专家论坛

大语言模型辅助医学系统综述：方法、发展方向和应用

黄衍楠[1]，桑浩然[1]，刘 宇[2]，马连韬[1]，朱英豪[1]
（1.北京大学，北京 100871；2.牛津大学，英国牛津 OX2 0JB）

马连韬，北京大学计算机软件与理论博士，北大软件工程国家工程研究中心研究型助理教授/助理研究员。长期从事医信交叉研究工作，重点研究开发具备高维数据处理能力的医疗领域专用软件平台，以可解释深度学习、大语言模型赋能临床工作与医学科研。主持国际/国家级科研课题4项；以项目骨干承担国家自然科学基金重点项目、国家重点研发计划项目等6项。主导研发医疗数据智能分析框架与诊疗辅助平台，以首页封面文章在*Cell Patterns*发表“腹膜透析患者可解释预后预测”研究；在人工智能与数据挖掘领域国际高水平会议期刊发表医信交叉研究20篇；研发“智慧诊疗辅助系统”获CCTV-1专题报道及国家卫健委《健康报》关注。

摘要 随着生物医学文献数量的爆炸式增长，传统的基于关键词匹配的检索方法日益难以满足临床与科研实践中对效率与精准性的双重需求。近年来，以ChatGPT和DeepSeek为代表的大语言模型，凭借其强大的自然语言处理能力，在医学系统综述领域展现出显著的应用潜力。然而，其固有的“幻觉”问题与知识更新滞后等挑战限制了其直接应用的可靠性。本文系统介绍了当前缓解大语言模型“幻觉”的6类核心技术路径，重点解释检索增强生成技术的原理与应用优势，并在综合梳理系统综述任务中的22篇代表性研究的技术特点与应用场景后，进一步指出基于“证据等级的结构化理解与生成”的大语言模型是未来的重要发展方向之一。本文旨在为医学研究人员与临床从业者提供系统性的参考，助力其科学高效地利用大语言模型提升医学文献信息处理效率与循证医疗决策质量。

关键词 大语言模型；医学系统综述；检索增强生成；提示工程
中图分类号:G434；R-4；TP18 **文献标志码**:A **文章编号**:1005-930X(2025)03-0323-09
DOI：10.16190/j.cnki.45-1211/r.2025.03.001

Empowering medical systematic reviews with large language models: methods, development directions, and applications

HUANG Yannan[1], SANG Haoran[1], LIU Yu[2], MA Liantao[1], ZHU Yinghao[1]. (1. Peking University, Beijing 100871, China; 2. University of Oxford, Oxford OX2 0JB, United Kingdom)

Abstract With the exponential growth of biomedical literature, traditional keyword-based retrieval methods are increasingly inadequate for meeting the dual demands of efficiency and precision in clinical and research contexts. In recent years, large language models (LLMs), exemplified by ChatGPT and DeepSeek, have demonstrated significant potential in supporting medical systematic reviews due to their powerful natural language processing capabilities. However, its inherent challenges such as the “hallucination” problem and lagging knowledge update limit the reliability of its direct application. This paper systematically introduces six core technical ap-

[基金项目] 国家自然科学基金资助项目(No. 62402017)
[通信作者] 马连韬，E-mail：malt@pku.edu.cn；朱英豪，E-mail：yhzhu99@gmail.com
[收稿日期] 2025-06-05

proaches currently used to mitigate hallucinations in LLMs, with a particular focus on explaining the principles and application advantages of retrieval-augmented generation (RAG). After comprehensively reviewing the technical characteristics and application scenarios of 22 representative studies in the context of systematic reviews, the paper further identifies LLMs capable of "structured understanding and generation based on levels of evidence" as one of the key future directions. The goal is to provide systematic guidance for medical researchers and clinical practitioners, helping them make scientific and efficient use of LLMs to enhance the efficiency of biomedical literature processing and the quality of evidence-based medical decision-making.

Keywords　large language models; medical systematic reviews; retrieval-augmented generation; prompt engineering

在医学研究与循证临床实践中，系统综述是获取高质量科学证据、支持临床决策和推动知识更新的核心环节。一个标准的系统综述流程包含文献初步检索、相关性筛选、数据提取、偏倚风险评估、证据合成和初步报告撰写，涉及大量文本阅读与重复性动作，费时费力。传统的基于关键词匹配与布尔逻辑运算的检索方法，尽管在当前实践中仍发挥着重要作用，但在语义理解的深度、表达歧义的处理以及信息过滤的精度等方面存在固有的局限性，难以充分满足多语言、多语境及多任务场景下的复杂检索需求[1]。与此同时，随着PubMed等权威数据库年均新增文献量突破150万篇[2]，研究人员和临床医生正面临前所未有的信息过载压力。

近年来，基于Transformer架构的大语言模型在自然语言处理领域取得了突破性进展，已展现出卓越的文本理解、语言生成及语义关联建模能力，并逐渐发展成为处理复杂语言任务的通用基础模型[3]，为医学系统综述的流程优化提供了新的途径，有望推动医学文献的检索、筛选与综合分析过程从传统的人工主导模式向自动化、智能化方向转变[4-5]。然而，大语言模型固有的"幻觉"（hallucination）问题[6]在系统综述这一高度重视事实准确性的任务中，可能成为关键瓶颈。基于医学领域数据的有监督微调（domain-specific supervised fine-tuning, SFT）、强化学习从人类反馈中学习（reinforcement learning from human feedback, RLHF）[7]和检索增强生成（retrieval-augmented generation, RAG）[8]等当前缓解大语言模型幻觉的主流方法在系统综述任务中表现各异。

本文旨在系统性地梳理与评述当前大语言模型在系统综述任务中的核心技术方法，分析其面临的挑战与未来发展方向，并针对医学研究者与临床医师提出实践建议，期望为相关领域的专业人士提供一份结构清晰、论证充分的技术参考。

1　方　法

1.1　大语言模型简述

大语言模型通过在超大规模语料库上进行自监督预训练，利用多层Transformer网络结构（一种基于自注意力机制的深度学习模型）[9]学习复杂的语言统计规律与深层语义结构，从而获得了强大的自然语言理解与文本生成能力。其核心机制在于通过上下文建模来预测序列中下一个最可能出现的词元（token）或句子，这种"下一个词元预测"（next token prediction）的范式，是当前主流大语言模型运作的基础[10]。正是这种预测能力，使得大语言模型能够模拟人类的语言表达和一定的推理能力，进而实现对人类自然语言表达的精确仿真。在交互式应用中，大语言模型可被抽象地视为一个函数$M(\cdot)$，该函数接收输入的文本序列，并生成一个在概率分布上最为相关的输出文本序列。这种能力使其能够表现出类似人类的理解、问答乃至文献阅读与摘要生成等复杂行为。

将通用大语言模型直接应用于严谨的医学文献检索任务时，面临最为突出的问题之一便是"幻觉"现象。幻觉是指模型可能生成表面流畅、逻辑自洽但与事实不符甚至完全错误的信息，例如编造不存在的文献、虚构研究结果或生成错误的引文。这种现象的根源在于：首先，大语言模型的知识主

要来源于其预训练数据，当用户查询的内容超出了其参数化知识的边界[11]（即模型“不知道”或其知识已过时）时，它仍会基于其学到的模式尝试生成一个“看起来合理”的答案[12]；其次，大语言模型在生成文本时通常会引入一定的随机采样机制（如temperature sampling）[13]以增加输出的多样性，但这有时也会导致偏离事实的陈述。举例来说，如果让大模型基于一篇非常经典、频繁出现在其训练语料中的参考文献进行回答，它大概率能够准确输出。然而，预训练语料的更新通常较为缓慢且成本高昂（训练一次大型大语言模型的成本可达数百万美元）。对于一些新兴的、不那么知名或在训练语料中出现频率较低的“长尾”文献，模型很可能无法准确回忆相关信息，从而开始生成表面合理但实则不准确的信息。

1.2 技术分类与讨论

理论上，从模型编码到模型训练，当前主流的能够缓解大语言模型的“幻觉”方法有6种（图1）。然而，其中需进行模型训练的方法（黑色实线）的成本效益往往不呈正比，因为它们技术门槛较高，同时，伴随着高昂的训练或人工标注成本，对于需要知识频繁更新、动态演进的医学领域而言，单纯依赖模型内部参数的更新是不现实的。

具体到文献系统综述这类对事实准确性和信息时效性要求极高的应用场景，针对幻觉问题，一种更为经济、可靠且高效的策略是为大语言模型“外挂”一个高可信度的、可实时更新的文献知识库（例如PubMed等权威医学数据库）。通过这种方式，知识库可以低成本、高频率地更新，而无需对庞大的大语言模型本身进行重新训练或微调，确保了知识的鲜活性和准确性，即RAG机制。RAG机制的基本流程图，见图2。

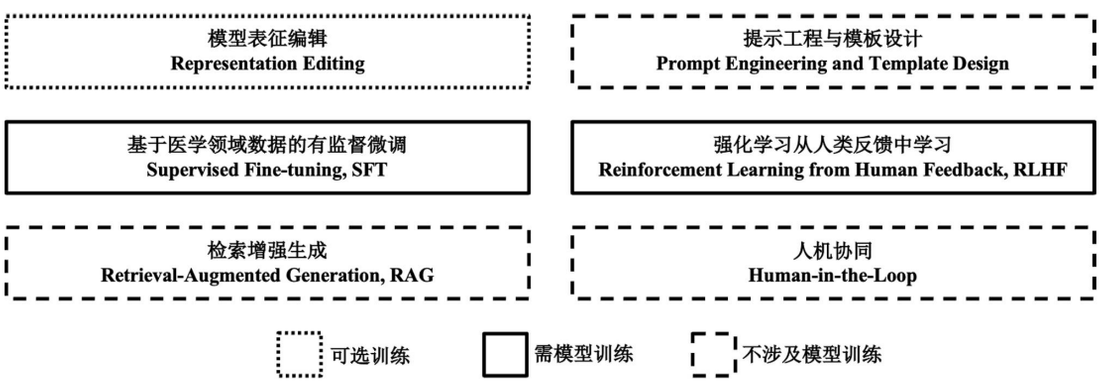

图1 方法汇总：应对大语言模型“幻觉”问题

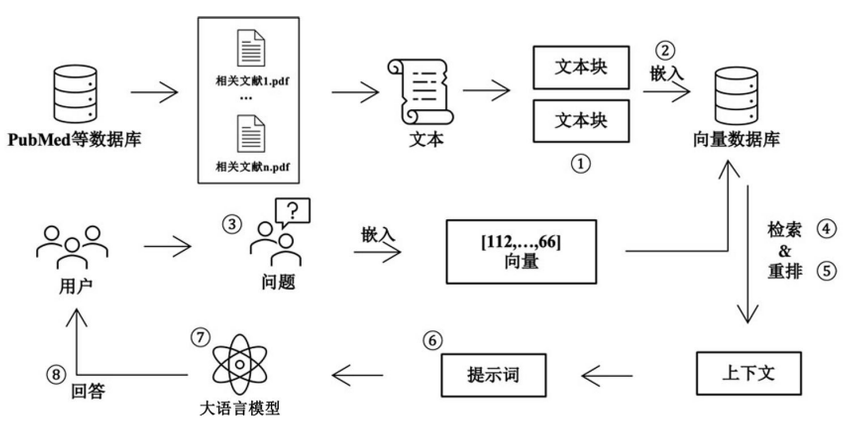

图2 RAG机制的基本流程示意图

如图2所示，RAG的基本工作流程如下：

（1）外部知识库转为向量数据库（图2中序号①、②）：系统首先将外部文献知识库（如PubMed）中的文献分割成合适的文本块（chunk），通过编码器模型（encoder，如Sentence-BERT[14]）转换为文献向量存入向量数据库（如FAISS[15]）中，这类数据库

能够高效地进行大规模向量的相似性搜索；(2)用户查询与向量化(图2中序号③)：用户以自然语言形式提出问题/查询，然后使用相同的或兼容的编码器模型处理成查询向量；(3)信息检索与重排(图2中序号④、⑤)：在向量数据库中计算查询向量与文献向量之间的语义相似度(常用余弦相似度或点积)，检索与用户查询语义最相关的若干文献片段，并对结果顺序按照一定的标准进行重排；(4)上下文构建与提示词增强(图2中序号⑥)：将这些检索到的相关文献片段作为“上下文”(context)信息，通常会将这些片段按照相关性排序后，拼接起来形成一段文本，将原始查询和检索到的上下文信息整合，构建成一个增强的提示(augmented prompt)，这个提示通常会明确指示大语言模型基于提供的上下文来回答问题，例如：“根据以下信息：〔检索到的上下文〕，请回答问题：〔原始查询〕”；(5)结果交付(图2中序号⑦)：大语言模型基于提供的上下文信息和原始查询，生成最终的、有据可循的答案或综述内容，并可选择性地提供所引用文献片段的来源链接，增强透明度和可验证性。

通过上述RAG流程，大模型的回答主要依据是实时从可靠外部知识库中检索到的具体文献内容，而非仅仅依赖其可能存在偏差或过时的内部参数化知识。关键技术简要说明见表1。

表1　关键技术简要说明

技术名称	简要说明
模型表征编辑[16-17]	表征编辑是一类在不重新训练模型的前提下，直接修改语言模型内部表示以更新其知识内容的技术，其通过精确定位并干预模型参数或激活状态中承载特定知识的部分，从而实现对单点事实(如人物、地点、医学常识等)的“微创式”更新或纠正
基于医学领域数据的SFT	指在通用大语言模型的基础上，利用特定领域(例如医学领域)的小规模、高质量语料进行二次训练，旨在提升模型对该领域专业术语、特定语境及专门任务的理解与生成性能
RAG[8]	核心思想是在大语言模型生成回答之前，先从一个可信的、通常是领域相关的外部可更新知识库中检索出与用户查询最相关的信息片段，然后将这些检索到的信息作为上下文提供给大语言模型，引导其基于具体、可溯源的外部证据生成时效性更强、可靠性更高的回答
提示工程与模板设计	关注于通过精心构造结构化、明确且具有引导性的输入指令，以指导大语言模型在特定任务中生成更为准确、聚焦和一致的输出
RLHF[7]	一种将人类偏好融入语言模型训练的技术，其通过让人工标注者对模型生成结果进行排序，训练一个奖励模型，再用该奖励信号指导语言模型的策略优化
人机协同	一种在实际应用中结合大语言模型与人工协作的策略，研究者借助模型处理流程中重复性强、劳动密集的任务(如文献筛选、数据提取)，同时在人类研究者把控关键决策节点与结果验证的基础上，提升整体效率、准确性与可控性

1.3　文献收集与归纳

本文通过Google Scholar数据库检索策略(例如，组合使用关键词“large language model”OR 大语言模型 OR GPT AND “medical literature”OR “systematic review” OR “evidence synthesis” AND retrieval OR search)并辅以文献追溯，限定时间范围为近期(主要为2024—2025年上半年发表或公开的预印本)，筛选出22篇具有代表性相关研究文献，并将其划分为3类研究类型：创新方法、实证研究、综述，并根据其关注的任务流程和技术路径进行归类汇总。

创新方法指研究者围绕某类文献处理任务，提出或实现了具备方法创新或系统构建特征的技术框架；实证研究强调基于现有工具或技术方案(GPT-4)在真实任务中的实用性验证与性能评估，研究者更多从“任务可达性”与“性能瓶颈”角度出发，通过精心设计提示工程、模版策略或人机协作流程，对文献筛选、数据提取等任务进行实测分析；综述文章聚焦于对特定子领域的已有研究进行系统性回顾与比较，总结大语言模型在系统评价、证据合成、自动化检索等任务中的方法演进、应用特点与挑战边界。

表2提供了研究全景，以便读者从任务目标、技术路径与应用场景维度获取关键研究线索。该表

汇总了当前在医学文献检索与系统评价任务中应用大语言模型的代表性研究，按研究类型分为创新方法、实证研究和综述文章三类。在技术标签上，主要方法涵盖RAG（检索增强生成）[18-19]、人机协同流程设计[29]、提示工程与模版设计[25-28, 30-32]以及领域适应与微调[24]，显示出当前研究在增强信息获取能力、提高任务适应性、优化用户交互等方面的多样探索。

表2　关键文献梳理汇总表

研究类型	参考文献	技术标签	应用场景分类
创新方法	[18]	RAG	全文检索与问答
创新方法	[19]	RAG	标题摘要/全文筛选与综合
创新方法	[20]	人机协同流程	全文数据提取
创新方法	[21]	人机协同流程	标题摘要筛选
创新方法	[22]	人机协同流程	证据综合与可视化
创新方法	[23]	人机协同流程	文献过滤
创新方法	[24]	领域适应与微调	学术文献搜索
实证研究	[25]	提示工程与模版设计	标题/摘要筛选
实证研究	[26]	提示工程与模版设计	文献引用生成
实证研究	[27]	提示工程与模版设计	标题摘要/全文筛选
实证研究	[28]	提示工程与模版设计	标题/摘要筛选
实证研究	[29]	人机协同流程	全文筛选
实证研究	[30]	提示工程与模版设计	系统评价流程优化
实证研究	[31]	提示工程与模版设计	标题/摘要筛选
实证研究	[32]	提示工程与模版设计	标题/摘要筛选
综述文章	[4]	不适用	不适用
综述文章	[33]	不适用	不适用
综述文章	[5]	不适用	不适用
综述文章	[34]	不适用	不适用
综述文章	[35]	不适用	不适用
综述文章	[36]	不适用	不适用
综述文章	[37]	不适用	不适用

1.3.1　任务评估指标　在本文纳入分析的研究中，多项工作[25, 27-28, 31]在评估大语言模型筛选医学文献的能力时，普遍采用灵敏度与特异性作为核心量化评价指标。灵敏度用于衡量模型正确识别“应纳入文献”的能力，即在所有实际应纳入的研究中，模型能够成功检出的比例，这直接关系到文献检索的全面性及后续系统评价的证据完整性；特异性则评估模型准确排除“不应纳入文献”的效能，即在所有实际不应纳入的研究中，模型能够正确识别并予以排除的比例，这反映了模型在节省人力成本、减少后续人工复筛工作量方面的潜力。综述文章[29]提到的多项近期工作的灵敏度与特异性在[89%，95%]区间内。

此外，Delgado-Chaves等[30]在其研究中引入了F1分数（F1-score）与马修斯相关系数（Matthews correlation coefficient, MCC）等综合性评价指标，用以更全面地评估模型在不同类别样本（纳入/排除）可能不均衡情况下的整体分类性能。Susnjak等[19]采用事实一致性评分和事实提取和验证（fact extraction and verification, FEVER）数据集的验证支持率作为评估指标：CGS分数越高，表明模型生成的陈述越是基于所提供的文献证据，产生的“幻觉”内容越少；FEVER支持率则量化了模型生成的陈述能够被检索文献内容所证实或支持的比例。

1.3.2　应用场景分类　Lieberum等[4]系统地回顾并归纳了大语言模型在系统综述各阶段的应用潜力：

(1)文献搜索与检索策略制定阶段:大语言模型可辅助研究人员生成更优化的搜索查询语句,或作为一种潜在的交互式实时搜索工具(尽管目前仍面临生成虚假引用等局限性);(2)文献筛选与选择阶段:包括对大量文献的标题和摘要进行初步相关性筛选,以及对筛选后的文献全文进行进一步的合格性评估;(3)数据提取与信息抽取阶段:大语言模型被用于从研究论文(包括正文文本、表格乃至图表内容)中自动提取关键信息要素,如研究人群特征、干预措施详情、主要及次要结局指标等。此外,研究还显示大语言模型可辅助生成文献摘要、进行初步的知识合成以构建综述草稿,甚至在研究初期辅助研究者构思新的研究问题或辅助撰写学术文本的部分章节。

在文献检索与筛选阶段,大语言模型通常能够展现出较高的召回率,但在特异性方面,不同模型及应用场景下的表现差异显著。Delgado-Chaves等[30]的实证研究表明,大语言模型可将筛选特定研究所需的人工工作量减少33%~93%,但其性能对纳入/排除标准的清晰度和复杂性较为敏感,在该研究中模型的MCC仅为0.349。

在数据提取与信息抽取阶段,大语言模型已显示出在保障数据提取完整性方面的潜力,并达到了一定程度的准确性。Ye等[20]提出的混合工作流程借助大语言模型从全文中自动提取关键信息,其关键信息要素的误分类率相较于纯人工提取降低了1.5%,提取完整率高达98%。

部分前瞻性研究已开始探索将大语言模型应用于更高层级的证据综合与可视化输出阶段。Wang等[22]提出的TrialMind系统能够自动生成森林图用于Meta分析结果的可视化,在62.5%~100%的测试案例中,其生成的图表被医学专家认为质量等同甚至优于人工制作,显示出较高的实用性与用户可接受度。Joos等[23]开发的LLMSurvey工具通过结合多模型共识机制,有效地将文献过滤的召回率提升至98.8%,同时将重要文献的误排率(错误排除率)控制在0.11%以内,整体表现在特定指标上显著优于传统人工方法设定的误差阈值。

2　发展方向

上述研究结果共同表明,大语言模型正逐步展现出在医学证据生成与加工全流程中多层次、多阶段深度嵌入的潜力。然而,其输出结果的可靠性、保障事实准确性与信息来源可追溯性,仍是未来技术迭代与优化的关键方向[4, 26, 32]。

领域自适应微调成本高昂;提示工程与模板设计不能满足文献系统综述领域“即时性”的需求;RAG自身亦存在检索质量的瓶颈效应[38],其最终生成内容的质量高度依赖于检索模块所检索信息的准确性与全面性,一旦检索环节未能有效捕获高质量的相关证据,即便生成器本身能力强大,也可能导致输出结果存在片面性甚至产生误导性信息。另外,现有在处理复杂医学语料时,RAG有时会表现出“肤浅概括”(shallow generalization)的倾向[39],即仅仅对输入文本进行表层的转述或重组,而缺乏深度的语义理解、多源信息的有效融合以及必要的批判性评估能力。

同时,未来医学文献处理系统不应仅停留在表层内容的语义提取层面,应进一步向“基于证据等级的结构化理解与生成”演进,从而为临床决策提供更具层次化、可解释性及可验证性的信息支撑。在通用文献检索场景下,尽管大语言模型能够高效地检索大量相关文献,但其通常难以自主且准确地评估文献的固有学术质量,例如区分不同研究设计类型的证据强度、识别参考文献的影响力(如期刊影响因子、分区)等。在循证医学的系统综述任务中的研究质量评估与偏倚风险环节,如何构建具备高级“证据辨别力”的大语言模型,也是一个待突破的方向。现有模型往往难以有效区分一篇高质量的系统评价或Meta分析与一篇证据等级相对较低的病例报告或专家意见,这可能导致其综合生成的结论在可信度与权威性上存在固有偏差。

3　实　践

为促进医学研究者与临床医生更高效、更可靠地利用大语言模型,本节将结合当前研究进展与实践经验,提出若干关键的技术应用建议与伦理考量。首先,提示工程是影响大语言模型输出质量的

核心环节之一，尤其在处理定义明确的结构化任务（例如，文献筛选、特定信息提取）时，精心设计的提示对模型行为具有显著的引导与约束作用。建议医学科研人员学习并掌握基础的提示设计原则与技巧，如清晰定义任务目标、为模型设定特定角色（例如，“假设你是一位经验丰富的循证医学专家”）、规定结构化的输出格式、提供充足且相关的上下文信息，并可尝试应用“思维链”（chain-of-thought, CoT）[40]等高级提示策略以激发模型的推理能力。合理构建提示词不仅有助于提升生成结果的准确性与任务相关性，还能在一定程度上降低“幻觉”现象的发生概率。

其次，使用者应始终保持批判性思维，强调对大语言模型输出内容（尤其是关键信息和结论）的独立核查与源头文献验证，将其定位为高效的“智能助手”，而非人类智能的“完全替代者”[34]。在当前技术发展阶段，医学研究人员应将大语言模型合理定位为“智能研究助理”或“辅助决策支持工具”，主要用于加速处理重复性高、劳动密集型任务（例如，文献摘要的初步筛选、结构化数据的初步提取等）。而在涉及研究方案设计、纳排标准制定、偏倚风险评估、最终结论解释等关键环节，仍应坚持人工主导与严格复核。通过构建“人机协同、优势互补”的智能化工作流程，即有效结合人类专家的深度领域知识、批判性思维与大语言模型强大的信息处理能力，有望在保障研究质量与结论可靠性的前提下，显著提升科研效率。

最后，随着大语言模型技术的快速演进，相关应用也在医学领域逐渐渗透[41－46]。王祖恒等[41]对人工智能在医疗保健中的应用实施路径进行了合理展望。同时，相关的伦理规范、操作标准及报告指南亦在不断发展与完善中。医学研究者与临床医生应主动关注并积极遵循这些新兴的行业共识与指南。例如，在文献检索公平性方面，Foley等[34]的研究指出，当前主流的人工智能文献检索工具主要依赖于英语语料库进行训练，且大语言模型在处理源自全球南方地区的事实性陈述时，其准确性可能显著低于处理全球北方地区的陈述，这可能导致全球南方国家的研究成果在学术检索和知识整合过程中被进一步边缘化，从而加剧全球知识获取与传播的不平等。随着大语言模型技术在医学领域的应用日益广泛，相关的伦理规范、操作标准和报告指南（例如，针对大语言模型在医学预测模型开发与报告中应用的TRIPOD-LLM声明[47]等）正陆续制定和发布。医学研究者和临床医生应主动学习并积极遵循相关标准，确保在自身的科研工作与临床实践中，能够安全、合规且负责任地应用大语言模型及相关人工智能技术。

4 展　望

本文综述了大语言模型在医学系统综述任务中的主要应用路径与研究进展，涵盖RAG、提示工程、领域适应、人机协同等关键技术。当前研究表明，大语言模型在提高流程效率、缓解人力负担方面已展现出实际可行性，特别是在文献筛选与信息提取等高重复性任务中具备显著优势。

随着大语言模型基础能力的持续提升，其在医学系统综述中的应用前景正日益清晰。未来研究可向加强对文献结构与证据等级的语义建模能力，使模型不仅“读懂”文献，更能“识别”其科学价值这一方向探索。更长期看，大语言模型有望重塑系统综述的方法学范式，推动其从以人工为中心的线性流程，转向以模型驱动、专家监督的智能化流程。在保障透明性、可重复性与伦理合规的前提下，大语言模型将成为未来医学知识更新与临床证据生成的重要基础设施。

参考文献：

[1] SCHOEFEGGER K, TAMMET T, GRANITZER M. A survey on socio-semantic information retrieval[J]. Computer science review, 2013, 8: 25-46.

[2] MEDLINE PubMed production statistics[EB/OL]. (2024-04-30) [2025-05-02]. https://www. nlm. nih. gov/bsd/medline_pubmed_production_stats.html.

[3] BOMMASANI R, HUDSON D A, ADELI E, et al. On the opportunities and risks of foundation models[EB/OL]. (2022-07-12) [2025-05-02]. https://arxiv. org/abs/2108. 07258.

[4] LIEBERUM J L, TOEWS M, METZENDORF M I, et al. Large language models for conducting systematic reviews: on the rise, but not yet ready for use-a scoping review[J]. Journal of clinical epidemiology, 2025, 181:

111746.

[5] SCHERBAKOV D, HUBIG N, JANSARI V, et al. The emergence of large language models as tools in literature reviews: a large language model-assisted systematic review[J]. Journal of the American medical informatics association, 2025, 32(6): 1071-1086.

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

[7] OUYANG L, WU J, XU J, et al. Training language models to follow instructions with human feedback[EB/OL]. (2022-05-04) [2025-05-02]. https://arxiv. org/abs/2203. 02155.

[8] LEWIS P, PEREZ E, PIKTUS A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks [EB/OL]. (2021-04-12) [2025-05-02]. https://arxiv. org/abs/2005.11401v4.

[9] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30: 5998-6008.

[10] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners[EB/OL]. (2020-07-22) [2025-05-02]. https://arxiv.org/abs/2005.14165.

[11] REN R Y, WANG Y H, QU Y Q, et al. Investigating the factual knowledge boundary of large language models with retrieval augmentation [EB/OL]. (2024-11-19) [2025-05-02]. https://arxiv.org/abs/2307.11019.

[12] BANG Y J, CAHYAWIJAYA S, LEE N, et al. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity[EB/OL]. (2023-11-28). [2025-05-02]. https://arxiv.org/abs/230.04023v4.

[13] RENZE M. The effect of sampling temperature on problem solving in large language models[C]. Miami: Findings of the Association for Computational Linguistics, 2024.

[14] REIMERS N, GUREVYCH I. Sentence-BERT: Sentence embeddings using siamese bert-networks[EB/OL]. (2019-08-27). [2025-05-02]. https://arxiv.org/abs/1908.10084.

[15] JOHNSON J, DOUZE M, JÉGOU H. Billion-scale similarity search with GPUs[J]. IEEE transactions on big data, 2021, 7(3): 535-547.

[16] MENG K, BAU D, ANDONIAN A, et al. Locating and editing factual associations in GPT[EB/OL]. (2023-01-13) [2025-05-02]. https://arxiv.org/abs/2202.05262.

[17] WANG T L, JIAO X F, ZHU Y H, et al. Adaptive activation steering: a tuning-free LLM truthfulness improvement method for diverse hallucinations categories[C]. Sydney: Proceedings of the ACM on Web Conference, 2025.

[18] LU K, LIANG Z, PAN D, et al. Med-R^2: Crafting trustworthy LLM physicians through retrieval and reasoning of evidence-based medicine[EB/OL]. (2025-01-21) [2025-05-02]. http://www.paperreading.club/page?id=279360.

[19] SUSNJAK T, HWANG P, REYES N, et al. Automating research synthesis with domain-specific large language model fine-tuning[J]. ACM transactions on knowledge discovery from data, 2025, 19(3): 1-39.

[20] YE A J, MAITI A, SCHMIDT M, et al. A hybrid semi-automated workflow for systematic and literature review processes with large language model analysis[J]. Future internet, 2024, 16(5): 167.

[21] LI Y, DATTA S, RASTEGAR-MOJARAD M, et al. Enhancing systematic literature reviews with generative artificial intelligence: development, applications, and performance evaluation[J]. Journal of the American medical informatics association, 2025, 32(4): 616-625.

[22] WANG Z, CAO L, DANEK B, et al. Accelerating clinical evidence synthesis with large language models. [EB/OL]. (2024-10-28) [2025-05-02]. https://arxiv. org/abs/2406. 17755.

[23] JOOS L, KEIM D A, FISCHER M T. Cutting through the clutter: The potential of llms for efficient filtration in systematic literature reviews[EB/OL]. (2025-04-28) [2025-05-02]. https://arxiv.org/abs/2407.10652.

[24] HE Y, HUANG G, FENG P, et al. PaSa: An LLM agent for comprehensive academic paper search[EB/OL]. (2025-05-27) [2025-05-02]. https://arxiv.org/abs/2501.10120.

[25] MATSUI K, UTSUMI T,AOKI Y,et al.Human-comparable sensitivity of large language models in identifying eligible studies through title and abstract screening: 3-layer strategy using GPT-．5 and GPT-4 for systematic reviews[J]. Journal of medical internet research, 2024, 26: e52758.

[26] GWON Y N, KIM J H, CHUNG H S, et al. The use of generative AI for scientific literature searches for systematic reviews: ChatGPT and microsoft Bing AI performance evaluation[J]. JMIR medical informatics, 2024, 12: e51187.

[27] CAO C, SANG J, ARORA R, et al. Development of prompt templates for large language model-driven screening in systematic reviews[J]. Annals of internal medicine,

2025, 178(3): 389-401.

[28] OAMI T, OKADA Y, NAKADA T A. Performance of a large language model in screening citations[J]. JAMA network open, 2024, 7(7): e2420496.

[29] CHEN H C, JIANG Z H, LIU X Y, et al. Can large language models fully Automate or partially assist paper selection in systematic reviews?[J]. British journal of ophthalmology, 2025: bjo-2024-326254.

[30] DELGADO-CHAVES F M, JENNINGS M J, ATALAIA A, et al. Transforming literature screening: The emerging role of large language models in systematic reviews[J]. Proceedings of the national academy of sciences of the United States of America, 2025, 122(2): e24119621.

[31] LI M, SUN J P, TAN X M. Evaluating the effectiveness of large language models in abstract screening: a comparative analysis[J]. Systematic reviews, 2024, 13(1): 219.

[32] TRAN V T, GARTLEHNER G, YAACOUB S, et al. Sensitivity and specificity of using GPT-.5 turbo models for title and abstract screening in systematic reviews and meta-analyses[J]. Annals of internal medicine, 2024, 177(6): 791-799.

[33] HAN B L, SUSNJAK T, MATHRANI A. Automating systematic literature reviews with retrieval-augmented generation: a comprehensive overview[J]. Applied sciences, 2024, 14(19): 910.

[34] FOLEY K, MCLEAN C, DE ZYLVA R, et al. Developing a critical imagination for how researchers can use artificially intelligent tools reflexively and responsibly during qualitative literature reviews[J]. International journal of qualitative methods, 2025, 24: 16094069251316249.

[35] KHRAISHA Q, PUT S, KAPPENBERG J, et al. Can large language models replace humans in systematic reviews? Evaluating GPT-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages[J]. Research synthesis methods, 2024, 15(4): 616-626.

[36] JIN Q, LEAMAN R, LU Z Y. PubMed and beyond: biomedical literature search in the age of artificial intelligence [J]. eBioMedicine, 2024, 100: 104988.

[37] PETERSEN K, GERKEN J M. On the road to interactive LLM-based systematic mapping studies[J]. Information and software technology, 2025, 178: 107611.

[38] AN Y W, CHENG Y H, PARK S J, et al. HyperRAG: enhancing quality-efficiency tradeoffs in retrieval-augmented generation with reranker KV-cache reuse[EB/OL]. (2025-04-03) [2025-05-02]. https://arxiv. org/abs/2504.02921v1.

[39] KIM J, VAJRAVELU B N. Assessing the current limitations of large language models in advancing health care education[J]. JMIR formative research, 2025, 9: e51319.

[40] WEI J, WANG X Z, SCHUURMANS D, et al. Chain-of-thought prompting elicits reasoning in large language models[EB/OL]. (2023-01-10) [2025-05-02]. https://arxiv.org/abs/2201.11903v6.

[41] 王祖恒, 韦春梦, 鲁文浩. 人工智能在智能医疗保健中的应用研究[J]. 广西医科大学学报, 2025, 42(1): 1-8.

[42] GAO J Y, ZHU Y H, WANG W Q, et al. A comprehensive benchmark for COVID-19 predictive modeling using electronic health records in intensive care[J]. Patterns, 2024, 5(4): 100951.

[43] MA L T, ZHANG C H, GAO J Y, et al. Mortality prediction with adaptive feature importance recalibration for peritoneal dialysis patients[J]. Patterns, 2023, 4(12): 10089.

[44] 程　京, 李　勐, 李　航, 等. 中医智能装备研究进展与思考[J]. 广西医科大学学报, 2023, 40(4): 523-532.

[45] 郑　雨, 李　呈, 胡贵平, 等. 机器学习技术在环境健康领域中的应用进展[J]. 广西医科大学学报, 2024,41(11): 1558-1564.

[46] 王　硕, 刘天语, 汪　琛, 等. 试论生成式人工智能的医疗应用能力与风险边界[J]. 医学与哲学, 2024, 45(12): 1-5.

[47] GALLIFANT J, AFSHAR M, AMEEN S, et al. The TRIPOD-LLM reporting guideline for studies using large language models[J]. Nature medicine, 2025, 31(1): 60-69.

本文引用格式：

黄衍楠，桑浩然，刘　宇，等.大语言模型辅助医学系统综述：方法、发展方向和应用 [J].广西医科大学学报，2025, 42(3)：323 -331. DOI：10.16190/j. cnki. 45-1211/r.2025.03.001

HUANG Y N, SANG H R, LIU Y, et al. Empowering medical systematic reviews with large language models: methods, development directions, and applications [J]. Journal of Guangxi medical university, 2025, 42(3)：323 -331. DOI：10.16190/j.cnki.45-1211/r.2025.03.001