知识驱动的混合检索增强生成方法在罕见病领域的应用研究

张　妁　方　安　娄　培　赵琬清　姚宽达　胡佳慧

（北京协和医学院/中国医学科学院医学信息研究所/图书馆　北京 100020）

〔摘要〕　**目的/意义** 探索知识驱动的混合检索增强生成方法，以解决大语言模型在医学垂直领域事实准确性不足与可靠度低的问题。**方法/过程** 构建以罕见病为代表的医学领域知识库，综合运用基于关键词匹配的稀疏检索与基于语义相似度的稠密向量检索，并采用倒数排序融合算法对双重检索结果进行优化排序，通过混合检索增强生成为大语言模型提供最优上下文。**结果/结论** 该方法的准确率、召回率等指标均优于单一检索与其他检索策略，并在多个大语言模型对比中具有良好的鲁棒性，有助于提升医学知识服务的智能化水平。

〔关键词〕　大语言模型；检索增强生成；混合检索；知识服务；罕见病

〔中图分类号〕R－058　　〔文献标识码〕A　　〔DOI〕10.3969/j.issn.1673－6036.2026.02.012

Study on the Application of Knowledge－driven Hybrid Retrieval－augmented Generation in the Field of Rare Diseases

ZHANG Xu, FANG An, LOU Pei, ZHAO Wanqing, YAO Kuanda, HU Jiahui

Institute of Medical Information/Medical Library, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100020, China

〔**Abstract**〕　**Purpose/Significance** To explore a medical knowledge－driven hybrid retrieval－augmented generation method (Med－HyRAG), so as to address the issues of insufficient factual accuracy and low reliability of large language model (LLM) in the medical vertical domain. **Method/Process** A medical knowledge base focused on rare diseases is constructed. The keyword matching－based sparse retrieval and semantic similarity－based dense vector retrieval are comprehensively applied, and the reciprocal ranking fusion algorithm is employed to optimize and rank the dual retrieval results, thereby providing an optimal context for LLM through hybrid retrieval－augmented generation. **Result/Conclusion** The indicators such as accuracy, recall of this method are superior to those of single retrieval and other retrieval strategies, and it has good robustness in the comparison of several LLMs. This study contributes to advancing the intelligence of medical knowledge services.

〔**Keywords**〕　large language model (LLM); retrieval－augmented generation (RAG); hybrid retrieval; knowledge service; rare diseases

〔修回日期〕　2025－10－15

〔作者简介〕　张妁，硕士研究生，发表论文 2 篇；通信作者：胡佳慧，副研究员。

〔基金项目〕　中国医学科学院医学与健康科技创新工程项目（项目编号：2021－I2M－1－056）；中央高水平医院临床科研专项（项目编号：2022－PUMCH－A－084）。

1 引言

大型语言模型（large language model，LLM）在通用领域表现出色[1-4]，但在高度专业化、知识体系复杂且对精确性要求严苛的医学垂直领域，特别是在数据稀疏的罕见病领域，仍存在局限性[5-6]。一方面，医学专业术语体系不仅规模庞大，且处于持续演化状态[7-8]，通用预训练数据难以实现对特定领域知识，尤其是新兴医学知识的全面覆盖，模型存在知识盲区，可能输出事实性错误[9-11]，临床决策风险加剧。另一方面，医学数据呈长尾分布特性，特定病种样本在训练集中比较稀缺，进一步加剧模型产生“幻觉”或输出事实性错误的风险[12-13]。

对此，检索增强生成（retrieval - augmented generation，RAG）可通过融合外部知识库提升 LLM 的可靠性[14-15]。然而，在复杂的医学场景中，传统 RAG 的单一检索策略难以同时满足精确医学术语匹配与多样化口语描述的需求[16-18]。一方面，临床实践高度依赖标准化、精确的医学实体，例如特定的药品名称、基因分型或诊断标准。对于此类查询，基于关键词匹配的稀疏检索因其对精确字符串的高敏感性而表现优异，能够确保关键术语不被曲解或忽略；但其无法理解语义层面的关联[19]。另一方面，患者或初级医师的提问往往是非结构化的自然语言，大部分为同义词、近义词和描述性短语。对于此类查询，基于语义相似度的稠密检索能够凭借其强大的语义泛化能力，理解查询的真实意图，即使查询与文档在字面上并无交集[20]；但其对特定专业术语的敏感度不足。稀疏检索和稠密检索均可能导致关键信息遗漏或产生偏差[14,19]。

针对上述挑战，本研究提出医学知识驱动的混合检索增强生成方法（hybrid retrieval - augmented generation method driven by medical knowledge，Med - HyRAG），并以罕见病领域为例，利用该方法融合两种检索策略的优势，既保证对专业术语的精确匹配，又兼顾对自然语言查询的语义理解；通过倒数排序融合（reciprocal rank fusion，RRF）算法对检索结果进行融合优化，为下游 LLM 生成提供高质量、高相关的上下文知识。

2 相关研究

针对 LLM 在知识密集型任务中固有的“幻觉”[20]及知识更新滞后问题[21-22]，Lewis P 等[14]于 2020 年提出 RAG 框架，旨在融合 LLM 内部知识与可检索的外部知识。该框架包含两个阶段流程：一是检索器根据用户输入，从海量文档语料库中召回最相关的知识片段；二是生成器将检索到的片段作为上下文信息，指导 LLM 生成基于证据的可靠响应。RAG 框架的性能高度依赖检索器的效率与准确性。传统的 RAG 多采用稀疏检索方法，如 TF - IDF[23]、BM25[19]，基于关键词匹配，计算效率高且可解释性强，在处理包含精确术语的查询时表现优异。但缺乏对词汇深层语义关系的理解能力，使其在应对同义词表述或口语化的自然语言查询时表现受限[24]。

随着表示学习技术的发展，基于语义相似度的稠密检索逐渐成为主流方法。其运用文本嵌入模型[25-26]，将查询和文档嵌入映射到同一高维向量空间，通过计算向量间的相似度衡量相关性，有效弥补了传统稀疏检索在语义理解上的不足。但其对特定的关键词不够敏感，尤其是在处理专业术语、代码片段或命名实体时可能出现偏差。

RRF 算法[27]是一种融合策略，避免了引入额外参数训练对特定数据集产生的过拟合风险，增强了方法的泛化能力。其对不同检索系统的原始得分并不敏感，而是更关注文档在各自排序列表中的相对位置，因此可以公平地整合异构的检索源。在医学场景中，只要在单一检索范式中获得较高排名，其最终的融合分数便会显著提升，该机制避免了单一检索源的系统性偏差。

此外，现有生物医学文献与训练数据中存在显著研究偏倚[28]。相较于癌症等研究热点领域，罕见病领域研究相对较少，导致 LLM 在预训练阶段难以充分学习其知识。为此，本研究设计并验证医学知识驱动的混合检索增强生成方法，选取罕见病作为医学垂直领域代表，以期优化医学知识服务的智能化水平，并提高其可靠程度。

3 研究方法

3.1 总体思路

本研究提出的 Med – HyRAG 方法包括医学知识库构建和混合检索增强生成两个核心环节，见图 1。其中，医学知识库构建是基础，其任务是将非结构化文本中的医学知识转化为机器可检索的结构化表示；在此基础上，混合检索增强生成环节响应用户实时查询，通过并行混合检索模块召回相关知识片段，并将其作为上下文注入基座 LLM，以生成最终回答。

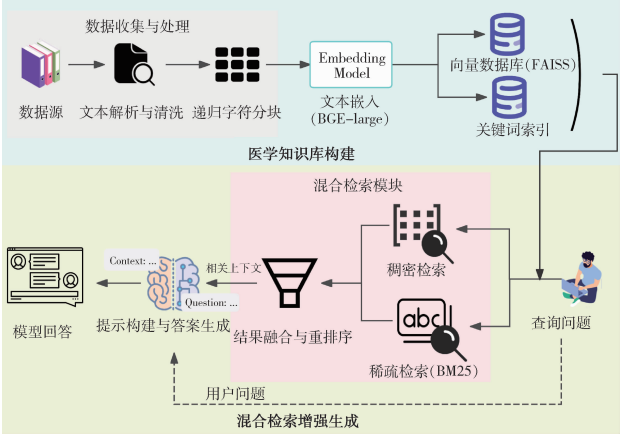

图 1 Med – HyRAG 整体框架

3.2 知识库构建

3.2.1 数据收集与处理 以国家卫生健康委员会发布的《罕见病诊疗指南（2019 年版）》[29-30] 和《软骨发育不全等 86 个罕见病病种诊疗指南（2025 年版）》[31] 为数据来源。在数据处理阶段，解析原始文档，提取纯文本内容，并进行数据清洗，如去除页眉、页脚、目录等无关信息。鉴于文本嵌入模型处理长文本的局限性以及检索单元的粒度需求，采用递归字符分割策略对文本进行分块。在此基础上，结合段落（/n/n）、单行换行符（/n）和句子结束符进行分割，确保文本块在语义上的完整性。为保证上下文的连续性，在相邻块之间设置 10% 的重叠。

3.2.2 文本嵌入 为实现医学知识的向量化表示，选取 3 个嵌入模型进行对比评估。MPNet – base – v2[32] 是基于 MPNet 架构的通用句子嵌入模型，在多种英文语义任务上表现优异，用以评估跨语言、通用模型在中文医学领域的基线性能。Text2vec – base – Chinese[33] 是经典的、专门为中文设计的通用词向量与句向量表征工具，在国内具有广泛的应用基础。BGE – large – zh – v1.5[34] 是由智源研究院研发的大规模中文嵌入模型，在多个中文语义检索评测基准上处于领先地位。对上述模型进行量化指标评测，并选择性能最优者作为后续所有实验的唯一嵌入模型。对于每个经过分块的文本段 d_i，通过所选模型得到高维语义向量 v_i。其中，*EmbeddingModel* 代表最终选定的嵌入模型。所有文本块的向量共同构成支持高效相似性检索的向量数据库。

$$v_i = EmbeddingModel(d_i) \tag{1}$$

3.3 混合检索增强生成

3.3.1 稠密检索 稠密检索旨在召回与用户查询语义相近的文本块。当用户输入查询时，首先使用与知识库文本块编码相同的嵌入模型将其转换为查询向量，随后通过计算余弦相似度，在向量数据库中检索并返回 K 个中最相似的文本块。

3.3.2 稀疏检索 采用 BM25 算法，基于查询词的逆文档频率和词频等信息，对文本块的相关性进行打分。将该算法的可调参数设为 $k_1 = 1.5$，$b = 0.75$。

3.3.3 结果融合与重排序 获得向量检索与关键词检索的两个排序列表后，运用 RRF 算法对结果进行整合与重排序。根据每个文档在不同检索结果列表中的排名计算其最终得分，排名越靠前，贡献的分数越高，其计算方式如下。

$$RRF - Score(C_i) = \sum_{s \in S} \frac{1}{k + rank_s(C_i)} \tag{2}$$

其中，S 为检索策略集合（即 {Vector, BM25}）；$rank_s(C_i)$ 表示文本块 C_i 在策略 S 结果中的排名，如果某文本块在向量检索中排名第 1 位，在 BM25 检索中排名第 10 位，则其对应的 $rank_{vector}(C_i)$ 为 1，$rank_{BM25}(C_i)$ 为 10。k 是平滑因子，用于平衡不同检索系统的贡献。经独立验证集测试，设定 $k = 60$ 时，归一化折损累计增益（normalized discounted cumulative gain, NDCG@ K）在 Top – K = 10 时达到最优。

RRF 作为一种无需训练的融合策略，通过关注文档的相对排名而非原始得分，公平整合异构检索源，有效提升方法的泛化性。

3.3.4 提示构建与答案生成 经过混合检索和重排序优化后，获得 K 个最相关的知识片段。将排序后的知识片段拼接成连贯的文本，与用户查询共同构建提示。为确保 LLM 能够准确、可靠地基于检索到的上下文生成答案，并最大限度地抑制“幻觉”，设计专门的提示模板，通过明确的角色设定、严格的任务指令和清晰的约束条件，引导模型的行为，见表 1。为验证 Med - HyRAG 的通用性，将构建的提示分别输入 3 个不同架构的基座 LLM，分别是 DeepSeek - V3.1[35]、Gemini - 2.5 - pro[36] 和 GLM - 4.5[37]，利用其上下文理解和零样本生成能力，对检索到的医学知识进行综合、提炼与组织，从而生成逻辑清晰、语言流畅且符合医学规范的回答。

表 1 提示结构示例

结构	内容
角色	你是一名严谨的、专业的罕见病领域医学问答助手
任务	请根据下面提供的“上下文信息”，用中文简洁、准确地回答“用户问题”
约束条件	（1）严格遵循上下文：你的回答必须完全基于“上下文信息”，禁止利用任何外部或你自身的先验知识。（2）信息不存在则明确指出：如果“上下文信息”中没有足够的内容来回答“用户问题”，请直接回复“根据提供的资料，无法回答该问题。”（3）避免推测与编造：严禁对上下文内容进行任何形式的推断、延伸或创造。（4）忠于原文：回答应尽可能地忠实于原文的表述
上下文信息	{retrieved_chunks}
用户问题	{user_question}
回答	……

4 实验及结果

4.1 实验数据

4.1.1 医学知识源 医学诊疗指南是指导临床实践、规范医疗行为的重要工具，为临床决策提供坚实知识基础。为构建权威且紧跟前沿的罕见病医学知识库，整合 2019 年版和 2025 年版罕见病指南作为外部知识源。前者全面覆盖我国《第一批罕见病目录》[38]中的 121 种罕见病，后者作为对前者的重要补充，新增 86 种罕见病的诊疗规范。由此构建的知识库共覆盖 207 种罕见病，包括这些疾病的定义、临床表现、诊断标准、治疗方案及预后等详尽信息。

4.1.2 评测数据集 基于 207 种罕见病的指南原文，由两名具有医学背景的研究人员独立进行评测数据集构建。为模拟真实的查询需求，问题被设计为多种类型，包括事实型、比较型、定义型和机制型。数据集中每个问答对均明确标注了其对应的原始文本块。为确保标注的客观性与一致性，引入质量控制流程，对标注数据进行跨标注者一致性检验（inter - annotator agreement，IAA）。针对核心标注任务“问答对 - 原始文本块”匹配，其 Cohen’s Kappa 值为 0.87，表明评测数据集一致性检验符合要求。由专家对评测数据集所有问答对进行交叉审核和终审。审核过程中，专家根据问题的质量、答案的准确性与原文的匹配度，将其评定为不同的质量等级。对于存在分歧的标注，通过集体讨论达成共识，以确保数据集的准确性、问题表述的清晰度以及答案的唯一性。最终构成的评测数据集示例数据，见表 2。

表 2 评测数据集示例（部分）

原始文本块编号	问题类型	问题	专家评审等级
chunk_233	事实型	根据 Schwartz 评分系统，LQTS 诊断中 QTc 间期的不同数值对应多少分	A
chunk_554	比较型	遗传性血管性水肿的长期预防治疗中，氨甲环酸相比达那唑有什么优势	A
chunk_1345	定义型	原发性轻链型淀粉样变诊断要满足几条标准	B
chunk_1004	机制型	线粒体脑肌病 MGT 染色的典型发现是什么	A

注：LQTS = 长 QT 综合征（long QT syndrome），MGT 染色 = 改良哥摩理三色（modified Gomori trichrome）染色。

4.2 评测指标

选取语义命中率和平均倒数排名（mean reciprocal rank，MRR）作为嵌入模型的评测指标。选取精确率（precision@）、召回率（recall@）、平均精度均值（mean average precision，mAP@）和 NDCG@ 综合评测检索策略。

4.3 实验结果

4.3.1 嵌入模型对比 嵌入模型对比结果，见表 3。BGE－large－zh－v1.5 与 Text2vec－base－Chinese 语义命中率相同，均优于以英文语料为主的 MPNet－base－v2。BGE－large－zh－v1.5 的 MRR 优于其他模型，表明其能更可靠地将相关知识排在更靠前的位置，这对 RAG 至关重要。因此选择 BGE－large－zh－v1.5 作为后续所有实验的唯一嵌入模型。

表 3 嵌入模型性能对比

嵌入模型	语义命中率	MRR
MPNet－base－v2	0.44	0.30
Text2vec－base－Chinese	0.48	0.28
BGE－large－zh－v1.5	0.48	0.37

4.3.2 检索策略性能对比 为进一步验证 Med－HyRAG 的有效性，引入假设性文档嵌入（hypothetical document embeddings，HyDE）[39] 作为对比基线。HyDE 旨在解决查询与文档表述不匹配的问题。其先利用一个指令遵循 LLM（本研究采用 DeepSeek－V3.1[35]），根据用户原始查询生成一篇风格与知识库文档风格相似的“假设性文档”；再对该富含上下文的假设性文档进行向量化，并以此生成的向量执行相似性检索，而非对原始短查询进行操作。对比稀疏检索（BM25）、稠密检索（Vector）、假设性文档嵌入检索（HyDE（DeepSeek－V3.1））以及本研究提出的混合检索（Med－HyRAG）4 种方法性能。采用 Precision@1 衡量检索准确性，采用Recall@10、mAP@10 和 NDCG@10 综合评估检索结果的全面性和整体排序质量。各项指标的性能对比，见表 4。Med－HyRAG 在所有指标上均显著领先。作为基线，BM25 和 Vector 展现了各自的检索倾向，但综合性能有限。先进的 HyDE 方法虽然通过 LLM 生成假设性文档小幅提升了排序质量，但其召回率甚至略有下降，这可能是由于 LLM 生成的文档在专业领域存在细节偏差。相比之下，Med－HyRAG 通过融合稀疏检索的精确性和稠密检索的语义理解能力，取得了最高的 Precision@1 和 Recall@10 值，证明其结果兼具准确性与全面性。因此，混合策略能有效弥补单一范式的不足，应对医学知识服务的双重挑战，为答案生成提供更高质量的上下文。

表 4 4 种检索方法性能对比

检索方法	Precision@1	Recall@10	mAP@10	NDCG@10
BM25	0.50	0.85	0.61	0.69
Vector	0.46	0.80	0.50	0.59
HyDE(DeepSeek－V3.1)	0.51	0.78	0.62	0.70
Med－HyRAG	0.53	0.88	0.65	0.71

4.3.3 不同基座模型性能对比 为验证 Med－HyRAG 框架并非仅对特定模型有效，应用 Med－HyRAG 的检索策略，并将 DeepSeek－V3.1、ChatGLM－4.5 和 Gemini－2.5－pro 作为基座模型进行性能评测。3 款不同架构的基座模型在接入 Med－HyRAG 后，检索性能表现出高度一致性，仅存在微小、可忽略不计的差别。综合来看，Med－HyRAG 具有良好的普适性，能够提供稳定的上下文环境，使不同架构、不同训练背景的 LLM 均能在此基础上表现出相近的性能。

4.3.4 生成答案质量对比 为了更直观地展示不同检索策略的差异，以 Gemini－2.5－pro 回答“ATTR－CA 患者 CMR 检查中钆延迟显像的典型表现是什么”为例进行分析。BM25 和 HyDE 策略生成的答案中，均包含对缩写“CA”的额外解释。虽然这个解释本身是正确的，但并非用户问题的直接要求。这在一定程度上表明，两种方法可能召回了包含该定义、相关但并非最核心的上下文，对 LLM 的最终输出造成了轻微的冗余干扰。相比之下，Vector 和 Med－HyRAG 生成的答案则更为简洁和聚焦。其没有引入对“CA”的额外解释，而是直接陈述典型表现。表明其召回的上下文质量更高，与用户问题的相关性更强，使 LLM 能够生成更精炼、更专业的回答。在 Vector 和 Med－HyRAG 之

间，Med－HyRAG 生成的答案以一段完整的陈述句呈现，其行文风格与医学指南中的专业表述更为贴近；而 Vector 生成的答案则采用了分段形式，虽然可读性较好，但在正式的医学问答语境下，Med－HyRAG 的表述更为权威。

5 结语

本研究针对 LLM 在医学垂直领域应用中面临的知识局限（易产生事实性错误）挑战，以及传统 RAG 方法难以同时兼顾专业术语精确匹配与多样化语义理解的不足，提出 Med－HyRAG 方法并进行验证。该方法以医学知识为基础，利用关键词匹配保障医学术语的检索准确性，结合语义向量检索理解用户查询意图的多样性，并通过倒数排序融合算法对结果进行智能优化重排。研究结果表明，相较于单一关键词检索和向量检索，Med－HyRAG 在所有评估指标上均优于单一检索基线，有效解决了医学专业术语处理和复杂查询理解的双重挑战。本研究有助于提升 LLM 在医学应用中的准确性与可靠度，促进医学知识服务的智能化。

本研究知识源主要聚焦罕见病诊疗指南，其文本结构与表述方式具有一定特异性。相较于研究充分的疾病，罕见病的相关知识在通用语料中覆盖率低且呈碎片化分布，使大语言模型在无外部知识辅助时，极易产生事实性错误。同时在其他疾病相关领域，本研究所提出的方法亦具有普适性。未来研究将探索融合影像学、病理报告等多模态数据的可行性。

作者贡献：张姒负责研究实施、论文撰写；方安负责提供指导；娄培负责研究设计；赵琬清、姚宽达负责数据处理；胡佳慧负责研究设计、论文审核。
利益声明：所有作者均声明不存在利益冲突。

参考文献

1 ZHAO W X, ZHOU K, LI J, et al. A survey of large language models [EB/OL]. [2025－10－09]. https://arxiv.org/abs/2303.18223.
2 ZHOU H, LIU F, GU B, et al. A survey of large language models in medicine: progress, application, and challenge [EB/OL]. [2025－10－09]. https://arxiv.org/abs/2311.05112.
3 YANG H S, WANG F, GREENBLATT M B, et al. AI chatbots in clinical laboratory medicine: foundations and trends [J]. Clinical chemistry, 2023, 69 (11): 1238－1246.
4 肖仰华，徐一丹．大规模生成式语言模型在医疗领域的应用：机遇与挑战 [J]．医学信息学杂志，2023，44 (9)：1－11.
5 SHYR C, HU Y, BASTARACHE L, et al. Identifying and extracting rare diseases and their phenotypes with large language models [J]. Journal of healthcare informatics research, 2024, 8 (2): 438－461.
6 PENG C, YANG X, CHEN A, et al. A study of generative large language model for medical research and healthcare [J]. NPJ digital medicine, 2023, 6 (1): 210.
7 RAJPURKAR P, CHEN E, BANERJEE O, et al. AI in health and medicine [J]. Nature medicine, 2022, 28 (1): 31－38.
8 胡振生，杨瑞，朱嘉豪，等．大语言模型在医学领域的研究与应用发展 [J]．人工智能，2023 (4)：10－19.
9 LONGWELL J B, HIRSCH I, BINDER F, et al. Performance of large language models on medical oncology examination questions [J]. JAMA network open, 2024, 7 (6): e2417641.
10 THIRUNAVUKARASU A J, TING D S J, ELANGOVAN K, et al. Large language models in medicine [J]. Nature medicine, 2023, 29 (8): 1930－1940.
11 MILLER R. A surgical perspective on large language models [J]. Annals of surgery, 2023, 278 (2): e211－e213.
12 XIONG G, JIN Q, LU Z, et al. Benchmarking retrieval－augmented generation for medicine [EB/OL]. [2025－10－09]. https://arxiv.org/abs/2402.13178.
13 KÖHLER S, SCHULZ M H, KRAWITZ P, et al. Clinical diagnostics in human genetics with semantic similarity searches in ontologies [J]. American journal of human genetics, 2009, 85 (4): 457－464.
14 LEWIS P, PEREZ E, PIKTUS A, et al. Retrieval－augmented generation for knowledge－intensive NLP tasks [EB/OL]. [2025－10－09]. https://arxiv.org/abs/2005.11401.
15 吴璇，付涛．检索增强生成技术研究综述 [J/OL]．计算机工程与应用，1－20 [2025－10－09]．https://link.cnki.net/urlid/11.2127.tp.20250610.1915.014.
16 KANATARIA N, PATEL K P, PATEL H N, et al. RAG－enhanced large language model for intelligent assistance from web－scraped data [C]. Coimbatore: 2024 9th International Conference on Communication and Electronics Systems

(ICCES), 2024.

17 ROBERTSON S, ZARAGOZA H. The probabilistic relevance framework: BM25 and beyond [J]. Foundations and trends in information retrieval, 2009, 3 (4): 333 - 389.

18 KARPUKHIN V, OĞUZ B, MIN S, et al. Dense passage retrieval for open - domain question answering [C]. Online: The 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020.

19 ARABZADEH N, YAN X, CLARKE C L A. Predicting efficiency/effectiveness trade - offs for dense vs. sparse retrieval strategy selection [C]. New York: The 30th ACM International Conference on Information & Knowledge Management, 2021.

20 JI Z, LEE N, FRIESKE R, et al. Survey of hallucination in natural language generation [J]. ACM computing surveys, 2023, 55 (12): 1 - 38.

21 RAWTE V, SHETH A, DAS A. A survey of hallucination in large foundation models [EB/OL]. [2025 - 10 - 09]. https://arxiv.org/abs/2309.05922.

22 HUANG L, YU W, MA W, et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions [EB/OL]. [2025 - 10 - 09]. https://arxiv.org/abs/2311.05232.

23 DAS B, CHAKRABORTY S. An improved text sentiment classification model using TF - IDF and next word negation [EB/OL]. [2025 - 10 - 09]. https://arxiv.org/abs/1806.06407.

24 LI Z, WANG Z, WANG W, et al. Retrieval - augmented generation for educational application: a systematic survey [J]. Computers and education: artificial intelligence, 2025, 8 (6): 100417.

25 DEVLIN J, CHANG M W, LEE K, et al. BERT: pre - training of deep bidirectional transformers for language understanding [C]. Minneapolis: The 2019 Conference of the North: Association for Computational Linguistics, 2019.

26 REIMERS N, GUREVYCH I. Sentence - BERT: sentence embeddings using siamese BERT - networks [C]. Hong Kong: The 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP - IJCNLP), 2019.

27 CORMACK G V, CLARKE C L A, BUETTCHER S. Reciprocal rank fusion outperforms condorcet and individual rank learning methods [C]. Amsterdam: The 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2009.

28 SERRA A, FRATELLO M, FEDERICO A, et al. An update on knowledge graphs and their current and potential applications in drug discovery [J]. Expert opinion on drug discovery, 2025, 20 (5): 599 - 619.

29 国家卫生健康委员会办公厅. 罕见病诊疗指南 (2019 年版) [EB/OL]. [2025 - 10 - 09]. https://www.nhc.gov.cn/yzygj/c100068/201902/073540e8f83b4a54a28684d23e2ae2f5.shtml.

30 LU Y, GAO Q, REN X, et al. Incidence and prevalence of 121 rare diseases in China: current status and challenges: 2022 revision [J]. Intractable & rare diseases research, 2022, 11 (3): 96 - 104.

31 国家卫生健康委员会办公厅. 软骨发育不全等 86 个罕见病病种诊疗指南 (2025 年版) [EB/OL]. [2025 - 10 - 09]. https://www.nhc.gov.cn/yzygj/c100068/202507/5b3f41180a42465eb9eec34597bacaf2.shtml.

32 SONG K, TAN X, QIN T, et al. MPNet: masked and permuted pre - training for language understanding [C]. Online: Advances in Neural Information Processing Systems 33 (NeurIPS), 2020.

33 XU M. Text2vec: text to vector toolkit [EB/OL]. [2025 - 10 - 09]. https://github.com/shibing624/text2vec.

34 XIAO S, LIU Z, ZHANG P, et al. C - Pack: packed resources for general Chinese embeddings [C]. New York: The 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2024.

35 DEEPSEEK - AI, LIU A, FENG B, et al. DeepSeek - V3 technical report [EB/OL]. [2025 - 10 - 09]. https://arxiv.org/abs/2412.19437.

36 COMANICI G, BIEBER E, SCHAEKERMANN M, et al. Gemini 2.5: pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities [EB/OL]. [2025 - 10 - 09]. https://arxiv.org/abs/2507.06261.

37 GLM - 4.5 Team. GLM - 4.5: agentic, reasoning, and coding (ARC) foundation models [EB/OL]. [2025 - 10 - 09]. https://arxiv.org/abs/2508.06471.

38 国家卫生健康委员会办公厅. 第一批罕见病目录 [EB/OL]. [2025 - 10 - 09]. https://www.nhc.gov.cn/yzygj/c100068/201806/bd1611850ff14bc8888c149567fe0a55.shtml.

39 GAO L, MA X, LIN J, et al. Precise zero - shot dense retrieval without relevance labels [EB/OL]. [2025 - 10 - 09]. http://arxiv.org/abs/2212.10496.