

检索增强生成 (RAG) 赋能医学人工智能综述

柳 汛^{*}, 刘述民

江西理工大学软件工程学院 江西南昌

【摘要】本文聚焦检索增强生成 (Retrieval-Augmented Generation, RAG) 技术在医学人工智能领域的核心价值与发展脉络, 系统剖析其技术演进历程、核心架构设计及多场景应用实践, 并构建多维度评估体系, 深入探讨当前面临的挑战与未来优化方向。通过整合分析 RAG 在临床决策支持、医学科研教育及患者健康管理等场景的应用案例, 验证了该技术在缓解大型语言模型 (LLMs) 知识滞后、幻觉生成及不可解释性等关键问题上的显著效能, 其通过动态检索外部权威知识库的方式, 大幅提升了医学内容生成的事实准确性、可追溯性与临床可靠性。研究表明, RAG 技术是推动医学人工智能可信化落地的核心路径, 但在多模态信息融合、医疗数据隐私保护、知识库动态更新及计算成本优化等方面仍需持续突破。本文为 RAG 技术在高风险医学场景的安全应用提供了理论支撑与实践参考, 助力智能医疗生态的高质量建设。

【关键词】大语言模型; 检索增强生成; 医疗 AI; 临床决策; 评估

【收稿日期】2025 年 8 月 15 日 **【出刊日期】**2025 年 9 月 18 日 **【DOI】**10.12208/j.aics.20250054

A review of retrieval enhancement generation (RAG) empowering medical artificial intelligence

Xun Liu^{*}, Shumin Liu

School of Software Engineering, Jiangxi University of Science and Technology, Nanchang, Jiangxi

【Abstract】This paper focuses on the core value and development context of Retrieval-Augmented Generation (RAG) technology in the field of medical artificial intelligence. It systematically analyzes its technical evolution, core architecture design, and multi-scenario application practices, constructs a multi-dimensional evaluation system, and deeply explores the current challenges and future optimization directions. By integrating and analyzing RAG's application cases in clinical decision support, medical research and education, and patient health management, this study verifies the significant effectiveness of the technology in alleviating key problems of Large Language Models (LLMs) such as knowledge latency, hallucination generation, and poor explainability. Through dynamically retrieving external authoritative knowledge bases, RAG greatly improves the factual accuracy, traceability, and clinical reliability of medical content generation. The research shows that RAG technology is a core path to promote the trustworthy deployment of medical artificial intelligence, but continuous breakthroughs are still needed in aspects such as multimodal information fusion, medical data privacy protection, dynamic knowledge base update, and computational cost optimization. This paper provides theoretical support and practical reference for the safe application of RAG technology in high-risk medical scenarios, and helps the high-quality construction of the intelligent healthcare ecosystem.

【Keywords】Large language models; Retrieval-augmented generation; Medical AI; Clinical decision-making; Evaluation

1 引言

生成式人工智能凭借其卓越的自然语言理解与文本生成能力, 正深刻重塑医学领域的研究范式与实践模式。大型语言模型 (LLMs) 在医学文献分析、临床

决策辅助、患者健康教育等场景中展现出巨大应用潜力, 为提升医疗服务效率、扩大医疗资源可及性提供了全新可能。然而, 医学领域对信息准确性与临床安全性的极致要求, 使得 LLMs 的固有局限日益凸显: 其一,

^{*}通讯作者: 柳汛

“幻觉”现象频发, 生成内容看似逻辑连贯却与医学事实相悖, 在临床场景中可能引发误诊误治风险; 其二, 医学知识更新迭代迅速, 静态训练数据导致模型存在明显知识滞后, 难以同步最新临床指南与科研成果; 其三, 决策过程的“黑箱”特性缺乏透明性与可追溯性, 严重影响医务人员对模型输出结果的信任度。

为破解上述难题, 检索增强生成 (RAG) 技术应运而生。该技术创新性地将检索机制与生成模型相结合, 在内容生成前主动从外部权威知识库中检索相关信息, 以“开卷考试”式的生成模式为模型提供实时、准确的上下文支撑。这种独特机制不仅有效提升了输出内容的事实可靠性, 更通过来源引用强化了结果的可追溯性, 使其在医学这一零容错率的高风险领域具备了不可替代的应用价值。

本文以 RAG 技术在医学领域的应用为核心, 构建了“技术基础-架构设计-应用实践-评估体系-挑战展望”的完整研究框架。首先梳理 RAG 技术的演进历程与理论基础; 其次解析医学 RAG 系统的核心架构及各模块功能; 随后全面展现其在临床、科研、教育、健康管理等多场景的应用价值; 进而建立科学的多维度评估体系; 最后剖析当前技术落地面临的挑战, 并展望未来发展趋势。本文旨在为医学研究者与从业者提供全面、系统的认知框架, 推动生成式人工智能在医疗领域的安全、负责任应用。

2 RAG 技术的演进历程

检索增强生成 (RAG) 技术的发展轨迹, 清晰映射了自然语言处理领域从规则驱动到智能生成的演进逻辑, 其核心是逐步实现“封闭静态记忆”向“开放动态检索”的模式转型。

2.1 早期问答系统阶段

早期问答系统以规则和模板为核心支撑, 采用“检索 - 匹配 - 生成”的简单流程: 先通过关键词匹配从静态知识库中检索相关文档, 再依据固定格式拼接生成答案。此类系统在处理简单明确的查询时具有一定可行性, 但在语义理解深度、上下文整合能力及生成灵活性方面存在显著局限, 难以应对复杂医学问题的多步骤推理需求。同时, 静态知识库无法及时更新, 导致系统难以追踪快速发展的医学知识, 严重限制了其在专业医疗场景的应用价值。

2.2 大型语言模型驱动阶段

Transformer 架构的出现推动自然语言处理技术实现跨越式发展, 大型预训练语言模型 (LLMs) 在文本生成的流畅性与多样性上取得突破性进展。然而, 纯

参数化生成模式的固有缺陷也随之暴露: 缺乏外部知识约束导致“幻觉”现象频发, 静态训练数据使其无法捕捉最新医学进展, 这些问题在高风险医学场景中被无限放大, 极大降低了模型的可信度与可用性^[2,3]。

2.3 检索增强生成成熟阶段

2020 年, Patrick Lewis 等人提出端到端联合训练的 RAG 技术框架, 将可微分检索器与生成器有机融合, 使模型在生成过程中能够动态访问外部知识库, 实现了从“闭卷”到“开卷”的本质转变^[1]。这一创新设计突破了模型内部记忆的局限, 通过实时引用外部权威信息, 有效提升了事实准确性并降低了幻觉发生率。

2.4 医学领域专业化阶段

RAG 在通用领域的成功迅速延伸至医学这一知识密集型高风险领域。针对医学知识专业性强、更新快、对证据可追溯性要求高的特点, 医学 RAG 技术持续迭代升级, 形成了多源知识整合、结构化推理、多跳检索及上下文优化等核心能力。其能够从 PubMed、临床指南、电子病历等多源数据库中精准检索信息, 并在生成答案时附带来源引用, 完美契合了医学领域的特殊需求, 推动 RAG 技术向专业化、可解释化及动态知识融合方向深度发展^[4]。

3 医学 RAG 系统的核心架构

医学 RAG 系统以“高可靠性、强可追溯性、临床适用性”为核心目标^[1], 其总体架构由用户输入解析、检索模块、生成模块、知识库管理与输出管理构成完整闭环^[5]。系统整体技术框架如图 1 所示。该系统在生成医学内容前主动检索权威外部知识, 如医学指南、PubMed 文献、电子病历 (EMR) 和医学知识图谱, 实现了实时信息更新与精准知识增强。与仅依赖内部参数记忆的传统 LLMs 相比, 医学 RAG 通过重排过滤、引用标注与合规治理确保输出的事实准确性和源头透明性, 从而更适配高风险的临床应用场景^[6]。

3.1 检索模块: 精准信息获取的基础

检索模块是医学 RAG 系统的核心基础, 承担着从多源异构医学知识库中高效获取相关信息的任务。其数据来源广泛, 包括 PubMed 等学术文献数据库、临床诊疗指南、电子病历系统、医学教科书及科研论文等多元资源。为实现高效检索, 所有数据需经过严格的清洗、标准化处理及向量化转换, 构建结构化检索索引^[7]。

在检索策略上, 系统突破传统关键词匹配的局限, 融合语义理解与多跳推理技术, 能够精准应对复杂临床查询, 尤其擅长处理 PICO (人群、干预、对照、结

局) 结构等专业医学问题。检索结果需经过重排优化与去噪处理, 通过算法模型筛选出最相关、最权威的信息

片段, 确保输入生成模块的内容既准确又高效, 为后续生成环节奠定坚实基础^[8]。

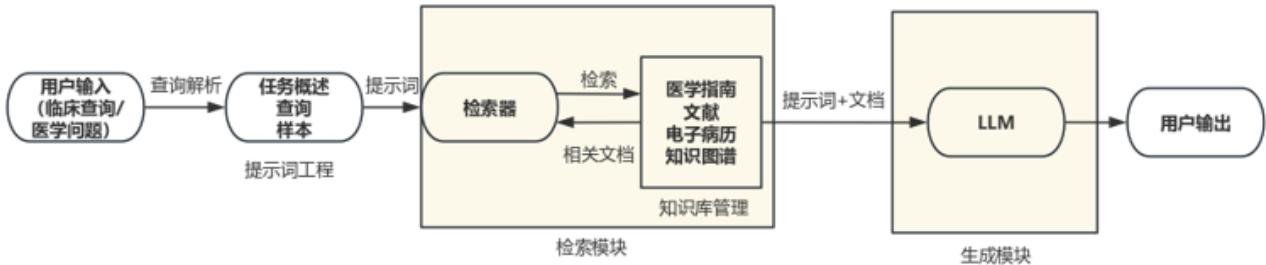


图 1 RAG 框架

3.2 生成模块: 专业内容生成的核心

生成模块负责将检索获取的权威信息与模型自身知识融合, 生成符合医学逻辑、结构规范的自然语言内容。模块在严格的提示词约束下运行, 确保输出内容符合医学专业规范, 可直接为临床诊疗、科研分析等场景提供参考。

为强化可追溯性, 每条生成内容均附带来源引用标注, 方便使用者回溯原始文献或指南依据。针对不同应用场景, 生成模块采用模块化设计策略: 在诊断支持场景中侧重推理逻辑的严谨性, 在治疗方案制定场景中突出证据的权威性, 在科研综述场景中强调信息的全面性, 在教育问答场景中注重解释的通俗性^[9]。此外, 通过融合医学知识图谱技术, 模块能够清晰呈现推理路径, 进一步提升系统的可解释性与临床安全性。

3.3 知识库构建与管理: 系统可靠运行的保障

知识库的质量直接决定医学 RAG 系统的输出可靠性, 其构建与管理需兼顾完整性、时效性与安全性。在数据整合方面, 系统需实现结构化与非结构化数据的高效融合, 妥善处理医学术语标准化、古文与现代语言转换 (如中医药领域) 等特殊问题, 确保知识覆盖的全面性。

在更新机制上, 系统采用增量更新与版本管理模式, 实时追踪医学指南修订、科研成果发布等动态信息, 确保知识库内容与最新医学进展保持同步。同时, 严格落实访问控制与隐私保护措施, 遵循 HIPAA、GDPR 等医疗信息安全法规, 对电子病历等敏感数据进行加密处理, 防范信息泄露风险, 为系统的合规运行提供保障。

3.4 系统优化与扩展: 功能升级的路径

医学 RAG 系统具备高度的灵活性与可扩展性, 可通过针对性优化适配不同医学场景的个性化需求。

在技术优化方面, Graph RAG 技术通过构建疾病、症状、药物及治疗方案之间的关联图谱, 实现了可解释的逻辑推理; 多模态扩展技术则融合医学影像、基因组数据及实验室检验结果, 为精准诊疗提供更全面的决策支持。

在应用扩展方面, 系统可深度嵌入临床工作流、科研管理系统、医学教育平台及健康管理工具等各类场景, 实现从基础信息查询到复杂决策支持的功能跃升。通过持续的技术迭代与场景适配, 医学 RAG 系统正从单一的问答工具, 逐步发展为支撑医疗全流程的智能辅助平台。

4 医学 RAG 技术的多场景应用图景

RAG 技术凭借其动态检索、精准生成、可追溯验证的核心优势, 在医学领域构建了覆盖临床实践、科研教育、健康管理的全方位应用生态, 为医疗行业的数字化转型提供了强大动力^[1]。

4.1 临床决策支持: 提升诊疗精准度与安全性

在临床实践场景中, RAG 技术成为智能临床决策支持 (CDS) 系统的核心引擎, 为医务人员提供实时、权威的决策参考。系统能够快速检索最新临床指南、循证医学证据及相似病例数据, 辅助医生进行鉴别诊断、制定个性化治疗方案及评估疾病预后^[8,10], 有效降低误诊漏诊风险。

与传统 LLMs 相比, RAG 驱动的 CDS 系统通过明确标注证据来源, 大幅提升了诊疗建议的可验证性与可信度。在甲状腺疾病管理、肿瘤精准治疗等具体场景中, 系统能够精准匹配患者病情与最新研究成果, 为治疗方案优化提供数据支撑^[11]; 在多组学分析领域, 通过整合基因组、转录组等多维度数据, 系统可揭示生物标志物与药物响应的关联规律, 为复杂疾病的个体化治疗提供科学依据^[12]。

4.2 医学科研与教育：加速知识创新与传承

RAG 技术显著革新了医学科研与教育的传统模式, 为知识发现与人才培养提供了高效工具。在科研场景中, 研究者可借助 RAG 技术快速检索跨学科、跨领域的相关文献, 自动归纳研究热点、追踪临床试验进展、整合分析研究数据, 有效降低文献调研成本, 为科研选题、方案设计与成果总结提供有力支持^[14]。

在教育场景中, RAG 驱动的智能导师系统能够根据学生的知识水平与提问需求, 从权威教材、学术文献中提取核心信息, 生成结构化、层次化的解释内容, 帮助学生建立循证思维与临床推理能力。针对中医药等特殊领域, 系统通过融合知识图谱与结构化检索技术, 有效解决术语混杂、古文解读困难等问题, 推动传统医学知识的标准化传承与普及化传播。

4.3 患者服务与健康管理：优化医疗资源可及性

RAG 技术为患者服务与健康管理提供了创新解决方案, 有效提升了医疗信息的可及性与可靠性。通过构建基于权威知识库的智能问答平台, 患者能够便捷获取个性化的健康指导、疾病科普、用药咨询等服务, 有效规避网络虚假医疗信息带来的健康风险。

在医疗资源匮乏的偏远地区, RAG 技术可通过远程健康咨询、异常情况监测等功能, 为当地居民提供持续的健康管理支持; 在老年群体护理中, 系统通过简化交互流程、优化内容呈现形式, 帮助老年人便捷获取健康服务, 改善生活质量。此外, RAG 技术的广泛应用有助于平衡不同地区、不同群体间的医疗资源差异, 推动医疗服务的公平化发展。

5 医学 RAG 系统的多维度评估体系

随着 RAG 技术在医学领域的规模化应用, 建立科学、系统的评估体系成为保障其临床可靠性、安全性和可解释性的关键。医学 RAG 系统的评估需兼顾技术性能与临床应用价值, 从检索模块、生成模块及系统整体表现三个维度构建多指标评估框架^[9]。

5.1 检索模块评估：聚焦信息获取质量

检索模块的评估核心是信息获取的相关性与有效性, 主要采用量化指标进行客观衡量。核心评估指标包括召回率 (Recall@K)、精确率 (Precision@K)、平均倒排排名 (MRR) 及归一化折损累积增益 (NDCG) 等, 分别用于评估检索结果的覆盖完整性、精准匹配度、排序合理性及整体质量^[14,15]。

针对医学场景的特殊性, 评估还需关注模块的复杂问题处理能力, 包括多跳推理能力、上下文理解能力及专业术语匹配精度, 确保在处理 PICO 结构查询、

罕见病诊断咨询等复杂场景时, 能够精准获取关键信息, 为生成模块提供高质量的输入支撑^[16]。

5.2 生成模块评估：兼顾准确性与适用性

生成模块的评估需兼顾形式质量与内容价值。在形式层面, 采用 BLEU、ROUGE、BERTScore 等传统指标评估文本的流畅性、连贯性与语言规范性; 在内容层面, 重点评估事实准确性与上下文一致性, 核心指标包括忠实度 (Factuality)、检索上下文依赖性等^[8], 确保生成内容与检索信息、医学事实高度吻合。

针对医学领域的专业需求, 采用 MIRAGE、MedHallu 等专用基准数据集, 测试模块在医学知识问答、临床案例分析等场景的表现。此外, 结合医学专业特点, 评估内容还需包括医学术语使用准确性、临床逻辑严谨性等专业指标, 确保输出内容符合医疗行业规范^[17]。

5.3 高风险临床场景专项评估

在高风险临床应用场景中, 评估需聚焦安全性与可靠性核心诉求。采用 RadGraph-F1 指标评估生成文本中解剖结构、病变描述的准确性; 通过 FactScore 指标衡量输出内容与检索上下文的一致性; 借助 RAGAS 等自我评估框架, 整合忠实度、语义相关性与不确定性估计, 实现幻觉风险的自动监测与预警^[18,19]。

同时, 评估体系需涵盖临床相关性、患者安全性、伦理合规性等维度, 通过模拟真实临床场景的测试案例, 验证系统在诊断建议、治疗方案推荐等关键环节的表现, 确保系统输出不会对患者健康造成潜在风险^[20]。

5.4 专家审查与多源校验：补充量化评估局限

单纯的量化指标难以全面反映医学 RAG 系统的实际应用价值, 人工审查与多源校验是必要补充。组建由临床专家、数据科学家、信息检索工程师构成的跨学科评估团队, 从医学专业性、临床实用性、逻辑合理性等维度进行定性评估^[21]。

通过多中心测试、真实场景试点应用等方式, 收集医务人员、科研人员及患者的反馈意见, 全面评估系统的用户体验与实际使用效果。针对多模态 RAG 系统、实时更新型知识库等特殊类型, 还需扩展评估维度, 涵盖多源数据整合准确性、更新时效性等专项指标, 为系统优化提供全面依据。

6 医学 RAG 技术的挑战与未来展望

尽管 RAG 技术在医学领域展现出巨大应用潜力, 但在高风险、高规范要求的医疗场景中, 其规模化落地仍面临多重挑战, 同时也孕育着广阔的发展空间。

6.1 当前核心挑战

6.1.1 数据隐私与安全风险

医学 RAG 系统需访问大量包含患者敏感信息的电子病历、影像数据及临床研究资料, 数据隐私保护面临严峻挑战。如何在满足系统检索需求的前提下, 严格遵循 HIPAA、GDPR 等医疗信息安全法规, 防范数据泄露、非法访问等风险, 成为系统合规运行的关键前提。

6.1.2 知识库构建与更新难题

医学知识具有更新快、来源杂、格式不一的特点, 知识库的构建需解决多源数据融合、术语标准化、古文与现代语言转换等问题。同时, 如何建立高效动态更新机制, 实现医学指南修订、科研成果发布与知识库内容的实时同步, 避免知识滞后, 是保障系统输出可靠性的核心挑战。

6.1.3 偏见与公平性问题

若检索数据存在种族、性别、地区等维度的偏差, RAG 系统可能在诊疗建议、健康指导中放大这种不平等, 影响医疗服务的公平性。如何识别并消除数据偏见, 确保系统对不同群体提供一致、公正的智能支持, 是伦理层面的重要挑战。

6.1.4 临床工作流整合障碍

现有医疗系统多为封闭架构, RAG 系统难以与其实现无缝衔接。如何在不干扰现有临床工作流程、不增加医务人员额外负担的前提下, 实现 RAG 功能的深度嵌入, 确保生成结果能够便捷融入诊疗决策过程, 是影响技术落地效果的关键因素。

6.2 未来发展方向

6.2.1 多模态信息融合技术突破

未来医学 RAG 系统将实现文本、影像、基因组数据、实验室检验结果等多模态信息的深度融合。通过整合计算机视觉、自然语言处理、生物信息学等多领域技术, 构建全方位的患者健康画像, 为精准诊疗提供更全面的决策支持。

6.2.2 智能体与自主推理能力升级

引入智能体 (Agent) 技术, 赋予 RAG 系统主动推理、多步骤决策的能力。系统能够自主识别临床问题、设计检索策略、整合分析信息、生成决策建议, 并根据反馈结果持续优化, 形成闭环智能辅助流程, 进一步提升临床应用价值。

6.2.3 隐私计算与安全技术融合

采用联邦学习、差分隐私、同态加密等先进隐私计算技术, 在保护数据隐私的前提下实现多中心数据共享与联合训练。通过技术创新平衡数据利用与隐私保护的矛盾, 为知识库的丰富性与安全性提供双重保障。

6.2.4 标准化与规范化发展

推动医学 RAG 技术的标准化建设, 包括数据格式标准、检索与生成规范、评估指标体系、临床应用指南等。通过建立统一的行业标准, 规范技术研发与应用流程, 降低技术推广成本, 促进智能医疗生态的有序发展。

参考文献

- [1] Patrick L, Ethan P, Aleksandra P, et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks[C]. Conference on Neural Information Processing Systems, 2020, 33: 9459-9474.
- [2] Lei H, Weijiang Y, Weitao M, et al. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions[J]. Computing Research Repository, 2023, abs/2311.05232.
- [3] Adam T K, Ofir N, Santosh S V, et al. Why Language Models Hallucinate[J]. arXiv preprint arXiv, 2025, 2509(04664).
- [4] Michael Klesel, H. Felix Wittmann. Retrieval-Augmented Generation (RAG)[J]. Business & Information Systems Engineering, 2025, 67(4): 551-561.
- [5] Xuejiao Z, Siyan L, Su-Yin Y, et al. MedRAG: Enhancing Retrieval-augmented Generation with Knowledge Graph-Elicited Reasoning for Healthcare Copilot[J]. Computing Research Repository, 2025.
- [6] Sudeshna D, Yao G, Yuting G, et al. Two-Layer Retrieval-Augmented Generation Framework for Low-Resource Medical Question Answering Using Reddit Data: Proof-of-Concept Study[J]. Journal of Medical Internet Research, 2025, 27.
- [7] Guangzhi X, Qiao J, Xiao W, et al. Improving Retrieval-Augmented Generation in Medicine with Iterative Follow-up Questions[C]. Pacific Symposium on Biocomputing, 2025, 30: 199-214.
- [8] Yucheng S, Shaochen X, Tianze Y, et al. MKRAG: Medical Knowledge Retrieval Augmented Generation for Medical Question Answering[C]. AMIA Annual Symposium Proceedings, 2024: 1011-1020.
- [9] Nghia T N, Chien V N, Franck D, et al. Comprehensive and Practical Evaluation of Retrieval-Augmented Generation Systems for Medical Question Answering[J]. Computing

- Research Repository, 2024, abs/2411.09213.
- [10] Junde W, Jiayuan Z, Yunli Q, et al. Medical Graph RAG: Towards Safe Medical Large Language Model Via Graph Retrieval-Augmented Generation[J]. Computing Research Repository, 2024, abs/2408.04187.
- [11] Guangzhi X, Qiao J, Xiao W, et al. Improving Retrieval-Augmented Generation in Medicine with Iterative Follow-up Questions[C]. Pacific Symposium on Biocomputing, 2025, 30: 199-214.
- [12] Sudeshna D, Yao G, Yuting G, et al. Two-Layer Retrieval-Augmented Generation Framework for Low-Resource Medical Question Answering Using Reddit Data: Proof-of-Concept Study[J]. Journal of Medical Internet Research, 2025, 27.
- [13] Kaiwen Z, Kwonjoon L, Teruhisa M, et al. ViCor: Bridging Visual Understanding and Commonsense Reasoning with Large Language Models[J]. Computing Research Repository, 2024.
- [14] Shahul E, Jithin J, Luis E, et al. RAGAS: Automated Evaluation of Retrieval Augmented Generation[C]. Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, 2024: 150-158.
- [15] YuHe K, Liyuan J, Kabilan E, et al. Development and Testing of Retrieval Augmented Generation in Large Language Models - A Case Study Report[J]. CoRR, 2024, abs/2402.01733.
- [16] Guangzhi X, Qiao J, Zhiyong L, et al. Benchmarking Retrieval-Augmented Generation for Medicine[C]. Annual Meeting of the Association for Computational Linguistics, 2024: 6233-6251.
- [17] Mario C, Lorenzo B, Valentin C, et al. Retrieval Augmented Generation Evaluation for Health Documents[J]. alphaxiv, 2025.
- [18] Guangzhi X, Qiao J, Xiao W, et al. Improving Retrieval-Augmented Generation in Medicine with Iterative Follow-up Questions[C]. Pacific Symposium on Biocomputing, 2025, 30: 199-214.
- [19] Juraj V, Annika D, Mai N, et al. Improving Reliability and Explainability of Medical Question Answering Through Atomic Fact Checking in Retrieval-Augmented LLMs[J]. alphaxiv, 2025.
- [20] Yuelyu Ji, Hang Zhang, Yanshan Wang. Bias Evaluation and Mitigation in Retrieval-Augmented Medical Question Answering Systems[J]. alphaxiv, 2025.
- [21] Sichu L, Linhai Z, Hongyu Z, et al. RGAR: Recurrence Generation-augmented Retrieval for Factual-aware Medical Question Answering[J]. Computing Research Repository, 2025, abs/2502.13361.

版权声明: ©2025 作者与开放获取期刊研究中心 (OAJRC) 所有。 本文章按照知识共享署名许可条款发表。

<http://creativecommons.org/licenses/by/4.0/>



OPEN ACCESS