智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

医学大语言模型的研发与应用系统综述

王璐, 丁慕菲, 周鹤, 何倩倩, 宋江典

引用本文:
王璐, 丁慕菲, 周鹤, 等. 医学大语言模型的研发与应用系统综述[J]. 智能系统学报, 2025, 20(6): 1295-1303.
WANG Lu, DING Mufei, ZHOU He, et al. Developing and employing large language models in medicine[J]. *CAAI Transactions on Intelligent Systems*, 2025, 20(6): 1295-1303.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202410020

DOI: 10.11992/tis.202410020
网络出版地址: https://link.cnki.net/urlid/23.1538.TP.20250901.1750.002

医学大语言模型的研发与应用系统综述

王璐[1,2], 丁慕菲[1], 周鹤[1], 何倩倩[1], 宋江典[1]
(1. 中国医科大学 健康管理学院, 辽宁 沈阳 110122; 2. 中国医科大学附属盛京医院, 辽宁 沈阳 110004)

摘 要: 自 2022 年 11 月 ChatGPT(chat generative pre-trained Transformer)问世以来, 针对医学应用场景的大语言模型(large language models, LLMs)相关研究逐渐成为热点。然而当前缺乏对医学大语言模型研发以及应用现状的系统分析。为了更好地理解这些专门为医学领域设计的 LLMs 并评估其应用价值, 本综述系统分析了截止至 2024 年 6 月 11 日, 在 PubMed、Google Scholar、arXiv、bioXiv 和 medRxiv 等数据库中发表的为医学领域开发的专有 LLMs, 同时对 LLMs 在临床应用场景中的相关应用研究进行了梳理。研究结果表明, 当前共计 129 项研究提出了基于医学相关语料研发的医学 LLMs, 而基于 LLMs 在临床应用场景中的应用涵盖了 LLMs 对医疗咨询的回应、不同模型间的比较、与专业医生的性能对比, 以及医疗从业相关人员对 LLMs 的观点等 4 类研究内容。综述结果表明, 通用型 LLMs, 如 ChatGPT、GPT-4 等在生成医疗记录时的准确性和完整性较高, 而专门针对某些疾病所研发的 LLMs 则更擅长回答特定病症的问题, 尽管它们的答复在全面性方面可能有所欠缺。医疗专家在辨别 LLMs 生成的文本与人类医生的文本时可能面临困难, 但 LLMs 对重复提问的回复存在变异性。此外, 从医学伦理角度看, LLMs 在易读性和可能涉及种族及地域偏见的传播方面存在挑战, 而且缺乏从患者或医疗保险提供商视角对 LLMs 可信度和责任等问题进行评估的研究。

关键词: 聊天机器人; 人工智能; 大语言模型; ChatGPT; 医疗保健; 临床诊断; 医疗咨询; 医疗信息学
中图分类号: TP18; R319 **文献标志码**: A **文章编号**: 1673−4785(2025)06−1295−09

中文引用格式: 王璐, 丁慕菲, 周鹤, 等. 医学大语言模型的研发与应用系统综述 [J]. 智能系统学报, 2025, 20(6): 1295–1303.
英文引用格式: WANG Lu, DING Mufei, ZHOU He, et al. Developing and employing large language models in medicine[J]. CAAI transactions on intelligent systems, 2025, 20(6): 1295–1303.

Developing and employing large language models in medicine

WANG Lu[1,2], DING Mufei[1], ZHOU He[1], HE Qianqian[1], SONG Jiangdian[1]
(1. School of Health Management, China Medical University, Shenyang110122, China; 2. Shengjing Hospital of China Medical University, Shenyang 110004, China)

Abstract: Since the introduction of ChatGPT (chat generative pre-trained Transformer) in November 2022, studies related to large language models (LLMs) for medical applications are increasing; however, a systematic review of this field is lacking. This review covered studies indexed in PubMed, Google Scholar, arXiv, bioXiv, and medRxiv up until June 31, 2024, and identified 129 medical LLMs. LLMs were evaluated in clinical contexts, including their responses to medical queries, performance comparison, and specialist evaluation. The results revealed that general-purpose LLMs, such as ChatGPT and GPT-4, demonstrate better accuracy in generating medical records, whereas disease-specific LLMs excel in niche areas but may lack comprehensiveness. Challenges include variability in responses, readability issues, and biases, with few studies on LLM trustworthiness from patient or insurance perspectives.

Keywords: chatbot; artificial intelligence; large language models; ChatGPT; health care; clinical diagnosis; medical consultation; medical informatics

随着大语言模型(large language models, LLMs)相关研究的兴起, 人工智能正引领教育、医疗和科学研究范式的变革。自 OpenAI 于 2022 年 11 月发布 ChatGPT(chat generative pre-trained Transformer) 以来, LLMs 在文本对话领域取得了巨大进步[1]。这些模型在某些特定场景下的表现甚至超越人类。与以往的自然语言处理技术不同的是, ChatGPT 使用 Transformer 架构, 通过自注意力机制并行处理不同的词元。加之奖励模型和微调

收稿日期: 2024−10−15. **网络出版日期**: 2025−09−02.
基金项目: 国家自然科学基金项目(92259104).
通信作者: 宋江典. E-mail: jdsong@cmu.edu.cn.

机制的引入，GPT 系列模型在多种语言任务中均有出色的表现，并在专业评测中达到了人类水平[2-3]。

在医疗领域，LLMs 相关研究已经证明了其在日常临床工作中的潜在价值，特别是在医疗相关文档的生成方面，例如标准化的临床记录、出院总结、放射学报告以及基于特定临床情境的同理心回应等。虽然 LLM 最初的设计并非专门针对临床应用场景，但通过微调或重新训练等方法，为特定疾病研发的专有 LLMs 以及如 Med-PaLM 和 GatorTron 等医疗基础模型[4-5]，已经在放射学、病理学和肿瘤学等领域中证明了自己的能力[6-8]。这些研究成果表明 LLMs 可能改变传统的临床实践模式。

随着医疗领域 LLMs 相关研究的日益增多，当前迫切需要对 LLMs 在医疗场景中的研发及应用进展进行归纳总结。尽管已有针对 LLMs 在临床、研究和教育方面的相关综述，但目前尚缺少专门针对临床场景中 LLMs 研发及相关应用的综述分析。因此，本综述将系统归纳医疗领域 LLMs 的研究现状，并阐明其在医学应用中的收益与风险。

1　综述方法

本综述严格遵循系统综述和荟萃分析报告的标准（preferred reporting items for systematic reviews and meta-analysis, PRISMA）。由于不涉及个人信息，因此无需伦理审批和参与者知情同意。

本研究基于 PubMed、Google Scholar、arXiv、bioXiv 和 medRxiv 数据库，检索时间为自各数据库建立至 2024 年 6 月 11 日期间，检索已发表的与临床医疗中 LLMs 研发和应用相关的原始文献。检索关键词包括“Large language model”“Chatbot”“ChatGPT”和“Medicine”。本文分析纳入文献标准如下：研究应为医疗领域 LLMs 的开发或应用方面的原创研究；形式包括期刊论文、原创报告、GitHub 上的程序代码或会议论文；对文献的语言和作者不作限制。由两位研究员独立根据纳入标准筛选文献，如有分歧则通过共识会议讨论解决。从每篇符合条件的文献中提取以下信息：LLM 的名称、发表的来源和日期、基座模型、研究目的、训练数据和模型参数。

2　医学 LLMs 的发展与应用

2.1　医学 LLMs 的发展

根据检索策略，现有 129 个为医学领域研发的 LLMs。其中，75.2% 的 LLMs（共计 97 个）都是基于英文医学资料训练而成，而针对中文、阿拉伯语和日语等多语种资料训练的 LLMs（共计 32 个，占 24.8%）在处理非英语环境下的医疗咨询方面显示出了良好的效果[9-12]。

医疗领域 LLMs 一般基于临床应用场景、生物实验室和医学文献中所获取的专业语料进行训练。这些模型的训练通常采用两种方式：一是基于通用模型如 GPT、BERT（bidirectional encoder representations from Transformers）[13]、ESM（evolutionary scale modeling）[14] 和 LLaMA（large language model meta artificial intelligence）[15]，结合关于特定疾病的语料库进行微调训练；二是使用数以百万计的医学文本或图像数据从头开始训练模型，如图 1 所示。例如，BioMedLM 是一个基于 GPT-2 模型架构的 LLM，包含了 27 亿参数，使用 PubMed 数据库中的摘要和全文进行训练，该模型专注于生物医学领域，其在回应生物医学咨询方面的表现与 GPT-3 相当[16]。此外，BioBERT、NYUTron、ESMFold 和 ProGen2[17-20] 等生物医学 LLMs 也将 BERT 和 ESM 作为基础模型。BioGPT 整合了 BERT 和 GPT 模型，在 1500 万篇 PubMed 摘要上进行训练以完成多项生物医学任务[21]。除了 GPT 和 BERT 以外，LLaMA 和 PaLM 也是常用的通用模型，经过微调后所研发的 Med-PaLM 和 MedAlpaca 等 LLMs 的表现已接近专业医生的水平[4,22]。在专门针对某些疾病所研发的 LLMs 研究中，LiVersa 使用 30 个针对肝病的美国肝病学会 (American Association for the Study of Liver Diseases) 指南和文档进行训练，研究结果能够为乙型肝炎和肝细胞癌提供监测治疗建议[23]。此外，针对心理学疾病基于 Alpaca 和 FLAN（finetuned language net）模型所研发的 Mental-Alpaca 和 MentalFLAN-T5 模型在处理心理健康问题时的表现优于 ChatGPT 3.5 和 4.0 版本[24]。

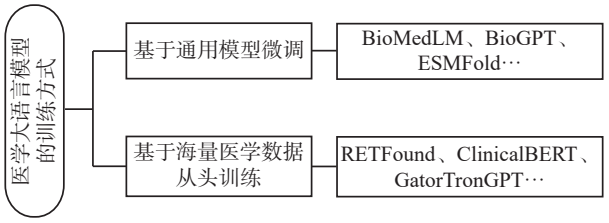

图 1　医学大语言模型的训练方式

Fig. 1　Training paradigms for medical LLMs

除了对现有模型进行微调之外，当前有研究者提出使用大量未经标注的医学数据从零开始训练 LLMs，从而满足多种医疗应用情景的需求。

例如, Zhou 等[25] 使用 160 万张未标注的视网膜图像数据进行训练, 提出 RETFound 模型, 在视网膜疾病的诊断及预测方面表现突出。而基于 GPT-3 开发的 GatorTronGPT 模型, 在训练过程中使用了 2270 亿个词元, 具有强大的医学文本生成能力[26]。此外, 当前已有研究提出专门应用于病理学领域的视觉语言基础模型, 该模型使用数百万张组织病理学图像和生物医学注释进行训练, 能够实现病理图像的分类、分割、描述文本生成, 以及通过描述性文本检索对应图像和通过图像检索对应文本的功能[27-28]。这些研究成果揭示了 LLMs 在医学领域的未来应用潜力。

2.2　LLMs 在临床实践中的应用

基于医学相关语料研发的医学 LLMs 目前正处于迅速发展阶段, 为了全面评估这些模型在临床场景中的应用潜力, 本研究聚焦于 4 个主要研究方向: 1) LLMs 对医疗咨询回应的准确性评估研究, 这一研究方向的价值在于确保 LLMs 提供的信息准确可靠, 这直接关系用户的健康和安全; 2) 不同模型间的比较分析, 旨在识别各个模型的优势与局限, 以指导医疗专业人员选择合适的模型以满足特定需求; 3) LLMs 与专业医生的性能对比, 有助于理解这些模型在实际临床场景中的应用与局限性, 明确 LLMs 在医疗决策过程中的定位; 4) 医疗从业相关人员对 LLMs 的观点分析, 通过了解医疗从业人员对 LLMs 的看法和接受度, 有助于识别实施过程中的障碍, 为制定有效的培训和教育策略提供依据。

2.2.1　LLMs 对医疗咨询回应的准确性评估研究

LLM 的文本对话方式为患者提供了获取医疗信息并与 LLMs 进行互动的窗口, 这一创新正在改变传统医疗咨询方式。Pan 等[29] 通过分析 Google Trends 上最常见的癌症相关搜索词, 评估了 LLM 聊天机器人输出的癌症相关信息的准确性。研究结果表明, 这些信息质量普遍较高, 并未发现误导性内容, 但在易于理解性和实际应用价值方面存在不足, 因此研究认为不能将 LLM 聊天机器人作为获取医疗信息的唯一渠道。而另一项研究指出 ChatGPT 在回答有关乳腺癌的预防与筛查、公共卫生问题、非酒精性脂肪肝以及胃食管反流病等问题时, 正确回答率超过 80%, 并且可理解性处于中等水平[30-33]。

在对 LLM 聊天机器人的临床诊断能力评估研究中, GPT-4.0 在肿瘤定位、识别以及进展推断方面相比于前一代技术(GPT-3.5)具有更好的性能[34]。当输入文本等单一模态信息时, ChatGPT 能够独立构建先进的机器学习模型以预测病原体基因和临床结局, 减少人为因素带来的偏差[35-36]。然而, 另一项研究指出, 基于 Diagnosis Please 数据集, 若只向 LLM 输入患者病史或医学图像等单模态信息, ChatGPT 的准确率约为 50%。但当涉及到心血管放射学相关疾病并且将多模态信息作为输入时, ChatGPT 的诊断准确率则比其他疾病更高[37]。

此外, 已有研究显示 LLMs 具备通过美国执业医师资格考试的能力[38], 但在其他专科医学测试中的结果却不尽相同[39-42]。例如, ChatGPT 4.0 在神经内科、神经外科和放射科考试中的表现超越了 ChatGPT 3.5, 但仍略低于医师的平均水平[42]。另有研究表明, 无论是 GPT-3.0 还是 GPT-4.0, 均未通过美国胃肠病学会自我评估测试[43]。以上结果表明 LLMs 在成为临床决策工具之前仍需进一步的提升与综合评估[44]。

另一方面, 虽然 LLMs 在回应医疗咨询方面的表现可嘉, 但其在提供临床建议和决策时涉及到的种族、性别和地域偏见问题也引起了广泛关注。特别是在为某些因种族或性别而具有更高疾病风险的群体提供医疗建议时, ChatGPT 表现出了一定的倾向性, 这可能会加剧弱势群体的疾病易感性。此外, 对不同地理区域的人群提供诊断和治疗建议时, 研究发现人群间存在显著差异, 比例超过三分之一[45]。LLM 聊天机器人有时会生成看似合理但实际不准确的“幻觉(错觉)”回答, 这可能会对提问者产生误导[46]。因此, 预评估和外部监督对于确保未来 LLMs 在临床应用中的可靠性将至关重要。

2.2.2　不同模型间的比较分析

为肝脏病设计的 LiVersa 专有模型使用检索增强生成架构以减少通用型 LLMs 产生的误导性信息, 在回答肝脏病学相关问题时, 准确率超过了 ChatGPT 4.0 和 LLaMA 2, 尽管其答复的全面性和安全性方面略有不足[23]。Benary 等[47] 通过分析 10 位虚拟癌症患者的临床案例, 指出 BioMedLM 在确定治疗方案方面的表现超越了 Perplexity、ChatGPT 和 Galactica 等模型, 对临床意义未明的分子变异能有效降低解读分歧。基于分子肿瘤委员会成员的人工评估结果显示, 上述 LLMs 生成的治疗方案与专家方案相比, 均表现出较高的识别准确率(置信度为 75%~80%)。此外, 以图像作为输入的放射学专有 LLM——Vicuna-13B, 在辨识不同医学图像数据库中放射学报告的 13 项特定结果时, 与现有的标注者达成了中等到显著程度的一致性水平[48]。

一项针对放射学报告中肺癌筛查的非专业问题的研究发现，通用型 LLMs 中的 ChatGPT 3.5 达到了 70.8% 的正确率，这一比率显著高于 Google Bard (Gemini) 的 19.1%，超出幅度达 51.7%[49]。值得注意的是，尽管 Microsoft Bing 和 Google Search 的正确率分别为 61.7% 和 55.0%，但 Google Search 与 ChatGPT 3.5 在答案一致性方面远远超过了其他模型（$P = 0.002$）。这一发现在另一项临床决策支持研究中得到进一步证实，Google Search 在诊断相关医疗状况时达到了最高的评分者间一致性[50]。在对比 ChatGPT 的不同版本时，研究发现 ChatGPT 3.5 和 ChatGPT 4.0 在处理常见疾病时的诊断性能都优于罕见疾病（$P < 0.000\,1$），二者在罕见疾病的诊断性能方面没有显著差异（$P = 0.050\,3$）[50]，但在甲状腺结节诊断等特定研究中，ChatGPT 4.0 展现出更优越的性能，其准确性不仅超越了 ChatGPT 3.5，也优于其他通用型 LLMs[51-52]。

然而，先前的研究也表明在评估通用型 LLMs 性能时可能会得出相互矛盾的结论。为了减轻放射科医生撰写报告的工作负担，Amin 等[53] 采用 Likert 量表评估 ChatGPT 3.5、ChatGPT 4.0、Google Bard 和 Microsoft Bing 生成的简化版放射学报告的质量。研究结果显示，ChatGPT 3.5 在报告简化的准确性为 86.0%，显著高于其他模型（均小于 83.3%），但 Google Bard 在提供补充信息以增强报告完整性方面表现最佳。相比之下，Li 团队和 Lim 团队的研究显示[54-55]，ChatGPT 4.0 在执行与精神病学认证考试、精神障碍鉴别诊断以及眼科近视自我纠正等相关任务时，性能优于 Google Bard、LLaMA 2 和 ChatGPT 3.5。这些不同的研究结果强调了在临床实践中应用 LLMs 之前进行全面评估的重要性。此外，围绕种族医学偏见的相关研究表明，现有的 LLMs 在评估与种族医学误解相关的 9 个问题时，包括 ChatGPT 3.0、ChatGPT 4.0、Google Bard 和 Anthropic Claude 在内的 4 种 LLMs 的回答存在强化种族化医学观念或传播种族刻板印象的倾向[56]。

2.2.3　LLMs 与专业医生的性能对比

当前，LLMs 在生成医疗记录方面已经达到了与专业人员相当的水平[57-58]。在一项针对眼科问题的研究中，8 位专业医生尝试辨别 ChatGPT 3.5 和眼科医生给出的建议[59]，发现两者提供的回复在错误信息的出现率或潜在风险方面没有显著差异。然而，有 21.0% 由聊天机器人生成的回答被误认为是人类医生撰写的，而人类撰写的答案则有 64.6% 被误判为 AI 生成。在一项通过 GPT 识别图像进行文本生成的甲状腺结节诊断研究中，当融合了多位资深医生观察结果与大型语言模型结果后，ChatGPT 4.0 表现出了与该融合分析结果相当的诊断能力[51]。Decker 等[60] 的研究还比较了 ChatGPT 3.5 和外科医生所设计的知情同意书的可读性和全面性，发现虽然 ChatGPT 3.5 生成的文件在可读性方面略显不足，但在内容的准确性和完整性，尤其是在药物治疗方案和潜在风险描述方面，与外科医生设计的文件不相上下。这些研究表明，LLMs 有潜力在临床实践中辅助完成知情同意文件的编写，为医生提供个性化的风险信息，从而减轻临床文书工作负担。

在对医学专有模型的评估方面，佛罗里达大学开发的临床医学文本生成模型 GatorTronGPT[26] 在语言可读性方面与人类表现没有显著差异。另一项研究利用 209 份美国放射学会适当性标准文件对 ChatGPT 3.5 和 LlamaIndex 模型进行微调训练，开发出 accGPT 模型，该模型显示出能够根据具体情况提供个性化的影像扫描推荐的能力。进一步研究显示，accGPT 模型提供的准确答案比例明显高于放射科医生以及 ChatGPT 3.5 和 ChatGPT 4.0，同时具有时间成本优势[61]。另一方面，通过对公共社交媒体上 195 个患者的问题进行研究，3 位医疗专业人员发现在回答公共关注的问题时，LLM 聊天机器人在取得更高的回复质量的同时，还显示出比医生回复更多的理解和关怀能力（$P<0.001$）[62]。

然而，ChatGPT 在回答相同问题时表现出的不一致性也引起了人们的担忧。在一项关于结肠镜检查的问题研究中，胃肠病学专家对 ChatGPT 的回答进行了评价，认为其在准确性、易理解性、科学充分性和满意度方面与非人工智能的回答相当。然而，当模型多次被询问相同的问题时，其生成的回答在相似性上存在显著变化，波动范围在 28%~77%[63]，这表明未来对于 ChatGPT 回复的可复现性评价可能是其在临床应用的一项重要参考指标。

2.2.4　医疗从业人员对 LLMs 的观点分析

虽然研究表明 LLMs 有助于提高临床工作的效率，但医疗从业人员对此看法不一。在精神卫生领域的一项研究中，美国精神病学协会的 138 名精神科医生中，有超过 70% 认为 ChatGPT 对提高工作效率和提供预咨询信息有帮助[64]。而在泌尿科医生中进行的另一项调查显示，只有半数受访者在学术研究中使用过 ChatGPT 或其他类似

的 LLMs, 而将这些工具实际应用于临床的医生更是寥寥无几。58.5% 的受访者认为 LLMs 的输出缺乏充分验证, 不应直接用于临床治疗[65]。还有研究指出, LLMs 有时会提供不准确或不完整的信息, 例如推荐在伤口处理时使用非处方抗生素, 这类信息需要医生进一步审核[66]。以上结果均表明目前在临床实践中更适合将 LLMs 定位为临床辅助工具使用。

更加专业化的提示词可以提高 LLMs 生成关键信息的准确性, 使患者能够更容易获取必要的医疗信息。然而, 目前的学术研究主要侧重于医疗专业人员对 LLMs 问答的评估分析, 而从患者和医疗保险供应商角度考虑 LLMs 临床应用前景的关注较少。另一方面, 由于患者通常无法轻易获得专业的医疗信息且他们对 LLMs 提供的回应感到满意的可能性更高, 这使得患者可能会越来越依赖通过 LLMs 获取医疗信息。例如, 对 LLMs 就黑色素瘤诊断和预后问题的回答进行评估, 发现 90% 以上的回答适合作为面向患者的信息[67]。此外, 患者对于经 ChatGPT 润色完善后的回答总结高度满意, 这些总结既详细又有很强的结构性, 提供了与实际医疗需求密切相关的信息[66]。然而, 过度依赖这些技术可能会影响临床决策的客观性, 使得医疗保险各方的责任界限变得模糊。目前, 已有医疗器械公司将人工智能技术整合到其产品和服务中[68], 但在将 LLMs 应用于临床实践之前, 必须全面而深入的验证这些 LLMs 在医疗领域的准确性。

3　医学 LLMs 的挑战与展望

随着 LLMs 在医疗领域的应用日益增多, 对其综合性能、合规性、伦理问题和可访问性的考量变得尤为重要。本研究旨在综合评估 LLMs 在医疗回应中的准确性, 并识别其在实际临床应用中面临的挑战。LLMs 的合规性和可靠性对于患者数据保护、医疗服务质量至关重要。提升模型性能和易用性则直接关系到用户对 LLMs 的接受程度和实际应用效果。伦理问题和数据保护是 LLMs 发展中不可忽视的方面, 它们关系到患者隐私和模型的社会责任。减少偏见和不平等问题对于实现医疗公正和提高医疗服务的普遍性至关重要。以下将对上述几个方面展开详细讨论, 并提出本研究的局限与医学 LLMs 的未来研究方向。

3.1　合规性和可靠性

根据之前在医疗卫生领域中整合新技术的相关警告, 确保 LLMs 遵循《健康保险流通与责任法案》(Health Insurance Portability and Accountability Act, HIPAA)以及与医疗保健提供者建立业务合作协议的重要性迫在眉睫[69]。一方面, 健康保险公司可能会质疑由 LLMs 生成的文档的可靠性, 因此 LLMs 在临床场景中的应用可能导致医生在向保险公司报销医疗费用时遇到更多阻碍。这一问题主要源于未能明确 LLMs 是临床场景的辅助工具还是决策者[65,70-71]。另一方面, 使用来自网络的繁杂数据进行训练会导致 LLMs 输出不稳定, 如果仅使用医学相关的权威文本训练可以提高其可靠性[72]。由于现阶段 LLMs 的训练数据来源不透明, 这进一步涉及到对现有知识产权的潜在侵犯。为了提高透明度, 部分倡议认为需要公开 LLMs 所使用的受版权保护的材料, 这种倡议已经对《人工智能法案》的制定产生了影响[73]。为了使未来 LLMs 更符合 HIPAA 的规定, 建立人与技术之间的信任, 并提高透明度, 这些措施将有助于医疗从业人员、人工智能开发者、临床医生和决策者更好地理解和利用集成到医院系统和医疗保健中的 LLMs[74-75]。

3.2　模型性能和易用性

当前研究指出, LLMs 在回答的准确性和一致性方面仍存在一定的问题[53]。使用高质量的医学语料库训练模型虽然能够在一定程度上克服以上问题, 但有研究显示, 近 40% 的医生在临床实践中使用 ChatGPT 或其他 LLMs 时仍然表示受到了限制, 未来 LLMs 应更加聚焦于解决医生在临床工作中遇到的相关实际问题[65]。同时, 尽管 LLMs 的医疗文本生成能力与医学专家相当, 但在临床应用场景中处理更深层次的医学问题时仍显现出缺陷, 比如基于影像结果进行临床具体推理等[40,76]。目前, LLMs 在回答的安全性和所提供答案的可读性等方面也在一定程度上阻碍了普通用户对 LLMs 的使用, 提升 LLMs 的易用性、回答的安全性及答案的可理解性仍是发展的重点[23,29]。此外, 尽管目前人们可以免费使用一些 LLMs(如 ChatGPT 3.5), 但 LLMs 的训练、部署和运行都需要一系列基础设施支持[77]。未来的开发可以通过不断优化算法降低资源消耗以适应不同医疗场景下的需求, 并提升 LLMs 在不同医疗环境中的经济可行性, 通过提供用户培训等措施提高 LLMs 的易用性, 确保它们在各种医疗环境中的应用。

3.3　伦理问题和数据保护

随着医学 LLMs 的迅速发展, 伦理问题日益凸显。研究表明, 超过 60% 的临床医生对与 LLMs 相关的伦理问题表示担忧, 包括抄袭、幻觉和错

误信息等问题[65]。此外，输入到 LLMs 的文本或图像中可能包含敏感且受保护的健康信息或未公开的数据，这些信息在 LLMs 中可能会有泄露的风险[78]。为了解决这些问题，未来的 LLMs 开发必须将伦理准则作为核心原则。这包括确保模型在处理敏感数据时的安全性和隐私保护，以及开发透明的数据处理和存储策略。此外，LLMs 的开发应遵循严格的伦理审查流程，包括对模型的潜在影响进行评估，确保其符合国际伦理标准和法律法规。

3.4　偏见和不平等问题

LLMs 在提供治疗方案时可能会无意中强化种族和地域偏见，这不仅违背了个性化推荐的目的，还可能进一步加剧医疗健康领域中被忽视的世界人口多样性问题[45]，这些潜在的危害将影响 LLMs 在未来的大范围临床应用[56,78]。为了减少这些潜在的不平等问题，未来的 LLMs 开发需要采取多元化和包容性的理念。这包括使用包含不同种族、性别和地理位置的数据集进行训练，以及在模型设计中考虑文化敏感性和多样性。此外，开发者应与医疗专业人员合作，确保 LLMs 的输出能够满足不同患者群体的需求。同时，监管机构和专业组织应制定明确的指导原则，以监督 LLMs 的开发和应用，确保所有用户都能公平地受益于 LLMs 提供的服务。总的来说，LLMs 在医疗领域的未来发展需要跨学科的合作、持续的研究和严格的监管。通过这些努力，我们有望实现 LLMs 在提高医疗服务质量方面的全部潜力，同时确保患者安全、保护患者隐私数据以及相关伦理规范的合规性。

4　研究局限性

本综述的覆盖范围受检索时间与来源所限：检索时间截至 2024 年 6 月 11 日，主要依托 PubMed、Google Scholar、arXiv、bioRxiv 和 medRxiv，围绕医疗领域相关问题的 LLMs 的研究报告数量已经激增至数千篇，且更新极为频繁，难以完全纳入最新进展，存在一定遗漏风险。现有证据在语种与地域上总体偏向英语语境和少数高收入国家场景，低资源语言与中低收入国家的数据相对较少，这可能影响对跨区域可推广性的判断。

纳入研究的类型与质量差异较大，样本量、对照设计与方法细节披露不一；同时，不同研究在任务设定、模型版本及评价指标方面异质性较高，难以进行严格的定量合并，本研究更多采用叙述性综合。部分 LLMs 以“家族/版本”形式快速迭代，部分工作未清晰记录版本号、对齐方法或训练数据变更，削弱了可复现性与横向可比性。对于参数规模等模型设定的归纳，主要依据相关文章中提供的确切数字，或据所用基座模型进行合理估算，但个别研究并未记录具体参数信息，进一步增加了综合判断的不确定性。

此外，灰色文献和无效结果公开不足，易造成发表偏倚；多模态任务、真实场景下的流程安全与不良事件仍缺少系统性报告；部分研究未充分报告评分者间一致性与不确定性。受闭源或合规限制，部分模型与训练数据难以开展外部审查与独立复现；少数数据集还可能存在数据泄露或样本重叠等问题。

上述因素均会削弱现有结论的稳健性和可推广性。因此，未来需要进行更新、更全面且规模更大的调查研究，以解决这些局限性。

5　结束语

本研究揭示了医疗领域 LLMs 的研发现状，并评估了它们在临床应用场景中的价值与风险。虽然当前 LLMs 在医疗领域前景广阔，然而为了消除医学 LLMs 可能存在的偏见，实现 HIPAA 合规性、确保参与者的可信度以及提高透明度的努力至关重要。这些措施将有助于医疗专业人员、LLMs 开发者、临床医生和决策者之间在将 LLMs 整合进医院系统和医疗保健方面尽快达成共识。通过这些努力，未来 LLMs 或将在医疗领域发挥更大的作用。

参考文献：

[1] OpenAI. Introducing ChatGPT[EB/OL]. (2022−11−30)[2025−08−29]. https://openai.com/index/chatgpt.

[2] BISWAS S S. Role of chat GPT in public health[J]. Annals of biomedical engineering, 2023, 51(5): 868−869.

[3] SHEN Yongliang, SONG Kaitao, TAN Xu, et al. HuggingGPT: solving ai tasks with chatgpt and its friends in hugging face[J]. Advances in neural information processing systems, 2023, 36: 1−27

[4] SINGHAL K, TU T, GOTTWEIS J, et al. Towards expert-level medical question answering with large language models[EB/OL]. (2023−05−16)[2025−08−29]. https://arxiv.org/abs/2305.09617.

[5] YANG Xi, CHEN A, POURNEJATIAN N, et al. GatorTron: a large clinical language model to unlock patient information from unstructured electronic health records[EB/OL]. (2022−12−16)[2025−08−29]. https://arxiv.org/abs/2203.03540.

[6] HUANG Zhi, BIANCHI F, YUKSEKGONUL M, et al. A

visual-language foundation model for pathology image analysis using medical Twitter[J]. Nature medicine, 2023, 29(9): 2307−2316.

[7] TIU E, TALIUS E, PATEL P, et al. Expert-level detection of pathologies from unannotated chest X-ray images *via* self-supervised learning[J]. Nature biomedical engineering, 2022, 6(12): 1399−1406.

[8] YALAMANCHILI A, SENGUPTA B, SONG J, et al. Quality of large language model responses to radiation oncology patient care questions[J]. JAMA network open, 2024, 7(4): e244630.

[9] CUI Yiming, YANG Ziqiang, YAO Xin. Efficient and effective text encoding for Chinese llama and alpaca[EB/OL]. (2024−02−23)[2025−08−29]. https://arxiv.org/abs/2304.08177.

[10] YANG Songhua, ZHAO Hanjie, ZHU Senbin, et al. Zhongjing: enhancing the Chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue[J]. Proceedings of the AAAI conference on artificial intelligence, 2024, 38(17): 19368−19376.

[11] PIERI S, MULLAPPILLY S S, KHAN F S, et al. BiMediX: bilingual medical mixture of experts LLM [EB/OL]. (2024−12−10)[2025−08−29]. https://arxiv.org/abs/2402.13253.

[12] SUKEDA I, SUZUKI M, SAKAJI H, et al. JMedLoRA: medical domain adaptation on Japanese large language models using instruction-tuning[EB/OL]. (2023−12−01)[2025−08−29]. https://arxiv.org/abs/2310.10083.

[13] DEVLIN J, CHANG Mingwei, LEE K, et al. BERT: pre-training of deep bidirectional Transformers for language understanding[EB/OL]. (2019−05−24)[2025−08−29]. https://arxiv.org/abs/1810.04805.

[14] LIN Zeming, AKIN H, RAO R, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model[J]. Science, 2023, 379(6637): 1123−1130.

[15] TOUVRON H, LAVRIL T, IZACARD G, et al. LLaMA: open and efficient foundation language models[EB/OL]. (2023−02−27)[2025−08−29]. https://arxiv.org/abs/2302.13971.

[16] BOLTON E, VENIGALLA A, YASUNAGA M, et al. BioMedLM: A 2.7B parameter language model trained on biomedical text[EB/OL]. (2024−03−27)[2025−08−29]. https://arxiv.org/abs/2403.18421.

[17] 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.

[18] JIANG L Y, LIU X C, NEJATIAN N P, et al. Health system-scale language models are all-purpose prediction engines[J]. Nature, 2023, 619(7969): 357−362.

[19] NIJKAMP E, RUFFOLO J A, WEINSTEIN E N, et al. ProGen2: exploring the boundaries of protein language models[J]. Cell systems, 2023, 14(11): 968−978.

[20] Facebook Research. ESM (evolutionary scale modeling): ESMFold code and model release[EB/OL]. (2024−08−01)[2025−08−29]. https://github.com/facebookresearch/esm.

[21] LUO Renqian, SUN Liai, XIA Yingce, et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining[J]. Briefings in bioinformatics, 2022, 23(6): bbac409.

[22] HAN Tianyu, ADAMS L C, PAPAIOANNOU J M, et al. MedAlpaca: an open-source collection of medical conversational AI models and training data[J]. (2023−04−14)[2025−08−29]. https://arxiv.org/abs/2304.08247.

[23] GE Jin, SUN S, OWENS J, et al. Development of a liver disease-specific large language model chat interface using retrieval-augmented generation[J]. Hepatology, 2024, 80(5): 1158−1168.

[24] XU Xuhai, YAO Bingsheng, DONG Yuanzhe, et al. Leveraging large language models for mental health prediction via online text data[EB/OL]. (2024−01−28)[2025−08−29]. https://arxiv.org/abs/2307.14385.

[25] ZHOU Yukun, CHIA M A, WAGNER S K, et al. A foundation model for generalizable disease detection from retinal images[J]. Nature, 2023, 622(7981): 156−163.

[26] PENG Cheng, YANG Xi, CHEN Aokun, et al. A study of generative large language model for medical research and healthcare[J]. NPJ digital medicine, 2023, 6(1): 210.

[27] LU M Y, CHEN B, WILLIAMSON D F K, et al. A visual-language foundation model for computational pathology[J]. Nature medicine, 2024, 30(3): 863−874.

[28] CHEN R J, DING Tong, LU M Y, et al. Towards a general-purpose foundation model for computational pathology[J]. Nature medicine, 2024, 30(3): 850−862.

[29] PAN A, MUSHEYEV D, BOCKELMAN D, et al. Assessment of artificial intelligence chatbot responses to top searched queries about cancer[J]. JAMA oncology, 2023, 9(10): 1437−1440.

[30] HAVER H L, AMBINDER E B, BAHL M, et al. Appropriateness of breast cancer prevention and screening recommendations provided by ChatGPT[J]. Radiology, 2023, 307(4): e230424.

[31] AYERS J W, ZHU Z, POLIAK A, et al. Evaluating artificial intelligence responses to public health questions[J]. JAMA network open, 2023, 6(6): e2317517.

[32] PUGLIESE N, WAI-SUN WONG V W, SCHATTENBERG J M, et al. Accuracy, reliability, and comprehensibility of ChatGPT-generated medical responses for patients with nonalcoholic fatty liver disease[J]. Clin gastroenterol hepatol, 2024, 22(4): 886−889.

[33] HENSON J B, GLISSEN BROWN J R, LEE J P, et al. Evaluation of the potential utility of an artificial intelligence chatbot in gastroesophageal reflux disease management[J]. The American journal of gastroenterology, 2023, 118(12): 2276−2279.

[34] FINK M A, BISCHOFF A, FINK C A, et al. Potential of ChatGPT and GPT-4 for data mining of free-text CT reports on lung cancer[J]. Radiology, 2023, 308(3): e231362.

[35] TAYEBI ARASTEH S, HAN Tianyu, LOTFINIA M, et al. Large language models streamline automated machine learning for clinical studies[J]. Nature communications, 2024, 15(1): 1603.

[36] YAN Chao, GRABOWSKA M E, DICKSON A L, et al. Leveraging generative AI to prioritize drug repurposing candidates for Alzheimer's disease with real-world clinical validation[J]. NPJ digital medicine, 2024, 7(1): 46.

[37] UEDA D, MITSUYAMA Y, TAKITA H, et al. ChatGPT's diagnostic performance from patient history and imaging findings on the diagnosis please quizzes[J]. Radiology, 2023, 308(1): e231040.

[38] KUNG T H, CHEATHAM M, MEDENILLA A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models[J]. PLoS digital health, 2023, 2(2): e0000198.

[39] MIHALACHE A, HUANG R S, POPOVIC M M, et al. Performance of an upgraded artificial intelligence chatbot for ophthalmic knowledge assessment[J]. JAMA ophthalmology, 2023, 141(8): 798−800.

[40] BHAYANA R, KRISHNA S, BLEAKNEY R R. Performance of ChatGPT on a radiology board-style examination: insights into current strengths and limitations[J]. Radiology, 2023, 307(5): e230582.

[41] ALI R, TANG O Y, CONNOLLY I D, et al. Performance of ChatGPT and GPT-4 on neurosurgery written board examinations[J]. Neurosurgery, 2023, 93(6): 1353−1365.

[42] SCHUBERT M C, WICK W, VENKATARAMANI V. Performance of large language models on a neurology board-style examination[J]. JAMA network open, 2023, 6(12): e2346721.

[43] SUCHMAN K, GARG S, TRINDADE A J. Chat generative pretrained transformer fails the multiple-choice American college of gastroenterology self-assessment test[J]. The American journal of gastroenterology, 2023, 118(12): 2280−2282.

[44] GOODMAN R S, RANDALL PATRINELY J, STONE C A Jr, et al. Accuracy and reliability of chatbot responses to physician questions[J]. JAMA network open, 2023, 6(10): e2336483.

[45] ZACK T, LEHMAN E, SUZGUN M, et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study[J]. The lancet digital health, 2024, 6(1): e12−e22.

[46] MCGOWAN A, GUI Yunlai, DOBBS M, et al. ChatGPT and Bard exhibit spontaneous citation fabrication during psychiatry literature search[J]. Psychiatry research, 2023, 326: 115334.

[47] BENARY M, WANG X D, SCHMIDT M, et al. Leveraging large language models for decision support in personalized oncology[J]. JAMA network open, 2023, 6(11): e2343689.

[48] MUKHERJEE P, HOU B, LANFREDI R B, et al. Feasibility of using the privacy-preserving large language model vicuna for labeling radiology reports[J]. Radiology, 2023, 309(1): e231147.

[49] RAHSEPAR A A, TAVAKOLI N, KIM G H J, et al. How AI responds to common lung cancer questions: ChatGPT vs Google Bard[J]. Radiology, 2023, 307(5): e230922.

[50] SANDMANN S, RIEPENHAUSEN S, PLAGWITZ L, et al. Systematic analysis of ChatGPT, Google search and Llama 2 for clinical decision support tasks[J]. Nature communications, 2024, 15(1): 2050.

[51] WU Shaohong, TONG Wenjuan, LI Mingde, et al. Collaborative enhancement of consistency and accuracy in US diagnosis of thyroid nodules using large language models[J]. Radiology, 2024, 310(3): e232255.

[52] SAVAGE T, NAYAK A, GALLO R, et al. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine[J]. NPJ digital medicine, 2024, 7(1): 20.

[53] AMIN K S, DAVIS M A, DOSHI R, et al. Accuracy of ChatGPT, Google Bard, and microsoft Bing for simplifying radiology reports[J]. Radiology, 2023, 309(2): e232561.

[54] LI D J, KAO Yuchen, TSAI S J, et al. Comparing the performance of ChatGPT GPT-4, Bard, and Llama-2 in the Taiwan Psychiatric Licensing Examination and in differential diagnosis with multi-center psychiatrists[J]. Psychiatry and clinical neurosciences, 2024, 78(6): 347−352.

[55] LIM Z W, PUSHPANATHAN K, YEW S M E, et al. Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard[J]. eBioMedicine, 2023, 95: 104770.

[56] OMIYE J A, LESTER J C, SPICHAK S, et al. Large language models propagate race-based medicine[J]. NPJ digital medicine, 2023, 6(1): 195.

[57] GERTZ R J, BUNCK A C, LENNARTZ S, et al. GPT-4 for automated determination of radiological study and protocol based on radiology request forms: a feasibility study[J]. Radiology, 2023, 307(5): e230877.

[58] GARCIA P, MA S P, SHAH S, et al. Artificial intelligence-generated draft replies to patient inbox messages [J]. JAMA network open, 2024, 7(3): e243201.

[59] BERNSTEIN I A, ZHANG Y V, GOVIL D, et al. Comparison of ophthalmologist and large language model chatbot responses to online patient eye care questions[J]. JAMA network open, 2023, 6(8): e2330320.

[60] DECKER H, TRANG K, RAMIREZ J, et al. Large lan-

guage model-based chatbot vs surgeon-generated informed consent documentation for common procedures [J]. JAMA network open, 2023, 6(10): e2336997.
[61] RAU A, RAU S, ZOELLER D, et al. A context-based chatbot surpasses trained radiologists and generic ChatGPT in following the ACR appropriateness guidelines[J]. Radiology, 2023, 308(1): e230970.
[62] AYERS J W, POLIAK A, DREDZE M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum[J]. JAMA internal medicine, 2023, 183(6): 589−596.
[63] LEE T C, STALLER K, BOTOMAN V, et al. ChatGPT answers common patient questions about colonoscopy[J]. Gastroenterology, 2023, 165(2): 509−511.
[64] BLEASE C, WORTHEN A, TOROUS J. Psychiatrists' experiences and opinions of generative artificial intelligence in mental healthcare: an online mixed methods survey[J]. Psychiatry research, 2024, 333: 115724.
[65] EPPLER M, GANJAVI C, RAMACCIOTTI L S, et al. Awareness and use of ChatGPT and large language models: a prospective cross-sectional global survey in urology[J]. European urology, 2024, 85(2): 146−153.
[66] JIN J Q, DOBRY A S. ChatGPT for healthcare providers and patients: practical implications within dermatology [J]. Journal of the American Academy of Dermatology, 2023, 89(4): 870−871.
[67] YOUNG J N, O'HAGAN R, POPLAUSKY D, et al. The utility of ChatGPT in generating patient-facing and clinical responses for melanoma[J]. Journal of the American Academy of Dermatology, 2023, 89(3): 602−604.
[68] BENJAMENS S, DHUNNOO P, MESKÓ B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database[J]. NPJ digital medicine, 2020, 3: 118.
[69] BROWN S. Partnerships between health authorities and Amazon Alexa raise many possibilities: and just as many questions[J]. Canadian medical association journal, 2019, 191(41): E1141−E1142.
[70] SCHULMAN K A, NIELSEN P K Jr, PATEL K. AI alone will not reduce the administrative burden of health care[J]. JAMA, 2023, 330(22): 2159−2160.
[71] MELLO M M, ROSE S. Denial: artificial intelligence tools and health insurance coverage decisions[J]. JAMA health forum, 2024, 5(3): e240622.
[72] HARRER S. Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine[J]. eBioMedicine, 2023, 90: 104512.
[73] MINSSEN T, VAYENA E, GLENN COHEN I. The challenges for regulating medical use of ChatGPT and other large language models[J]. JAMA, 2023, 330(4): 315−316.
[74] SEZGIN E, SIRRIANNI J, LINWOOD S L. Operationalizing and implementing pretrained, large artificial intelligence linguistic models in the US health care system: outlook of generative pretrained transformer 3 (GPT-3) as a service model[J]. JMIR medical informatics, 2022, 10(2): e32875.
[75] SUN L, HUANG Y, WANG H, et al. Trustllm: trustworthiness in large language models[EB/OL]. (2024−01−10)[2025−08−29]. https://arxiv.org/abs/2401.05561.
[76] CABRAL S, RESTREPO D, KANJEE Z, et al. Clinical reasoning of a generative artificial intelligence model compared with physicians[J]. JAMA internal medicine, 2024, 184(5): 581−583.
[77] BHAYANA R. Chatbots and large language models in radiology: a practical primer for clinical and research applications[J]. Radiology, 2024, 310(1): e232756.
[78] LI Hanzhou, MOON J T, PURKAYASTHA S, et al. Ethics of large language models in medicine and medical research[J]. The lancet digital health, 2023, 5(6): e333−e335.

作者简介:

王璐, 博士研究生, 主要研究方向为医学数据分析、自然语言处理。发表学术论文论文 7 篇。 E-mail: luwang@sj-hospital.org。

丁慕菲, 硕士研究生, 主要研究方向为医学图像处理。 E-mail: dingmuou@163.com。

宋江典, 副教授, 博士, 中国计算机学会数字医学分会执行委员, 主要研究方向为医学图像处理与人工智能, 主持国家自然科学基金项目 2 项。发表学术论文 34 篇。 E-mail: jdsong@cmu.edu.cn。