支持多源异构文档即插即用的医学知识库系统设计

王鹏[1]，叶均明[2,*]，陈宇聪[1]，韩春春[1]，王耿彬[1]，戎伟鑫[1]
（1 大数据与人工智能中心，网络信息科，江门市中心医院，广东省江门市，529000）
（2 网络信息科，江门市中心医院，广东省江门市，529000）
通讯作者：叶均明，47480355@qq.com

【摘要】目的/意义 构建支持多源异构即插即用的医学文档知识库系统，实现从精准知识检索到专业问答的一体化服务，以提升知识检索效率与应用便捷性。**方法/过程** 在本地搭建多源异构文档知识库，基于动态权重混合检索策略，实现知识库文档标准化的解析提取及个性化的即插即用；利用检索增强技术（RAG）融合本地医学知识库与大语言模型，通过多级缓存、流式响应等优化机制，构建高效、可溯的医学专业问答生成系统。**结果/结论** 基于自建的多源异构医学知识测试集进行验证，结果表明：相较于单一的向量检索或关键词检索策略，本系统的检索准确率得到显著提升；同时，依托大语言模型的内容生成能力，可满足医务人员即时知识获取与智能问答的实际需求。

关键词：医学知识库；即插即用；混合检索；检索增强生成；智能问答

Design and Implementation of a Hot-swappable Medical Knowledge Base System Based on Hybrid Retrieval Strategy

WANG Peng[1], YE Junming[2,*], CHEN Yuchong[1],HAN Chunchun[1],WANG Gengbin[1],RONG Weixin[1],
(1 Center for Big Data and Artificial Intelligence, Network Information Department, Jiangmen Central Hospital, Jiangmen, Guangdong Province, 529000, China)
(2 Network Information Department, Jiangmen Central Hospital, Jiangmen, Guangdong Province, 529000, China)

【Abstract】 Objective: To construct a plug-and-play multi-source heterogeneous medical document knowledge base system, enabling an integrated service from precise knowledge retrieval to professional question answering, thereby enhancing retrieval efficiency and application convenience. Methods: A local multi-source heterogeneous document knowledge base was built, achieving standardized document parsing and extraction with personalized plug-and-play capability through a dynamic weighted hybrid retrieval strategy. By integrating the local medical knowledge base with a large language model (LLM) via Retrieval-Augmented Generation (RAG), an efficient and traceable medical question answering system was developed through mechanisms such as multi-level caching and streaming responses. Results: Validation on a self-built multi-source heterogeneous medical knowledge test set demonstrated that, compared to pure vector-based or keyword retrieval strategies, the system significantly improved retrieval accuracy. Leveraging the LLM's generation capabilities, it effectively meets the urgent needs of medical professionals for instant knowledge acquisition and intelligent question answering.

Keywords:medical knowledge base;plug-and-play;hybrid retrieval; retrieval-augmented generation; Intelligent Question and Answer

1 引言

医学是知识与数据密集型领域，具有知识专业性强、更新快且载体多元异构的特点。医学教材、学术论文、临床指南、药物说明书等大部分以 Word 或 PDF 格式存储，临床试验数据则多以 Excel 格式归档[1-2]。传统检索策略在应对

此类格式多样、专业性突出、语义要求高的场景时，普遍存在格式兼容性差、语义匹配弱、检索效率低等问题[3-6]。

当前医疗知识管理主要面临三方面挑战：其一，多源异构文档的统一处理存在瓶颈。数据来源涵盖 DOCX、PDF、XLSX、MD 等多种格式，传统方法难以实现跨格式无损提取与语义统一表示；其二，医学知识库动态适配性不足。传统知识库多采用固定架构，难以支持医务人员按岗位或科室需求动态调整与实时更新，亦难以满足即时应用需求；其三，从精准检索到智能问答的跨越存在困难。即使检索到相关片段，仍需人工阅读、筛选与整合，效率有待提升[7]。因此，面向多源异构医学文档的知识库系统与智能检索技术研究具有重要现实意义。

近年来，随着语义检索与人工智能技术的快速发展，国内外学者针对医疗知识库构建和检索技术的优化开展了大量研究[8-10]。其中，Google Health 提出了基于大模型的医疗知识检索框架；解放军总医院研发了面向临床应用的指南知识库；清华大学提出了多格式医疗文档的解析模型；北京协和医院结合大模型与检索增强技术（Retrieval-Augmented Generation，RAG）[11]搭建了智能化医学问答系统。上述研究为临床、科研等场景下的知识获取提供了技术参考，但多数研究聚焦于单一格式的文档处理或固定领域知识库设计，对知识库动态扩展支持及个性化即插即用能力的关注相对较少。

基于目前研究现状及医疗领域的实际需求，本文设计了一个支持知识文档即插即用的智能医学知识库系统。本文将硬件领域的“即插即用”概念引入医学知识库，特指知识源（文档、文件夹等）在不中断服务的前提下被动态识别与更新的能力。与传统固定架构知识库需人工维护（如重启服务）不同，本系统通过文件监测与增量更新技术，允许用户按需随时增删知识文档，实现知识资产的"即添即用"与动态迭代。为此，本文开展的工作包括：（1）设计多源异构文档解析引擎，支持知识文档的动态加载与实时索引更新；（2）提出面向医疗文本的动态混合检索模型，通过查询自适应权重提升检索精度；（3）集成大语言模型构建问答生成模块，实现从检索到解答的有效衔接。实验验证了系统在检索精度、响应速度及问答质量方面的有效性。

2 系统设计与实现

2.1 总体架构

医学知识来源分散、格式异构且存在个性化整合需求，本文提出一种支持知识源动态接入与管理的模块化架构，用于构建从多源异构文档处理到智能问答生成的端到端知识库系统。如图 1 所示，该系统自底向上分为数据处理层、存储管理层、服务核心层和应用接口层。各层之间通过标准化接口进行通信，以保证系统的松耦合性、可扩展性良好支持。

数据处理层作为系统的基础，用于多源异构医学文档的解析、预处理以及知识化组织。本层内置 12 种常见文档格式（如 DOCX、PDF、XLSX、PPTX 等）的专用解析器，可根据文档格式自动调用相应处理模块，实现文本内容、表格数据及基本结构等信息的标准化提取；存储管理层负责系统核心数据存储和组织，该层依据不同的业务场景需求，采取相应的存储策略以支撑多样化的数据访问模式，并支持知识库的动态更新；服务核心层作为系统的智能中枢，主要包括混合检索引擎模块与智能问答生成模块。检索引擎模块可同时执行向量语义检索与增强型关键词检索任务，并采用动态权重融合算法对两种检索方式的结果进行整合，输出 K 个最相关的文档片段；应用接口层为前端应用或外部系统提供统一的 API 服务，封装了文档上传、知识库管理、混合检索、智能问答等核心功能接口。

图 1 系统整体架构

2.2 核心模块设计

2.2.1 多源异构文档解析引擎

不同格式文档在编码、结构、语义等方面存在显著差异，传统方法通常针对单一格式设计专用解析器，在应对多格式场景时易出现“解析逻辑碎片化”与“输出表示非统一”的问题。为此，本文提出分层解析方法，将流程划分为协议层适配与语义层统一两个阶段：协议层依据格式提取原始内容，语义层映射为统一中间表示。新增格式仅需实现协议适配器，语义层可复用。协议层将文档归纳为文本类、富文本类、表格类及特殊格式类，每类配置差异化提取规则，同类共享相近解析逻辑。

文本类（TXT、MD）文档结构简单，解析难点在于编码格式的识别与处理。系统首先通过统计方法完成编码检测，并建立编码回退机制(优先使用 UTF-8 编码，若解析失败则依次尝试 GBK、GB2312 等常见编码)。

富文本类包括 PDF、DOCX、PPTX 等类型文本，包含复杂的排版样式和逻辑结构，需在解析时同时兼顾内容与结构的完整性[13]。针对跨页语句断开问题，算法通过识别句末标点及下行文字特征，实现智能语句拼接。DOCX 文档使用 python-docx 库按文档对象模型顺序提取段落、表格、标题等结构。如遇表格，将其内容转化为带有表头与行列描述的格式化文本（如“表 1：患者基本信息|姓名：张三|年龄：45 岁”），以确保表格信息的可检索性与可读性[14-16]。PPTX 文档则通过 python-pptx 库按“幻灯片→形状（Shape）”的层级遍历提取文本，并在每段文本前附加幻灯片序号标记（如“[Slide3]”），以维持演示文稿的上下文顺序与逻辑。

表格类格式（XLSX、CSV）以结构化数据为主，重点解析数据的行列关系与可追溯性。系统将每行数据转为制表符分隔的文本流，并在元数据中记录工作表名称。例如，检验报告可能表示为“工作表：血常规报告\n 行 1：项目\t 结

果\t 单位\t 参考范围”。通过将结构化数据转化成这种一维数据形式，可在最大程度上保留数据的上下文信息，便于后续检索结果的解读与回溯。

对于特殊格式（HTML/RTF），解析时去除格式化字符，保留纯文本内容。对于 HTML，系统通过过滤<script>、<style>等非内容标签，提取<body>内有效文本，可选择性保留<p>、<li>等语义标签。对于 RTF 文档，则通过正则表达式匹配并移除 RTF 控制符（如\par、\b）及转义序列，从而提取出纯文本内容。

经协议层解析，各类文档的输出被转换为统一的中间表示形式，为后续的文档分块、向量嵌入及检索溯源提供标准化的数据输入。

2.2.2 文档分块与向量嵌入

本模块采用语义感知的自适应分块策略，并建立向量化索引。

（1）文档分块策略

首先对文本进行预处理，使用 jieba 分词及中英文标点规则识别完整语句边界，以句子为基本语义单位。实现方式是采用滑动窗口分块的策略，设定窗口大小为 500 字符、重叠长度为 100 字符。在中文医学文本中，500 字符约含 3 至 5 个句子，构成完整语义单元；100 字符重叠可保留上一块末尾 1 至 2 个核心句子，作为块间语义桥接，避免上下文断裂。若当前块已包含完整句子且加入下句将超限，则结束当前块，以其末尾 100 字符作为下一块的开头。对于长度超过块限制的单句，则将其独立成块。

每个文本块均有对应的元数据信息，包括唯一标识符、来源文件信息、所在位置信息和文档类型等，以支持精准溯源及增量更新。

（2）向量化与索引构建

本模块将离散的文本块转换为可高效检索的向量表示，并构建支持快速相似度搜索的向量索引，为后续混合检索提供语义层面的支撑。

针对医学领域术语专业性强、语义表达复杂的特点，本模块选用 text-embedding-ada-002 作为基础嵌入模型。该模型具备较强的语义表征能力，能够将不同长度的文本块统一映射为 1536 维的高维向量空间表示，在兼顾计算效率的同时，较好地保留了文本的语义信息。

为提升相似度计算的精度与稳定性，系统对所有向量执行 L2 范数归一化，使余弦相似度等价于向量点积，降低模长差异影响。归一化后的向量集通过导入 FAISS 库中以构建高效索引。

2.2.3 混合检索引擎

本模块融合向量语义检索与增强型关键词检索，兼顾语义深度与术语匹配精度。

（1）向量语义检索模块

向量语义检索模块目的是捕捉查询块与文档块之间的深层语义关联。假设查询变量为 Q，首先利用预训练的嵌入模型将其映射为高维向量 vq，然后基于 FAISS 构建的文档块向量索引，执行近似最近邻搜索，返回与 vq 余弦相似度最高的 K 个候选文档块。该过程形式化表示如下：

$$Top\ K = arg\max_{D_i \in D} \frac{vq \cdot v D_i}{\| vq \| \vee \| v D_i \| \vee \|} \quad (1)$$

其中，D 表示文档块全集，$v D_i$ 为文档块 D_i 对应的向量表示。该方法能够有效处理医学领域常见的同义词现象（如“心肌梗死”与“心梗”）及表述模糊的语义查询。

（2）增强型关键词检索模块

为弥补向量检索在精确术语匹配上的不足，本文同时构建基于 BM25 算法的关键词检索模块，查询块 Q 与文档块 D 的相关性得分为 $BM25(D,Q)$。为进一步提升算法在医学知识库领域的性能，本文采用字段加权、模糊匹配与容错等增强策略。

基于文档结构的信息密度理论[17]，文档中不同结构位置的文本，对主题的概括能力存在显著差异，标题、文件名、正文依次递减，因而定义权重向量 $w_f=[w_{title}=3.0, w_{filename}=2.0, w_{content}=1.0]$ 为对应权重，其中 F={title,filename,content}，最终得分由各字段得分加权求和得出。

$$BM25_{enhanced}(D,Q)=\sum_{f\in F} w_f \cdot BM25(D_f,Q) \quad (2)$$

模糊匹配与容错。通过计算归一化编辑距离 $norm_{ed}(q,t)$（参考近似字符匹配理论[18]，设定阈值 $\theta=0.3$）识别拼写变体与简称，并对匹配得分施加衰减系数 $\alpha=1-norm_{ed}(q,t)$，增强检索的鲁棒性。

$$Score_{fuzzy}(D,q,t)=\begin{cases}\alpha\cdot BM25(D,\{t\}), & if norm\_ed<\theta\\ 0, & otherwise\end{cases} \quad (3)$$

（3）动态权重融合与结果生成

本文采用动态权重融合策略，首先对向量检索得分 S_v、关键词检索得分 S_k 进行 Min-Max 归一化分别得到 S_v' 和 S_k'，以消除量纲影响。然后根据查询块特征动态分配融合权重 λ（向量检索权重）与 $1-\lambda$（关键词检索权重）。根据 Perez 等[19]的研究，长查询（≥20 字符）通常包含更丰富的语义信息，更适合由稠密检索主导，设定 λ=0.7，以强化语义理解能力；而包含精确标识符（如药品代码）的查询则更适合由稀疏检索主导，设定 λ=0.4，以突出精确匹配优势。候选文档块结合排序与去重算法，其最终得分计算为：

$$S_{final}(D_i)=\lambda\cdot S_v'(D_i)+(1-\lambda)\cdot S_k'(D_i) \quad (4)$$

最后，依据 S_{final} 降序排序后，采用基于 SimHash 的近似去重算法移除内容高度重复的文档块，最终输出 *TopK* 个高质量、多样化的检索结果。该混合检索引擎通过两种检索机制的互补与智能融合，在医学文本检索任务中实现了语义深度与匹配精度的有效统一。

2.2.4 智能问答生成模块

本模块是实现从知识检索到专业答案输出的关键，该模块基于 RAG 框架，将混合检索引擎返回的相关文档块与用户查询有机结合，通过大语言模型生成结构清晰、依据可溯的专业答案。

当系统收到查询后，调用混合检索引擎获取前 K 个文档块，将其文本及元数据与用户查询按预设 Prompt 模板组装为模型输入。该模板要求模型仅基于给定资料生成答案，并标注信息来源，确保可追溯性。例如，用户询问药物用法时，模型严格依据说明书片段回答，并以"根据《药品说明书》第 X 条"格式标注来源。模块支持对接 Qwen3、DeepSeek 等通用模型及 Dify 平台，可按需灵活切换，并通过配置温度、惩罚系数等参数控制生成行为。

系统采用 SSE 流式响应，将生成内容逐字推送至客户端。用户可在极短时间内看到首个字符，有效减少等待感知；渐进式反馈模拟人工输入过程，增强交互的自然性与可预期性；用户可根据初步生成内容判断答案方向，必要时可中断请求。

2.2.5 系统性能优化策略

为提升系统在大规模数据处理与高并发访问下的性能与稳定性，本系统实施了针对性的性能优化方案。

（1）多级缓存机制

对检索引擎及问答生成层分别采用 LRU 策略做查询缓存与结果缓存。检索

引擎查询缓存可使高频次检索请求的响应时延由原来的 1.2s 降低至 0.15s 以下；对检索引擎及问答缓存，在高频次检索或问答过程中都可以避免调用大模型，大幅度节省了处理时长。

（2）增量式知识更新

通过监测文件修改时间，智能识别新增或变动的文档，仅对其执行解析、分块与索引更新操作，避免全量重建带来的资源消耗，支持知识库的高效、低延迟迭代。

3 实验与结果分析

3.1 测试数据集构建

为验证系统在多格式医学知识文档处理与检索方面的有效性，本文构建了一个具有代表性的医学领域混合文档测试集（120 篇文档），该测试集包含：50 篇 PDF 格式的医学研究论文及临床实践指南、30 份 DOCX 格式的药物说明书及药理信息文档、20 份 Excel 格式的临床试验数据集与统计分析报告，以及 20 份 PPTX 格式的医学教材课件及教学资料。

3.2 性能评估

3.2.1 精度评估指标

本实验采用 P@K（Precision at K）作为评估指标，P@K 是信息检索领域核心评估指标之一，用于衡量检索系统在返回的前 K 个结果中的相关性。该指标直接反映用户在实际使用中最关注的前 K 个检索结果的相关性，与医务人员期望快速获取最相关信息的真实需求高度契合；本实验中设置 K=1、3、5 三个取值，分别测试系统在返回首个结果、前 3 个结果及前 5 个结果时的准确率。对比不同检索方法，结果如表 1 所示。

本文提出的动态权重混合检索模型在各项指标上均优于单一的向量检索或关键词检索，在 P@3 上达到 82.3%，相较于向量检索（76.5%）和单一关键词检索（68.7%）分别提升 5.8 个百分点和 13.6 个百分点。在 P@1 上，混合检索较向量检索提升 10.3%，说明该算法能更准确地将最相关内容置于首位。相比之下，关键词检索效果不佳，表明仅依赖字面匹配难以应对医学术语的同义词、缩略语等语义模糊问题，同时也验证了动态权重策略的有效性。

表 1 不同检索方式准确率对比(P@K)

检索方式	P@1	P@3	P@5
向量检索	72.1%	76.5%	78.3%
关键词检索	65.3%	68.7%	71.2%
本文提出的方法	**79.5%**	**82.3%**	**83.6%**

3.2.2 不同文档格式的检索准确率

如表 2 所示，系统对 PDF 学术论文和 DOCX 临床指南的处理效果更佳（P@3 分别为 84.1% 和 83.2%），对 Excel 表格数据的检索效果略低（79.5%），主要因表格内容转换为文本后部分结构信息丢失，但整体仍保持较高精度。PPTX 教学课件的检索精度 P@3 为 81.8%，验证了解析引擎对演示文稿格式的有效处理。

表 2 不同文档类型检索准确率对比(P@3)

文档类型	样本数	P@3
PDF（学术论文）	50 篇	84.1%
DOCX（临床指南）	30 份	83.2%
Excel（检验数据）	20 份	79.5%
PPTX（教学课件）	20 份	81.8%

3.2.3 智能问答质量评价

（1）测试查询集构建。由两位具有医学信息学背景的研究人员（非本文作者）从医院日常临床咨询中收集 50 个真实查询，涵盖药物信息查询（如“阿托伐他汀的用法用量”）、治疗指南查询（如“2 型糖尿病的一线治疗方案”）、医学检验指标解读（如“HbA1c 大于 7%的意义”），分别占比 40%、35%、25%。由该两位研究人员依据答案获取方式将查询分为简单（单一文档片段，20 个）、中等（2-3 个片段整合，20 个）及困难（跨文档综合推理，10 个）三个难度等级，若有分歧通过讨论进行解决。

（2）评测流程与结果。三位具有中级及以上职称的临床医师（平均工作年限 5 年，均未参与系统开发与测试集构建）参与评测。评价采用双盲设计（评测人员不知晓答案来源系统，且互不知晓彼此的评分结果），三者独立评分，评测顺序随机化。每位评测人员从专业性、完整性、准确性及可读性四个维度对大模型生成答案进行评分，采用 5 分 Likert 量表。50 个测试查询的平均得分分别为 4.2 分、4.0 分、4.3 分、4.1 分（图 2），表明系统整体生成的答案质量较好，能够满足临床医学问答的实际需求。

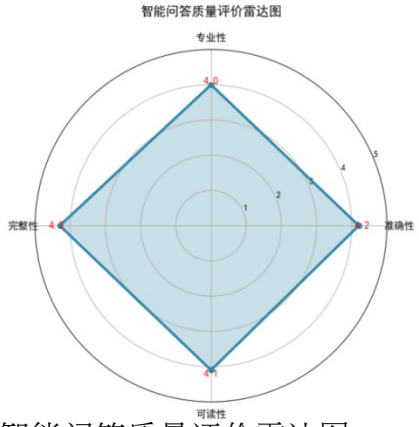

图 2 智能问答质量评价雷达图

3.2.4 系统的响应性能评估。

搜索响应的平均时间为 1.2 秒（无缓存），缓存命中后降为 0.15 秒，问答生成的时间依据答案的长度有所差别，一般在 3-8 秒之间，流式输出首字响应时间小于 1 秒。

3.3 系统效果

本系统部署的智能助手交互界面如图 3 所示，右右侧功能面板集成知识库状态管理、内网部署模式切换、对话历史维护及上下文控制等核心模块。对话区域以时间线形式呈现交互过程，支持用户与系统之间的多轮对话。界面底部设有结构化输入框，以引导用户输入查询内容，确保用户在临床或科研场景中能够快速、便捷地获取结构化医学知识支持。系系统采用本地化部署，保障数据安全性与响应实时性，为院内专业知识服务提供可靠平台。

图 3 系统交互界面

4 讨论与结论

针对医学知识管理中的多源异构文档处理与智能检索需求，本文设计并实现了一个融合多格式文档解析、混合检索及专业问答生成的知识库系统。该系统采用模块化、支持**即插即用扩展**的文档解析架构，可对多种主流医学文档格式进行统一处理与语义化组织，具有良好的可扩展性。

本文采用的动态权重混合检索模型融合了向量语义检索与增强型关键词检索两种技术路径，在自建医学文档测试集上的 P@3 检索准确率达到 82.3%，相较于单一向量检索或关键词检索有一定提升。系统将检索结果与大语言模型生成能力相结合，可生成来源可靠的结构化回答。此外，系统通过多级缓存、增量更新及流式响应等工程优化，在实际部署中获得良好的响应性能与交互体验。

未来工作将围绕多模态信息处理、复杂推理优化、领域适应性增强及结果可信度评估等方面持续深化，尤其探索集成 OCR 图像识别、医学影像分析等新模态数据的处理，以适应医学领域不断发展演进的技术需求。

参考文献

[1] 雒增月,尹裴.基于多源异构信息融合的肺癌专病知识图谱构建研究[J].建模与仿真, 2025, 14(10):10.DOI:10.12677/mos.2025.1410603.

[2] 张振,周毅,杜守洪等.医疗大数据及其面临的机遇与挑战[J].医学信息学杂志, 2014, 35(6):2-8.DOI:10.3969/j.issn.1673-6036.2014.06.001.

[3] 李桃,郑西川,蒋伏松.基于知识库的临床决策支持系统的设计与应用[J].医疗卫生装备, 2019, 40(5):4.DOI:10.19745/j.1003-8868.2019112.

[4] 王彩云,郑增亮,蔡晓琼,等.知识图谱在医学领域的应用综述[J]. 生物医学工程学杂志，2023,40 (5):1040-1044.

[5] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[6]蒋献,王涵亦,杨诗婷,等.基于大语言模型与知识图谱融合的多跳问答技术研究[J].电信科学, 2025, 41(11):67-83.

[7]焦鹏程,钱欣玮,张弛,蔡叔梅.基于大模型的本地知识库技术研究[J]. 2025.

[8]王文湖,韦昌法.基于大语言模型和知识库的阿尔茨海默病智能问答系统构建研究[J].世界科学技术-中医药现代化, 2025(3).

[9]胡佳慧,李姣,姚宽达,等.大语言模型融合知识图谱的医学问答系统构建研究[J].中国数字医学, 2024, 19(6):91-95.DOI:10.3969/j.issn.1673-7571.2024.06.017.

[10]王斌,刘涛,王广志等.支持新型冠状病毒肺炎的中医智能处方推荐和知识库系统[J].中国数字医学, 2020, 15(5):3.DOI:CNKI:SUN:YISZ.0.2020-05-010.

[11] Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for

knowledge-intensive nlp tasks[J]. Advances in Neural Information Processing Systems, 2020, 33: 9459-9474.

[12] Johnson J, Douze M, Jégou H. Billion-scale similarity search with GPUs[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 5354-5362.

[13] 姚琳,马鹏飞,尚丹梅,等.基于 Python 程序设计的学科知识图谱构建研究[J].中国管理信息化, 2025, 28(12):127-130.

[14] Xu Y, Li M, Cui L, et al. LayoutLM: Pre-training of text and layout for document image understanding[C]. Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020: 1192-1200.

[15] Yasunaga M, Leskovec J, Liang P. LinkBERT: Pretraining language models with document links[J]. arXiv preprint arXiv:2203.15827, 2022.

[16] Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining[J]. Bioinformatics, 2020, 36(4): 1234-1240.

[17] Trotman A. Choosing document structure weights. Information Processing and Management 41(2):243–264

[18] Navarro G, Raffinot M. Flexible pattern matching in strings: practical on-line search algorithms for texts and biological sequences[M]. Cambridge: Cambridge University Press, 2002.

[19] Perez J, Zhou J, Le N, et al. Entropy-Based Dynamic Hybrid Retrieval for Adaptive Query Weighting in RAG Pipelines[C]. ICML 2025 Workshop on Vector Databases, 2025.

利益冲突声明

所有作者均声明不存在利益冲突。

作者贡献声明

王鹏负责系统总体架构设计、混合检索算法研究及论文撰写；叶均明负责研究方案设计、技术指导及论文修改；陈宇聪负责多源异构文档解析引擎开发与实验数据采集；韩春春负责知识库构建与向量嵌入模块实现；王耿彬负责智能问答生成模块开发与性能测试；戎伟鑫负责系统集成与前端交互界面开发。