工程与管理科学
Engineering and Management Science
第 8 卷◆第 6 期◆版本 1.0◆2026 年
文章类型：论文|刊号（ISSN）：2705-0637(P) / 2705-0645(O)

# 基于 RAG 的企业级智能知识库系统设计与实现

何子晗
扬州宇安电子科技股份有限公司 技术研发中心 江苏扬州 225000

DOI:10.32629/ems.v8i6.20585

[摘 要] 针对企业技术文档分散、传统关键词检索语义理解能力不足、知识获取效率低的问题，本文设计并实现了一套基于检索增强生成（Retrieval-Augmented Generation，RAG）的企业级智能知识库系统。研究方法上，系统采用数据层、检索层、生成层的三层架构，以本地化部署的 DeepSeek 大语言模型为生成核心，结合 Milvus 向量数据库与 Elasticsearch 关键词检索引擎构建混合检索策略，并引入交叉编码器 Rerank 对候选文档精排。在 2347 份企业技术文档、200 组评估问答对上的对比实验表明，混合检索结合 Rerank 方案的 Recall@10 达到 92.3%，较纯向量检索基线提升 18.7 个百分点，MRR 提升 0.211，端到端平均响应时间控制在 2.1 秒以内。研究结论表明，所提方案可有效提升企业私域知识的检索精度与问答准确性，已在企业内部稳定上线，为技术密集型企业的知识资产数字化管理提供了可行的工程实践范式。

[关键词] 检索增强生成；大语言模型；混合检索；Rerank；知识库

[中图分类号] TP391 [文献标识码] A

## 1 引言

企业技术文档分散存储、关键词检索语义理解能力不足、经验型知识依赖口耳相传，工程技术人员大量工作时间耗费在信息检索上，严重制约研发效率。近年来，以 DeepSeek 为代表的大语言模型（LLM）展现出强大的自然语言理解与生成能力，但存在知识截止、幻觉生成及无法获取私域知识等固有缺陷。检索增强生成（RAG）通过在推理阶段动态检索外部知识库并将相关片段注入提示词，将检索精确性与生成语言能力有机结合，为企业私域知识问答提供了可行路径[1]。

本文主要工作与贡献：（1）设计并实现了一套面向企业技术文档的本地化部署 RAG 知识库系统，保障数据安全；（2）提出了覆盖文档清洗、语义分块、元数据标注的数据规范化处理流程；（3）构建了混合检索结合 Rerank 的两阶段检索框架，并通过对比实验验证其有效性；（4）系统已在企业实际环境中稳定运行，取得显著应用效果。

图1 系统总体架构图

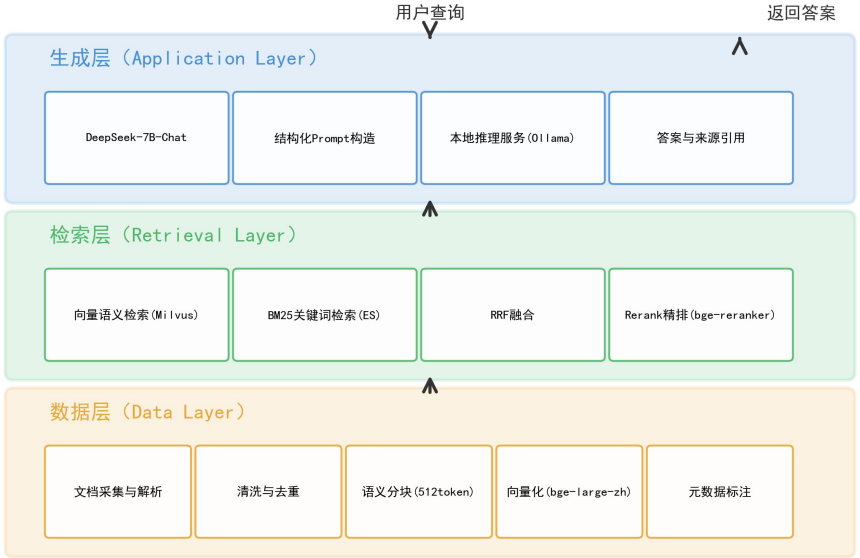


图 1 系统总体架构图

## 2 系统总体架构设计

### 2.1 系统功能需求分析

本系统面向企业技术研发人员，须支持自然语言专业问答（含来源追溯）、对 2000 余份多格式技术文档（PDF/Word/Markdown 等）的语义级全文检索，以及文档持续导入与知识动态积累三类核心需求，并要求对电子信息、嵌入式、显控

等领域专业术语具备较强的理解能力。

2.2 三层系统架构

系统采用三层架构，从底层到顶层依次为：数据层，负责文档采集、解析、清洗、分块与向量化，并分别写入 Milvus 向量库与 Elasticsearch 关键词索引；检索层，并发执行向量语义检索与 BM25 关键词检索两路召回，融合去重后经 Rerank 模型精排，输出高质量上下文；生成层，以 DeepSeek 大语言模型为核心，结合用户问题与检索上下文构造 Prompt，调用本地推理服务生成回答并附带来源引用。系统总体架构及数据流关系如图 1 所示。

2.3 技术选型

系统主要技术选型如表 1 所示。其中 DeepSeek-7B-Chat 经 Ollama 框架部署于企业内网 GPU 服务器，实现完全本地化推理，保障数据不出内网[2]；Milvus 2.3 支持十亿级向量与毫秒级 ANN 检索[4]；Elasticsearch 8.x 配合 ik_max_word 中文分词实现精确关键词匹配；BAAI/bge-large-zh-v1.5 在中文语义相似度基准 CMTEB 上表现优异[3]。

表 1　系统核心技术选型

| 模块 | 选型 | 说明 |
|---|---|---|
| 大语言模型 | DeepSeek-7B-Chat（Ollama 部署） | 本地化推理 |
| 向量数据库 | Milvus 2.3（HNSW 索引） | 毫秒级 ANN 检索 |
| 关键词引擎 | Elasticsearch 8.x（ik_max_word） | BM25 精确匹配 |
| Embedding | BAAI/bge-large-zh-v1.5（1024 维） | 中文语义编码 |
| Rerank | BAAI/bge-reranker-large | 交叉编码精排 |

## 3 知识数据处理与规范化

知识数据质量是 RAG 系统性能的决定性因素之一。原始企业文档格式不统一、内容冗余、术语混乱，若直接入库将严重影响检索精度。

3.1 数据采集与清洗

系统支持主动推送与定时爬取两种接入方式，采集 PDF、Word、Markdown、HTML 等多格式文档。清洗阶段依次完成：使用 PyMuPDF、python-docx 等解析器按格式提取文本并保留章节层级；识别并删除页眉页脚、水印、版权声明等噪声及空白页；统一字符编码为 UTF-8 并归一化全/半角符号；基于 MinHash 局部敏感哈希对相似度超 0.85 的近似重复片段进行合并去除。

3.2 文档分块策略

文档分块（Chunking）质量直接决定检索粒度。系统采用分层分块策略：对具有清晰标题层级的文档优先按章节进行结构化分块，单块约 400~600 字以保持语义完整；对无明显结构的文档采用固定长度滑动窗口分块，块大小 512 token，重叠窗口 128 token，避免边界截断；对超长段落采用递归切分，优先在自然段、句号等语义边界处切分。实验表明，结构化分块结合 128 token 重叠的组合策略，相较纯固定长度分块在 Recall@5 指标上提升约 6.2 个百分点。

3.3 元数据标注规范

每个文档分块入库时附带结构化元数据，以支持过滤检索与结果溯源，字段规范如表 2 所示。检索阶段支持基于元数据的预过滤（Pre-filtering）。

表 2　文档分块元数据字段规范

| 字段名 | 类型 | 说明 |
|---|---|---|
| doc_id / chunk_id | String | 文档及分块唯一标识 |
| source / category | String | 来源系统、文档类别 |
| title / section | String | 文档标题、所属章节 |
| create_time / update_time | DateTime | 创建与更新时间 |
| version / author | String | 版本号、作者 |

3.4 专业术语统一处理

针对"显控"与"显示与控制"、"FPGA"与"现场可编程门阵列"等术语不一致问题，系统建立了约 800 条术语条目的企业术语映射表，在文档清洗与查询预处理阶段同步进行规范化替换；并基于领域语料对 Embedding 模型进行轻量级对比学习微调，提升专业术语的语义表示质量。

表 3　主流中文 Embedding 模型对比

| 模型 | 维度 | Recall@10 | 推理速度（句/秒） |
|---|---|---|---|
| text2vec-base-chinese | 768 | 71.4% | 312 |
| m3e-base | 768 | 74.8% | 285 |
| m3e-large | 1024 | 77.2% | 198 |
| bge-base-zh-v1.5 | 768 | 78.6% | 274 |
| bge-large-zh-v1.5 | 1024 | 83.1% | 176 |

## 4 检索增强策略优化

检索质量是 RAG 系统性能的核心瓶颈。本章从 Embedding 选型、混合检索、Rerank 及召回率优化四个维度展开。

4.1 Embedding 模型选型与对比

本文在企业技术文档评估集上对若干主流中文

Embedding 模型进行了对比评估，结果如表 3 所示。综合检索精度与推理效率，最终选用 bge-large-zh-v1.5 作为生产模型，其在中文语义相似度任务上表现最优[3]。

4.2 混合检索策略

纯向量检索在精确术语匹配（如产品型号、错误代码）上存在弱点，而纯 BM25 缺乏语义泛化能力。混合检索结合两者优势，本系统采用倒数排名融合(Reciprocal Rank Fusion，RRF）算法对两路结果进行合并：

$$\mathrm{RRF\_Score}(d)=\sum_{r\in R}\frac{1}{k+\mathrm{rank}_r(d)}$$

4.3 对比实验

设计四组对比实验，除检索策略外其余配置完全一致：策略 A（基线） 纯向量检索（bge-large-zh-v1.5 + Milvus HNSW）；策略 B 纯 BM25 关键词检索（ES + IK 分词）；策略 C 混合检索（向量+BM25，RRF 融合，无 Rerank）；策略 D（本文方案） 混合检索 + bge-reranker-large 精排。

4.4 结果分析

各策略的评估结果如表 4 所示。

表 4　各检索策略评估结果对比

| 检索策略 | Recall@5 | Recall@10 | MRR | 响应时间(s) |
|---|---|---|---|---|
| 策略 A：纯向量检索 | 67.5% | 73.6% | 0.612 | 0.8 |
| 策略 B：纯关键词检索 | 61.3% | 68.9% | 0.574 | 0.4 |
| 策略 C：混合检索 | 74.8% | 83.7% | 0.701 | 1.0 |
| 策略 D：混合检索+Rerank | 85.4% | 92.3% | 0.823 | 2.1 |

实验结论如下：（1）混合检索的增益显著。相较纯向量检索，混合检索（策略 C）的 Recall@10 提升 10.1 个百分点（73.6%→83.7%），MRR 提升 0.089，验证了向量与 BM25 互补能有效覆盖单一策略的盲区；纯向量在所有指标上均优于纯 BM25，但对精确型号编码（如"BM-2031-A 错误代码"）类查询 BM25 仍占优。（2）Rerank 对排名位置改善突出。引入交叉编码器精排后（策略 D），Recall@10 进一步从 83.7%提升至 92.3%，MRR 由 0.701 跃升至 0.823，说明相关文档被显著前置，用户更易在头部获得有效信息。（3）响应时间在可接受范围。策略 D 端到端 2.1 秒（检索约 0.8s，生成约 1.3s），相较基线增加约 1.3 秒，但满足知识检索场景对实时性的要求。

## 5 结论

本文针对企业技术知识管理需求，设计并实现了一套基于检索增强生成的企业级智能知识库系统。系统以本地化部署的 DeepSeek-7B 为生成核心，构建了数据层、检索层、生成层三层架构，提出了覆盖采集、清洗、分块、向量化的数据标准化流程，并设计了混合检索结合 Rerank 的两阶段检索框架。实验结果表明，本文方案在企业技术文档评估集上达到 Recall@10 为 92.3%、MRR 为 0.823，较纯向量基线分别提升 18.7 个百分点和 0.211，平均响应时间 2.1 秒。系统已在企业内部正式上线运行，覆盖技术研发中心、软件设计中心等多个部门，有效降低了信息检索成本、提升了研发效率与知识复用率。后续工作将聚焦多模态文档处理、对话式追问与领域轻量级微调三个方向。

## [参考文献]

[1]Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks[C]//Advances in Neural Information Processing Systems, 2020, 33: 9459-9474.

[2]Gao Y, Xiong Y, Gao X, et al. Retrieval-augmented generation for large language models: A survey [EB/OL]. (2023-12-18)[2025-04-10]. https://arxiv.org/abs/2312.10997.

[3]Xiao S, Liu Z, Zhang P, et al. C-Pack: Packaged resources to advance general Chinese embedding[EB/OL]. (2023-09-14)[2025-04-10]. https://arxiv.org/abs/2309.07597.

[4]Wang J, Yi X, Guo R, et al. Milvus: A purpose-built vector data management system[C]// Proceedings of the 2021 International Conference on Management of Data (SIGMOD), 2021: 2614-2627.

[5]Cormack G V, Clarke C L A, Buettcher S. Reciprocal rank fusion outperforms condorcet and individual rank learning methods[C]//Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2009: 758-759.

作者简介：何子晗（1998—），男，江苏扬州人，本科，工程师，研究方向为软件工程与人工智能应用。