反馈的低分案例自动加入 BadCase 池，定向地周期性迭代检索评估系统。

可拓展性设计

01

离线化本地部署

系统选用的Qwen系列语言、视觉以及嵌入模型，均可将线上模型接口替换为同系列开源模型，实现“零外网依赖、零调用费用”。只需一台 GPU 服务器即可在本地完成检索、生成全流程，满足数据不出域、低延迟、高并发场景，真正做到“插上电就能跑”。

02

垂直领域Fine-Tuning

用少量高质量“标准问答对”对开源底座进行微调，把模型的输出格式、用词风格、引用方式统一到公司规范。短期内即可把“自由发挥”收敛成“标准答案”，让一线客服、审核、质检都能直接复用，降低后期人工校正 60% 以上。

03

无缝兼容GraphRAG

现有结构化 JSON 天然携带实体-关系信息，可直接映射为知识图谱节点与边，无缝切换到 GraphRAG 范式。无需重构索引，仅需新增一层图查询接口，即可实现多跳推理、关系可视化与可解释溯源，为后续复杂问答、决策支持奠定基础。

用户案例与技术沉淀

01

用户案例

该系统最终版本上线以来，已经被不包括作者在内的二十余位全国各大高校的学生以及研究员投入使用。对采集到的百余份端到端使用案例进行多维度系统性评估，得到综合得分4.265分（0~5）。该系统能够有效避免模型“幻觉”对研究工作的负面影响，显著提升了知识管理质量。

02

技术沉淀

该项目自最初版本上线，至今已完成五次迭代更新。为不同行业提供差异化的的业务嵌入式RAG解决方案，包括但不限于基于workflow开发的企业智能客服类、金融投资分析类、医疗临床决策类、教育个性化辅导类以及法律案例分析类等。