·综述·

检索增强生成技术在医学人工智能中的应用与前沿探索

金　哲[1]，邹　健[1]，李　逍[1]，吕嘉欣[1]，胡中旭[2]，冯　达[1]

1. 华中科技大学同济医学院药学院（武汉 430030）
2. 华中科技大学机械科学与工程学院（武汉 430074）

【摘要】随着大语言模型（LLM）的蓬勃兴起，深度学习的自然语言生成能力在医学领域展现出了显著价值。然而模型参数的“封闭性”易导致其产生“幻觉”，难以对最新知识作出准确回答，且推理过程缺乏透明度与溯源。检索增强生成（RAG）技术通过在生成阶段主动连接外部文献数据库、知识图谱等信息源，大幅降低了 LLM 对过时训练数据的依赖，并在回答中引入可验证的证据和实时更新的知识。在医学领域，RAG 技术有效解决了文献检索与临床决策支持中的高准确度与可追溯性需求，广泛应用于新药研发、药物警戒、罕见病诊疗等方面。结合强化学习、多模态处理与合规隐私保护等新兴技术，RAG 技术正朝着更加开放、高度定制化的方向演进，为医学信息检索与辅助决策提供了全新的智能化解决方案。

【关键词】大语言模型；检索增强生成；“幻觉”问题；医学信息检索

【中图分类号】TP 391　　**【文献标识码】**A

Application and frontier exploration of retrieval-augmented generation technology in medical artificial intelligence

JIN Zhe[1], ZOU Jian[1], LI Xiao[1], Lyu Jiaxin[1], HU Zhongxu[2], FENG Da[1]

1. School of Pharmacy, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430030, China

2. School of Mechanical Science & Engineering, Huazhong University of Science and Technology, Wuhan 430074, China

Corresponding author: FENG Da, Email: fengda@hust.edu.cn

【Abstract】With the rapid rise of large language models(LLM), the natural language generation capabilities of deep learning have demonstrated significant value in the medical field. However, the "closed nature" of model parameters makes them prone to generating "hallucinations", making it difficult to provide accurate answers to the latest knowledge, and the reasoning process lacks transparency and traceability. Retrieval-augmented generation (RAG) technology addresses these issues by actively connecting external information sources such as document databases and knowledge graphs during the generation process. This significantly reduces the dependence of LLM on outdated training data and introduces verifiable evidence and

DOI: 10.12173/j.issn.1005-0698.202503219

基金项目：国家自然科学基金面上项目（72274071）；湖北省自然科学基金面上项目（2023AFB1067）；国家自然科学基金国际（地区）合作与交流项目（2023YFVA1004）

通信作者：冯达，博士，副教授，博士研究生导师，Email：fengda@hust.edu.cn

real-time knowledge updates into their responses. In the medical field, RAG technology effectively addresses the high-accuracy and traceability requirements of literature retrieval and clinical decision support. It is widely applied in areas such as drug discovery, pharmacovigilance, and the diagnosis and treatment of rare diseases. By integrating emerging technologies such as reinforcement learning, multimodal processing, and compliant privacy protection, RAG technology is evolving towards a more open and highly customizable direction, providing innovative intelligent solutions for medical information retrieval and decision-making support.

【Keywords】Large language model; Retrieval-augmented generation; Hallucination problem; Medical information retrieval

近年来，大语言模型（large language model，LLM）已经在中药学、护理学等医学类专业的深度学习领域研究取得了显著进展[1-4]。然而，随着其在真实生产环境中的进一步部署，模型本身依然面临两方面的短板：其一为“幻觉”现象，当缺乏真实参考时，LLM 往往会凭借语言模式编造并输出并不存在的事实，潜藏较大的误导风险；其二为过时知识难题，模型只能依赖预训练数据，在遇到新出现的文献、前沿成果或专业深度问题时，往往难以发挥作用[5-8]。与此同时，LLM 的内部推理过程常被视为“黑箱”，对于高风险及高精准性的医学应用场景，用户不仅需要获得可靠的事实性回答，还期望查阅明确且可验证的引用来源。为弥补 LLM 在“事实性”与“实时性”上的不足，学术界与工业界普遍认为检索增强生成（retrieval-augmented generation，RAG）是一条切实可行的技术路径[9]。RAG 通过将外部数据库（如文献库、知识图谱等）接入 LLM 的生成过程，使模型在回答时不再囿于内部训练参数，而是能够实时检索并引用多来源的最新信息。凭借“生成前先检索、生成中融入检索内容”的思路，RAG 有效降低了无据可循的编造风险，并为专业领域的知识查询提供了更高的可信度与可溯源性。基于其灵活与可扩展的特质，RAG 目前已逐渐成为提升 LLM 实际应用能力的重要模块[10]。然而目前对于 RAG 在医学领域的研究还处于起步阶段，相关文献数量仅百余篇。基于此，本文对 RAG 的核心概念、技术框架及其在专业领域应用中所面临的瓶颈与挑战进行梳理和总结，重点探讨其在医药场景下的具体应用，包括新药研发与药物警戒、临床决策支持与罕见事件研究等多个方面。最后，结合最新研究动向与实践经验，对 RAG 的未来发展及其在医学信息检索、公共卫生决策中进一步深化的可能性进行了展望，以期为读者提供从理论到实务的一体化参考与启示。

1 RAG技术的起源与发展

RAG 技术由学者 Lewis 等[11]于 2020 年正式提出，初衷是解决 LLM 在处理知识密集型任务时所面临的“幻觉”问题以及专业知识盲区。LLM 尽管拥有庞大的参数规模和对自然语言的深度理解能力，但其所能掌握的知识常常随着训练过程的结束而被“封装”在模型内部，一旦外界出现新的相关知识，模型就难以做出准确回答，甚至会以臆测或编造形式填补“知识空白”[12-13]。这在学术研究或严肃场景中是不可接受的。不仅如此，模型回答中缺乏可靠引用与可追溯来源，也给用户带来极大的不信任感[14-15]。

为了解决 LLM 在“事实性”与“时效性”方面的弱点，RAG 遵循了“让模型随时查资料、再进行回答”的基本思路，在生成文字的过程中主动与外部数据库或知识库交互，根据用户提交的问题从海量文档中筛选出最相关的内容，再将这些精炼信息纳入语言模型的上下文之中，让最终的回答不再仅依赖于模型内置的旧知识。通过这一机制，RAG 在实际应用中能够显著提升回答的可信度与可验证性。由于 LLM 只需基于检索到的实时信息进行“表达”与“推断”，在遇到超出自身训练范围或需要最前沿文献支撑的问题时，也能及时整合最新资料进行权威回答，从而为需要高准确率与高时效性的医学场景提供了重要技术保障。

尽管在理念上看似简单，RAG 要想真正发挥优势并不容易。其核心挑战在于如何在繁杂的外部数据库中高效、精准地检索最相关的文

档，以及如何在回答生成时平滑地将检索信息嵌入语言模型的推理过程之中。如果检索到的文档本身不准确或含有干扰信息，就可能导致模型生成更为混乱的回答；而若检索信息无法与原有问题或上下文对齐，也会增大误用乃至编造的可能性[16]。与此同时，医学领域出于隐私与合规需求，不同级别的数据常存在严格的访问限制和披露要求，RAG 在检索过程中如何进行细粒度的权限管理、在模型生成时如何保证不泄露隐私与敏感内容，都是在实际部署和规模化应用时必须攻克的难题。图 1 为 RAG 技术的简要工作流程。

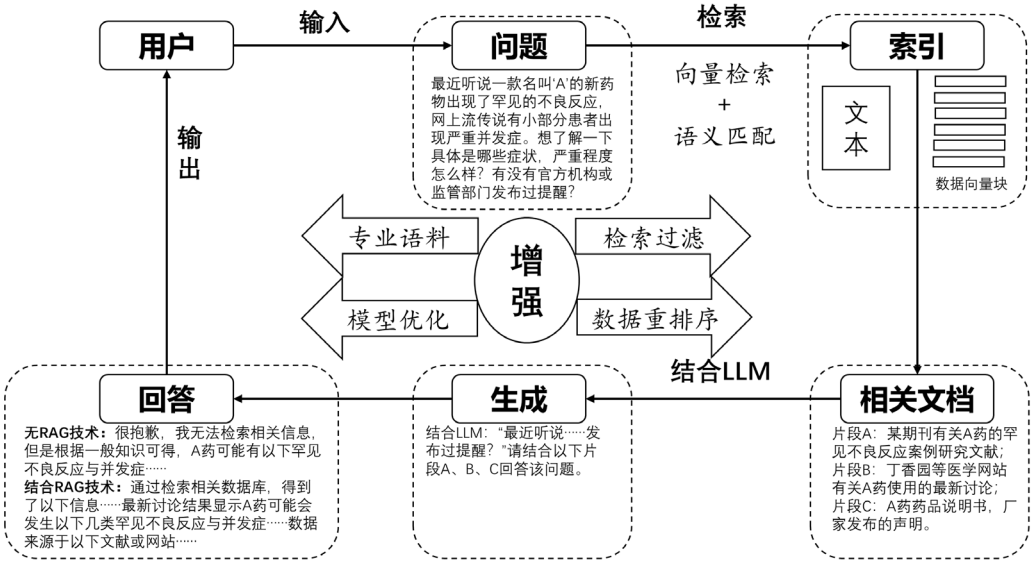

图1 RAG 技术的工作流程

Figure 1. RAG technology workflow

2 RAG 的核心组成与运行机制

从整体框架上看，RAG 技术通常由“检索模块”“生成模块”“增强策略”三大核心组件组成，其中检索和生成是最关键的功能模块，而增强策略则贯穿于预训练、微调和推理全过程，以保证检索数据与模型解答的有效衔接[9-10]。

2.1 检索模块

检索模块的使命是从外部数据库中快速筛选出与用户问题相关度最高的文档或知识片段。为提升检索精准度，RAG 常使用基于向量检索和语义匹配的技术手段，将文档与查询转换为向量后计算相似度。医学场景中，由于存在大量专业术语、同义词、缩写以及多语言文献的情况，检索模块往往会结合专门领域模型来构建更高质量的向量表示，从而更好地识别文本语义与上下文关联。检索过程中还可考虑多轮迭代检索、自适应检索等高级策略，以不断优化返回结果的相关性。

2.2 生成模块

生成模块的功能是对用户的问题以及检索模块返回的外部文档进行综合，最终输出符合自然语言表达的回答或文本内容。相比于纯粹依赖内置参数的生成式模型，RAG 在生成环节可以显式引用检索到的文献或事实，使回答具备更高可信度与可溯源性。一般而言，该模块基于主流的 LLM（如 GPT 系列、T5 等）进行定制化开发，也可与强化学习（reinforcement learning from human feedback，RLHF）相结合，以更好地控制输出风格和提升专业度。对于医学场景，生成模块往往需要配合严格的提示设计，以保证回答符合临床或学术规范，并能够在文本中标明文献出处、数据来源等关键信息。

2.3 增强模块

增强策略可以贯穿于检索与生成的各个阶段，包括预训练期加入专业语料进行知识增强、微调期结合对比学习或人类反馈进行模型优化、推理期对检索结果进行过滤或重排序，以确保最终输出的准确性与易读性。这些措施能够进一步降低医学领域常见的信息噪音与过度简化风险，帮助 RAG 在面对高专业门槛和高准确度需求时依然保持稳健的表现。同时，对于医学隐私与敏

感数据的保护，也往往需要在增强策略中进行相应设计，确保系统在检索或生成时不泄露受保护信息。

3 RAG技术未来发展

结合现阶段文献与工具栈的最新探索，可以发现 RAG 的后续研究逐渐向以下几个方向深化与拓展。首先，在检索阶段加强对多模态信息的支持已成为重要趋势，当前研究不再局限于文本检索，而是引入图像、表格、结构化数据甚至多语言知识库的动态融合。例如，Qilin-Med 项目基于中医医学语料构建多通道检索入口，显著提高了中文医学问答任务中对结构信息和语义细节的掌握能力 [17]。在生成过程中融合 RLHF 机制的研究持续升温，使得模型能在不确定或风险内容出现时主动调整对外部文献的使用方式。例如，在医疗低资源场景中，Das 等 [18] 提出的“双层 RAG 框架”可根据 Reddit 等社交媒体中的非结构数据实时筛选药物不良反应证据，并通过人类标注反馈持续优化生成逻辑。Long 等 [19] 在 Bailicai 系统中构建了适应医疗领域术语的词表对齐机制，确保检索片段与问题语义的精准对应，在 RLHF 微调阶段体现出良好泛化能力。为响应合规与隐私保护需求，RAG 系统越来越重视对外部数据库访问的权限控制与敏感字段的识别屏蔽。在药物安全领域，Li 等 [20] 构建了一个集成 ChatGPT 与 RAG 的临床辅助决策系统，在支持新药重定位任务的同时，严格遵守隐私数据最小暴露原则，并在回答中保留溯源路径与引用说明。

与此同时，研究者还在探索结构增强与攻击防御方面的新技术。Wang 等 [21] 通过一种新颖的以路径为中心的池化操作来引入结构信息。它以即插即用的方式无缝集成到现有的知识图谱 RAG 方法中，从而实现更丰富的结构信息利用。Sui 等[22] 设计了一种应对 RAG 系统新型攻击的扰动机制，动态优化恶意内容，以响应检索到的上下文中的变化。Cheng 等 [23] 通过迭代重构受认知启发的 RAG 系统，提高了 LLM 在解决复杂问题方面的准确性。Jiang 等 [24] 设计了双语基准数据集和轻量级裁判模型，显著减轻了 RAG 中的“幻觉”现象。伴随检索框架、语义索引、微调方法和增强模块的持续迭代，RAG 正成长为下一代 LLM 应用生态中不可或缺的关键组件，不仅可支持复杂的医学知识推理，还在安全性、可信性和实用性层面为 LLM 在医疗等高风险领域的落地提供了坚实支撑，对实际生产环境和多样化应用需求均有深远影响。

4 RAG与其余技术的对比与融合

除 RAG 之外，近年来也涌现出多种旨在增强语言模型事实性与检索能力的技术路线，它们在缓解幻觉问题方面各具优势，具体的技术对比见表 1。

表1 RAG技术与其他技术的对比

Table 1. Comparison of RAG technology with other technologies

技术名称	简要原理	应对幻觉策略	与RAG对比	应用场景
RAG	在生成时对外部文档数据库或知识库进行检索，将检索结果作为上下文融合进回答	回答中引入可验证的证据和实时知识，避免凭空编造	/	知识密集型问答、摘要、多领域检索等；如医疗文献问答、金融报告解读等
API调用增强	训练LLM在必要时调用外部工具/接口，并将返回结果合入生成过程	使用工具提供的准确结果替代模型猜测，如调用搜索查询或计算，大幅减少错误生成	与RAG一样访问外部知识，但RAG检索文本内容，API增强则执行特定功能调用；API方式更灵活，适用于计算、多步决策等	复杂查询、技术支持、多轮对话、数学推理、代码生成等需实时工具支持的场景
知识注入	在预训练或微调阶段将结构化或专业知识显式嵌入模型，使模型权重中携带更多领域信息	增强模型对专业知识的掌握，依靠模型内在知识作答，减少对常识的猜测；但所注入知识时效性受限，可能无法覆盖最新信息	RAG动态检索外部信息，知识注入将知识固化在模型中；适合离线场景，对极高时效性要求的任务不如RAG	医疗诊断、法律问答、专业问答等需丰富领域知识但可离线处理的场景
知识图谱增强生成	在生成过程中结合知识图谱的结构化信息，如实体关系路径或子图查询，将知识图谱中的事实作为生成依据	强制模型参考图谱事实路径，输出符合结构化语义的答案，减少与知识矛盾的回答	与RAG类似利用检索获取知识，不同点是检索对象为图谱关系。知识图谱方法结构化程度高，适用于多跳推理；可与RAG互补	需要精确领域知识关联的任务，如医疗诊断解释、知识图谱问答、多步推理、实体关系问答等

续表1

技术名称	简要原理	应对幻觉策略	与RAG对比	应用场景
检索增强验证	回答生成后进行检索式事实核查：对生成答案再次检索相关文档并进行比对或打分，根据验证结果筛选或修正答案	将检索到的真实文献/数据与模型输出比对，对矛盾或无证据部分进行过滤或提示，从而剔除高概率幻觉内容	RAG在生成时引入证据，检索验证则是生成后的再评估；两者可串联构建“检索－生成－验证”流程	精确性要求极高的领域，如医学建议、法律咨询等需要最终答案可查证的场景
记忆增强LLM	引入外部“记忆”模块，存储历史交互和上下文信息，使模型具有长时记忆能力	持续跟踪对话或任务状态，维持信息一致性；通过外部记忆回忆关键事实，可减少信息丢失和矛盾输出，从而降低幻觉	RAG侧重检索外部知识库，记忆增强侧重跨轮或跨任务保持上下文；两者可结合，如将检索结果写入记忆	多轮对话助手、长期项目跟踪、个性化服务等需记忆用户信息和历史数据的场景

注：API. 应用程序编程接口（application program interface）。

4.1 API调用增强

近年来，研究者提出了让大模型通过调用外部工具来提高准确性的方案。如 Toolformer 模型通过自监督微调，使模型学会何时以及如何调用工具，从而显著提升了零样本任务的表现，该模型在各种任务中调用计算器、问答系统、搜索等工具后，零样本性能大幅改善而不会丧失语言建模能力[25]。类似地，Gorilla 模型针对编写 API 调用进行了训练，其性能超过了 GPT-4，在结合文档检索后能够快速适应文档更新并大幅降低“幻觉”输出[26]。最新工作 ToolLLM 构建了包含 1.6万多种 RESTful API 的指令微调数据集，通过该数据集训练出的模型不仅能执行复杂指令，还能在零样本场景下对新 API 具有很强的泛化能力，其性能已达到可比拟 ChatGPT 的水平[27]。API 调用增强的特点是“以调用代替推理”，能够精准地利用外部工具返回结果，显著降低计算错误和事实性幻觉，适用于需要精确算术或逻辑的场景，与 RAG 强调文档引用的方法形成互补。

4.2 知识注入方法

知识注入（knowledge injection）方法通过在预训练或微调阶段引入领域数据，让模型掌握特定领域知识。常见做法包括指令微调、持续预训练等。有研究表明指令微调能提升模型的总体质量和推理能力，但并不等同于“教给模型新知识”，真正注入新知识通常需要无监督的持续预训练，继续用领域语料训练模型[28]。对通用 LLM 进行领域再训练，可以有效强化其对特定领域事实的记忆，不过这类方法也面临“灾难遗忘”问题，需要使用较低学习率和记忆正则化等技术来缓解。近期研究对比发现，对于需要最新知识的任务，RAG 往往优于仅依赖微调的方法，而知识注入方法更适用于更新频率低、领域较固定的专业场景。

4.3 知识图谱增强生成技术

知识图谱增强生成（knowledge graph-enhanced generation）通过在生成阶段引入图谱结构信息来约束输出，有效减少主观臆断。近期提出的 KAG（knowledge augmented generation）框架结合了知识图谱和向量检索，通过构建互索引、逻辑引导推理等机制，将结构化知识与 RAG 结合，在多跳问答任务上优于传统方法——例如在 HotpotQA 数据集上 F1 分数提升了33.5%[29]。另一项工作 SURGE 则首先检索对话上下文相关的子图，然后通过扰动嵌入和对比学习确保生成的对话严格反映图谱事实，实验结果显示生成的知识性对话质量显著提升[30]。另外也有研究[31]指出将知识图谱作为推理提示或生成约束，可以有效提高事实一致性。因此，在药物相互作用、症状病因链等结构化领域问答中，结合知识图谱的生成模型往往能够精确地跟随图谱路径给出答案。需要强调的是，知识图谱增强方法常与 RAG 相辅相成：如利用知识图谱检索到关联路径后再供语言模型生成答案，可形成更强的“结构化知识检索＋生成”流程[32-33]。

4.4 检索增强验证

部分研究采用了检索增强验证（retrieval augmented verification）机制，即在回答生成后引入独立的事实核查模型对回答进行二次筛选或评分，从而筛除高幻觉概率内容[34]。在实际应用中，一种思路是用检索工具先定位相关证据，再由专门的核查模型判断回答的正确性。如 Provenance 方法针对 RAG 输出计算真实性分数，不借助大型模型而是使用紧凑的 NLI 模型来判别回答与检索到的上下文是否一致，从而实现低延迟的事实检验，其在多种数据集上对幻觉的检测准确率很高[35]。总体来说，检索增强验证的关键是“先生成，

后核实”，它能通过额外的证据支持机制对抗幻觉，但也会增加系统的推理开销。

4.5 记忆增强 LLM

记忆增强 LLM（memory-augmented LLM）通过为模型增加可读写记忆组件，扩展其上下文容量和记忆能力，代表性的 MemoryLLM 模型为 Transformer 每层引入了大量记忆标记，总参数规模达 10 亿，通过特殊的更新和生成机制实现对过去信息的保存，其在 16K Token 长文本上性能优于 Llama-2-7B，但在超过 20K Token 后效果下降[36-37]。后来出现的 M+ 等模型进一步引入分层长时记忆机制，将被丢弃的记忆分片保存到长期存储中，使得整体的知识保留长度扩展到十数万 Token。另一类以记忆池为基础的方法则将对话历史和摘要以文本形式存储为内存块，并使用检索机制根据上下文补充相关内容。还有研究将对话内容存入外部数据库或知识图谱中，如 MemLLM 通过生成函数调用的方式查询知识图谱，实现所谓“三元记忆”。这些记忆增强方法的共同点是减少长期对话或文档处理过程中的前后信息丢失，提高生成的一致性，但其训练复杂度和系统设计也显著增加[38-39]。

5 RAG技术在医药与罕见事件研究领域的应用场景概述

RAG 的思路之所以在医药领域受到格外关注，核心原因在于医药领域对信息的权威性、更新速度和可追溯性有极高要求，任何错误信息或延误都有可能带来严重的后果。RAG 的突出优势在于其外部知识库能够基于最新数据进行实时索引更新，这意味着保持检索模块与医学文献数据库、药品监管平台或公共卫生机构等数据源的联通，模型的回答就能以最新文献或通告为参考，有效避免因知识陈旧而带来的误判。同时在生成回答的过程中引用检索结果的具体出处，也能让临床医生或研究者在需要进一步验证时快速溯源到原始文献，从而兼顾效率与安全性。

5.1 临床决策支持系统

临床决策支持系统（clinical decision support system，CDSS）是一类用于帮助医生做出决策以改善临床表现和患者护理的信息系统[40-41]。在 CDSS 中，RAG 通过将 LLM 和外部检索相结合，为医护人员提供诊断辅助、用药建议以及个性化病情分析时，能随时查询并嵌入医学指南、期刊论文或专家共识等权威资料。如 Remy 等[42]提出的 BioLORD-2023 框架融合 LLM 与临床知识图谱，显著提升了模型在病例总结、患者分流、诊断分类等任务中的表现，尤其在多轮问诊与罕见病辅助识别场景中取得突破。另一项由 Kadhim 等[43]主导的研究展示了如何利用 RAG 增强炎症性肠病研究的临床数据处理，支持 CDSS 在疾病亚型划分和治疗推荐上的新能力。相比传统的基于规则或静态知识库的系统，RAG 对偏罕见的疾病信息、最新药物不良反应报告以及特定患者人群的个案数据也能实现动态检索。原本需要大量人工时间去查询和翻阅海量文献的过程被大幅简化，从而在紧急情况下为患者争取宝贵的诊治机会。

5.2 新药研发与药物警戒

RAG 技术能在宏观和微观层面上为新药研发与药物警戒提供多维度支持。药物研发团队需要持续监测全球范围内的同类药物研发进度、专利布局与临床试验设计方案，这些信息往往散落在各大专利库、数据库以及专业刊物中。RAG 能够在研发人员输入查询后迅速对不同数据库进行检索，将各种版本的进展信息整合到回答里，既能生成对比性强的简要分析，也能在需要时提供原始数据索引。而在药物警戒阶段，当出现罕见不良反应个案或特定人群中发生与已上市药物相关的新问题时，RAG 可以根据搜索到的警示通报、文献报道以及社会媒体发布的信息，将不同来源的数据交叉匹配，呈现出潜在风险的综合画像或趋势变化情况，为药监部门和医疗机构作进一步调查或紧急决策提供快速参考。例如 Choi 等[44]构建的 Malade 框架引入多智能体协作机制，在多轮溯源与语境校验基础上提升了药品风险评估的精度与透明度。该系统还具备政策支持功能，能够生成可溯源的药品安全监测报告，辅助政策制定与公众通告。这些技术进展表明，RAG 在新药研发与药物安全管理中的角色正从“辅助工具”向“认知智能中枢”演化，推动从文献驱动到数据融合驱动的药物创新与监管范式转型。

5.3 罕见事件研究

在罕见事件和突发公共卫生事件中，RAG 技术展现出其在舆情追踪与多源信息整合方面

的独特价值。传统监测系统依赖关键词匹配和静态规则，难以处理来自政府通报、媒体报道、社交平台等异构数据。相比之下，RAG 可实时检索确诊数据、政策调整、研究简报与药品应急信息，并生成带有时间戳与来源说明的结构化内容，显著提升应急响应的效率与准确性，为公共决策与信息发布提供有力支持。Zheng 等[45]的研究进一步证实，RAG 架构可通过接入临床试验注册平台与蛋白靶点数据库，为罕见病研究提供小样本辅助分析能力，尤其在药物重定位与候选基因筛选中表现出较高的精准性。而 Nassiri 等[46]对遗传性罕见代谢疾病的研究则表明，RAG 有助于提升高召回率特征识别能力，在早期症状建模和知识提取方面效果显著。这一思路可拓展至突发公共卫生事件中的社交媒体信号处理与谣言溯源，为“事件级模型监测”提供技术基础。由此可见，RAG 在罕见事件研究中的优势不仅体现在知识检索效率提升上，更体现在其生成内容的可追溯性、时效性与结构化，体现出其作为应急工具的巨大潜力。

5.4 其余应用

RAG 在医学教育、患者自助问答系统乃至医学法规与政策解读方面也颇具潜力。医学生或基层医生常需要查阅海量文献和指南来了解某种疾病的前沿进展或治疗思路，如果采用具备 RAG 功能的智能问答引擎，当他们输入问题时，系统就能即时检索标准教科书、临床试验平台以及学会共识文件等数据源，再由模型将这些信息串联成便于理解与记忆的语段，并附上参照文献的位置索引[47-48]。对患者自助问答而言，RAG 能引用到科普和公共卫生官网的权威信息，还能通过专业词汇与通俗表达间的适配在一定程度上规避医疗误导的风险。Wang 等[49]的系统性综述指出，当前 RAG 驱动的医疗问答系统已在健康宣教、常见症状辨识和预防行为提示等方面取得了可推广的实证基础。医疗政策与法规分析的场景，RAG 通过检索官方文档或案例法规库，能够将已发布的政策要点、最新修订条款以及解读意见一并提取进回答文本中，有助于政策制定者或合规部门在信息繁杂的监管环境里迅速理清要点与对比差异[50]。整体来看，RAG 技术从根本上打破了语言模型以往对内部参数知识的“封闭性”限制，把外部信息当成生成过程的支柱来增强回答的可信度和新鲜度。对于要求实时掌握新信息、深度整合海量专业数据以及追求明确依据溯源的医药领域应用场景而言，这一点显得尤为宝贵。透过结合多类型数据源并嵌入可溯源内容的输出方式，RAG 不仅让生成式人工智能（artificial intelligence，AI）用于关键决策领域的安全与可靠性进一步提升，也为临床医生、研究人员以及公共卫生管理者带来了更加轻便、高效且可持续的信息支持模式。

6 局限性与展望

回顾从 LLM 到 RAG 技术的发展脉络，可以发现它们的结合为医学与罕见事件研究等高风险、高专业度领域带来了极具冲击力的创新模式。RAG 通过开放式地连接外部数据源，在回答生成的过程中注入专业文献、实时统计和权威报道等关键信息，相较于传统只依赖模型参数的生成方式，显著降低了“幻觉”问题的发生率，同时为学术研究和临床决策提供了可追溯的数据参考。从宏观角度看，这种“检索—生成—增强”的框架几乎完全改变了 LLM 在应用场合的角色，它不再只是一个会讲故事的作家，而是兼具知识积累、实时查询和语言表达多重能力的复合式 AI 助手。

尽管 RAG 技术在提升大语言模型事实性与时效性方面展现出巨大潜力，其在医学与罕见事件等高专业、高风险领域的全面落地仍面临诸多挑战。首先是检索精度与效率的平衡问题，若外部文献检索结果本身质量不高，不仅可能引入低可信度甚至无关的信息，还会违背 RAG 增强精准度与可溯源性的初衷；其次，多源信息的冲突与一致性判断依然缺乏稳健机制，不同来源之间可能存在语义不符或结论对立，当前系统尚难以在生成端自动实现交叉验证与冲突消解；第三，医学数据的隐私保护与合规监管压力日益增强，如何在检索与生成阶段对敏感字段进行动态识别、屏蔽与脱敏，防止模型泄露患者隐私，是实现 RAG 临床部署的前提。

进一步而言，RAG 在技术层面还面临若干深层次的结构性局限。其一，医学场景下检索结果碎片化严重，模型往往需要跨文档整合复杂信息，当前 RAG 在多段文本语义融合与一致性保持方面仍显薄弱；其二，医学术语高度复杂且含混，检索模块在处理缩写、同义词、多语言术语

时准确性不足，易导致语义偏移与检索误差；其三，尽管回答中引入了文献片段，RAG 在事实一致性核查机制上仍不健全，可能出现“形式有据、实质错解”的带证幻觉问题；其四，模型生成端对检索内容存在一定“依赖性”，若文献不完整或偏离核心问题，则容易输出“拼贴式”或空洞答案，特别是在低资源领域或冷门疾病场景中表现不佳。此外，RAG 系统构建需整合稳定的索引系统与高性能算力资源，其部署门槛与运维成本高，对中小型医疗机构而言存在可及性问题。最后，尽管 RAG 强化了文献引用能力，其整体仍属“黑箱式”推理，缺乏可解释性与透明推理路径，难以满足医疗决策对过程信任与验证的实际需求。

在 RAG 技术的未来展望层面，与 RLHF 结合的动态检索机制、多模态数据与文本检索的融合，以及对生成环节进行可解释性强化等思路，都是促使 RAG 更深层次发展的潜在方向。例如，通过强化学习让模型根据回答过程中的不确定度如语言概率、上下文歧义，自适应地发起新一轮检索，或在回答后进行自动的“事实核查”，可以减少幻觉和不准确引用；再如，医学诊断往往需要结合影像、基因组测序等多模态信息，RAG 如果能调取影像检索、影像标注等深度学习模型协同工作，就有望在临床场景里形成更完备的辅助诊断体系。考虑到医学和公共事件研究常常要面对高度动态的外部环境，RAG 只要能在数据安全与合规层面进一步得到保障，必然会成为未来医疗信息化和社会突发事件管理的重要支撑底座。

综合而言，RAG 在现阶段已经为专业领域信息检索和精准回答奠定了良好基础，并不断朝着模块化、多模态和可持续更新的方向前进。随着医学界对时效性与高可信度的需求不断升级，这种“检索 – 生成”融合的模式将进一步扩散到药物研发、健康管理、医疗教育以及公共卫生危机应对等更多细分场景，推动医学与研究工作者完成从海量信息到决策洞察的跨越，也为罕见事件的快速响应和深入洞察提供了一条可行且前景广阔的技术路径。

利益冲突声明：作者声明本研究不存在任何经济或非经济利益冲突。

参考文献

1 Zhao WX, Zhou K, Li J, et al. A survey of large language models[EB/OL]. (2025-03-11) [2025-08-01]. https://doi.org/10.48550/arXiv.2303.18223.

2 李欣桐，马素芬，张丰聪，等. 中医药领域大语言模型的研究进展与应用前景 [J]. 南京中医药大学学报，2024, 40(12): 1393-1403. [Li XT, Ma SF, Zhang FC, et al. Research progress and application prospect of large language model in the traditional Chinese medicine[J]. Journal of Nanjing University of Traditional Chinese Medicine, 2024, 40(12): 1393-1403.] DOI: 10.14148/j.issn.1672-0482.2024.1393.

3 施呈昊，屠馨怡，史佳伟，等. 大语言模型临床实践应用范围综述 [J]. 医学信息学杂志，2024, 45(9): 19-26. [Shi CH, Tu XY, Shi JW, et al. A scoping review of the application of large language models in clinical practice[J]. Journal of Medical Informatics, 2024, 45(9): 19-26.] DOI: 10.3969/j.issn.1673-6036.2024.09.003.

4 刘泓泽，王耀国，唐圣晟，等. 医学大语言模型的应用现状与发展趋势研究 [J]. 中国数字医学，2024, 19(8): 1-7, 13. [Liu HZ, Wang YG, Tang SS, et al. Research on the current application and development trend of medical large language model[J]. China Digital Medicine, 2024, 19(8): 1-7, 13.] DOI: 10.3969/j.issn.1673-7571.2024.08.001.

5 熊灏. 融合知识图谱的大语言模型幻觉问题研究 [D]. 广州：广东工业大学，2024.

6 Ji Z, Lee N, Frieske R, et al. Survey of hallucination in natural language generation[J]. ACM Comput Surv, 2023, 55(12): 1-38. DOI: 10.1145/3571730.

7 Kandpal N, Deng H, Roberts A, et al. Large language models struggle to learn long-tail knowledge[EB/OL]. (2023-07-27) [2025-08-01]. https://doi.org/10.48550/arXiv.2211.08411.

8 Zhang Y, Li Y, Cui L, et al. Siren's song in the AI ocean: a survey on hallucination in large language models[EB/OL]. (2023-09-24) [2025-08-01]. https://doi.org/10.48550/arXiv.2309.01219.

9 Gao Y, Xiong Y, Gao X, et al. Retrieval-augmented generation for large language models: a survey[EB/OL]. (2024-03-27) [2025-08-01]. https://doi.org/10.48550/arXiv.2312.10997.

10 Cheng M, Luo Y, Ouyang J, et al. A survey on knowledge-oriented retrieval-augmented generation[EB/OL]. (2025-03-17) [2025-08-01]. https://doi.org/10.48550/arXiv.2503.10677.

11 Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks[EB/OL]. (2021-04-12) [2025-08-01]. https://doi.org/10.48550/arXiv.2005.11401.

12 Xu H, Yang Z, Zhu Z, et al. Delusions of large language models[EB/OL]. (2025-03-09) [2025-08-01]. https://doi.org/10.48550/arXiv.2503.06709.

13 Barros S. I think, therefore I hallucinate: minds, machines, and the art of being wrong[EB/OL]. (202503-04) [2025-08-01]. https://doi.org/10.48550/arXiv.2503.05806.

14 Bo JY, Wan S, Anderson A. To rely or not to rely? evaluating interventions for appropriate reliance on large language models[EB/OL]. (2025-03-09) [2025-08-01]. https://doi.

org/10.48550/arXiv.2412.15584.

15 Thorne S. Understanding and evaluating trust in generative AI and large language models for spreadsheets[EBOL]. (2024-12-18) [2025-08-01]. https://doi.org/10.48550/arXiv.2412.14062.

16 Levy S, Mazor N, Shalmon L, et al. More documents, same length: isolating the challenge of multiple documents in RAG[EB/OL]. (2025-03-06) [2025-08-01]. https://doi.org/10.48550/arXiv.2503.04388.

17 Ye Q, Liu J, Chong D, et al. Qilin-med: multi-stage knowledge injection advanced medical large language model[EB/OL]. (2024-04-17) [2025-08-01]. https://doi.org/10.48550/arXiv.2310.09089.

18 Das S, Ge Y, Guo Y, et al. Two-layer retrieval-augmented generation framework for low-resource medical question answering using reddit data: proof-of-concept study[J]. J Med Internet Res, 2025, 27: e66220. DOI: 10.2196/66220.

19 Long C, Liu Y, Ouyang C, et al. Bailicai: a domain-optimized retrieval-augmented generation framework for medical aApplications[EB/OL]. (2024-07-24) [2025-08-01]. https://doi.org/10.48550/arXiv.2407.21055.

20 Li D, Jiang N, Huang K, et al. From questions to clinical recommendations: large language models driving evidence-based clinical decision making[EB/OL]. (2025-05-15) [2025-08-01]. https://doi.org/10.48550/arXiv.2505.10282.

21 Wang H, Feng Y, Xie X, et al. Path pooling: train-free structure enhancement for efficient knowledge graph retrieval-augmented generation[EB/OL]. (2025-05-27) [2025-08-01]. https://doi.org/10.48550/arXiv.2503.05203.

22 Sui R. CtrlRAG: black-box adversarial attacks based on masked language models in retrieval-augmented language generation[EB/OL]. (2025-03-10) [2025-08-01]. https://doi.org/10.48550/arXiv.2503.06950.

23 Cheng Y, Zhao Y, Zhu J, et al. Human cognition inspired RAG with knowledge graph for complex problem solving[EB/OL]. (2025-03-09) [2025-08-01]. https://doi.org/10.48550/arXiv.2503.06567.

24 Jiang Z, Sun M, Zhang Z, et al. Bi'an: a bilingual benchmark and model for hallucination detection in retrieval-augmented generation[EB/OL]. (2025-02-26) [2025-08-01]. https://doi.org/10.48550/arXiv.2502.19209.

25 Schick T, Dwivedi-Yu J, Dessì R, et al. Toolformer: language models can teach themselves to use tools[EB/OL]. (2023-02-09) [2025-08-01]. https://doi.org/10.48550/arXiv.2302.04761.

26 Patil SG, Zhang T, Wang X, et al. Gorilla: large language model connected with massive APIs[EB/OL]. (2023-03-24) [2025-08-01]. https://arxiv.org/abs/2305.15334.

27 Qin Y, Liang S, Ye Y, et al. ToolLLM: facilitating large language models to master 16000+ real-world APIs[EB/OL]. (2023-10-03) [2025-08-01]. https://doi.org/10.48550/arXiv.2307.16789.

28 Ovadia O, Brief M, Mishaeli M, et al. Fine-tuning or retrieval? comparing knowledge injection in LLMs[EB/OL]. (2024-01-30) [2025-08-01]. https://doi.org/10.48550/arXiv.2312.05934.

29 Liang L, Sun M, Gui Z, et al. KAG: boosting LLMs in professional domains via knowledge augmented generation[EB/OL]. (2024-09-26) [2025-08-01]. https://doi.org/10.48550/arXiv.2409.13731.

30 Kang M, Kwak JM, Baek J, et al. Knowledge graph-augmented language models for knowledge-grounded dialogue generation[EB/OL]. (2023-03-30) [2025-08-01]. https://doi.org/10.48550/arXiv.2305.18846.

31 Lavrinovics E, Biswas R, Bjerva J, et al. Knowledge graphs, large language models, and hallucinations: an NLP perspective[EB/OL]. (2024-09-21) [2025-08-01]. https://doi.org/10.48550/arXiv.2411.14258.

32 Hamza A, Abdullah, Ahn YH, et al. LLaVA needs more knowledge: retrieval augmented natural language generation with knowledge graph for explaining thoracic pathologies[EB/OL]. (2024-12-19) [2025-08-01]. https://doi.org/10.48550/arXiv.2410.04749.

33 Zhao X, Liu S, Yang S, et al. MedRAG: enhancing retrieval-augmented generation with knowledge graph-elicited reasoning for healthcare copilot[EB/OL]. (2025-06-27) [2025-08-01]. https://doi.org/10.48550/arXiv.2502.04413.

34 Jiang J, Chen J, Li J, et al. RAG-Star: enhancing deliberative reasoning with retrieval augmented verification and refinement[EB/OL]. (2024-12-17) [2025-08-01]. https://doi.org/10.48550/arXiv.2412.12881.

35 Sankararaman H, Yasin MN, Sorensen T, et al. Provenance: a light-weight fact-checker for retrieval augmented LLM generation output[EB/OL]. (2024-11-01) [2025-08-01]. https://doi.org/10.48550/arXiv.2411.01022.

36 Wang Y, Krotov D, Hu Y, et al. M+: extending MemoryLLM with scalable long-term memory[EB/OL]. (2025-05-30) [2025-08-01]. https://doi.org/10.48550/arXiv.2502.00592.

37 Wang Y, Gao Y, Chen X, et al. MEMORYLLM: towards self-updatable large language models[EB/OL]. (2024-05-26) [2025-08-01]. https://doi.org/10.48550/arXiv.2402.04624.

38 Li S, Zhang K, Liu Q, et al. MindBridge: scalable and cross-model knowledge editing via memory-augmented modality[EB/OL]. (2025-03-04) [2025-08-01]. https://doi.org/10.48550/arXiv.2503.02701.

39 Xu M, Liang G, Chen K, et al. Memory-augmented query reconstruction for LLM-based knowledge graph reasoning[EB/OL]. (2025-03-07) [2025-08-01]. https://doi.org/10.48550/arXiv.2503.05193.

40 Muhiyaddin R, Abd-Alrazaq AA, Househ M, et al. The impact of clinical decision support systems (CDSS) on physicians: a scoping review[J]. Stud Health Technol Inform, 2020, 272: 470-473. DOI: 10.3233/SHTI200597.

41 Sutton RT, Pincock D, Baumgart DC, et al. An overview of clinical decision support systems: benefits, risks, and strategies for success[J]. NPJ Digit Med, 2020, 3(1): 17. DOI: 10.1038/s41746-020-0221-y.

42 Remy F, Demuynck K, Demeester T. BioLORD-2023: semantic textual representations fusing LLM and clinical knowledge graph

insights[EB/OL]. (2023-11-27) [2025-08-01]. https://doi.org/10.48550/arXiv.2311.16075.

43 Kadhim A Z, Green Z, Nazari I, et al. Application of generative artificial intelligence to utilise unstructured clinical data for acceleration of inflammatory bowel disease research[EB/OL]. (2025-03-07) [2025-08-01]. https://doi.org/10.1101/2025.03.07.25323569.

44 Choi J, Palumbo N, Chalasani P, et al. MALADE: orchestration of LLM-powered agents with retrieval augmented generation for pharmacovigilance[EB/OL]. (2024-08-03) [2025-08-01]. https://doi.org/10.48550/arXiv.2408.01869.

45 Zheng Y, Koh HY, Yang M, et al. Large language models in drug discovery and development: from disease mechanisms to clinical trials[EB/OL]. (2024-09-06) [2025-08-01]. https://doi.org/10.48550/arXiv.2409.04481.

46 Nassiri K, Akhloufi MA. Recent advances in large language models for healthcare[J]. BioMedInformatics, 2024, 4(2): 1097-1143. https://doi.org/10.3390/biomedinformatics4020062.

47 王之 . 多模态中文大语言模型医疗问答研究与应用 [D]. 武汉 : 湖北工业大学 , 2024.

48 张玉铭 , 李红岩 , 郎许锋 , 等 . 基于检索增强生成技术的中医药问答大语言模型的构建 [J]. 南京中医药大学学报 , 2024, 40(12): 1375-1382. [Zhang YM, Li HY, Lang XF, et al. Construction of traditional Chinese medicine question-answering large language model based on retrieval-augmented generation technology[J]. Journal of Nanjing University of Traditional Chinese Medicine, 2024, 40(12): 1375-1382.] DOI: 10.14148/j.issn.1672-0482.2024.1375.

49 Wang L, Wan Z, Ni C, et al. A systematic review of ChatGPT and other conversational large language models in healthcare[J/OL]. medRxiv[Preprint], [2025-08-01]. DOI: 10.1101/2024.04.26.24306390.

50 梁泽宇 . 多选项医疗问答模型研究与问答系统实现 [D]. 北京 : 北京交通大学 , 2023.

收稿日期：2025 年 03 月 31 日 修回日期：2025 年 08 月 04 日
本文编辑：冼静怡 周璐敏