医疗健康领域中对话式人工智能的评估范式:系统综述

廖委真　韩优莉　马骋宇
首都医科大学公共卫生学院　北京　100069

【摘　要】目的:系统梳理医疗健康对话式人工智能(Artificial Intelligence,AI)的评估范式,为促进医疗AI评估体系构建及评估方法改进提供参考。方法:采用系统综述方法,分析医疗健康对话式AI的评估范式,包括评估对象、评价指标、评估方法等。结果:共纳入60篇文献,评估对象以通用大模型为主,评价指标涵盖技术性能、信息质量、临床效果、用户体验、伦理与安全五个维度,已有研究采用的评估指标差异较大。存在测评问题与应用场景匹配度不高、评估者角色单一等问题。结论:当前医疗健康对话式AI评估体系尚不完善,未来应从模型类型的覆盖面、评估指标体系的综合性、评估方法的标准化、测试内容的可操作性及评估语言的可扩展性等方面,完善医疗健康对话式AI的评估范式。

【关键词】大语言模型;对话式AI;评价指标;评估体系;医疗健康
中图分类号:R-1　文献标识码:A　doi:10.3969/j.issn.1674-2982.2025.07.010

Evaluation paradigms for conversational AI in healthcare: Systematic review

LIAO Wei-zhen, HAN You-li, MA Cheng-yu
School of Public Health, Capital Medical University, Beijing 100069, China

【**Abstract**】Objective: This study aims to systematically review the current evaluation paradigms of conversational AI in healthcare and provide insights to facilitate the development of a comprehensive evaluation framework and methodological advancements in this field. Methods: A systematic review was conducted by searching the PubMed and Web of Science databases to analyze the existing evaluation paradigms of healthcare conversational AI, including evaluation subjects, assessment metrics, and evaluation methodologies. Results: A total of 60 studies were included in this review. The findings indicate that most evaluation subjects focus on general-purpose large language models. The assessment metrics cover five key dimensions: technical performance, information quality, clinical effectiveness, user experience, and ethics and safety. However, there were significant differences in the evaluation criteria used in existing studies. There were also issues such as a low degree of alignment between the evaluation questions and the application scenarios, as well as a lack of diversity in the roles of the evaluators. Conclusions: The current evaluation framework for healthcare conversational AI remains underdeveloped. Future improvements should focus on broadening model coverage, enhancing the comprehensiveness of evaluation indicators, standardizing evaluation methods, improving the operationalizability of test content, and expanding the scalability of evaluation languages.

【**Key words**】Large language model; Conversational AI; Evaluation index; Evaluation system; Healthcare

1 研究背景

对话式人工智能(Artificial Intelligence,AI),又称虚拟助手、聊天机器人或对话代理等,是能够识别并处理语音或文本信息,理解用户意图并以自然语言响应的人工智能应用。[1]早期的对话式AI基于规则驱动,理解能力有限[2],无法进行高质量的人机交互。近年来,随着深度学习和大语言模型(Large

* 基金项目:国家社会科学基金项目(24BGL273)
作者简介:廖委真(2000年—),女,硕士研究生,主要研究方向为数字健康、智慧医疗。E-mail:liaoweiizhen@163.com
通讯作者:马骋宇。E-mail:machengyu@ccmu.edu.cn

Language Model，LLM）的发展，对话式 AI 的语言理解与生成能力大幅度提升，使其在医疗健康领域展现出广泛的应用前景。[3]

医疗健康领域中的对话式 AI（以下简称“医疗健康对话式 AI”）以生成式人工智能（Generative AI）技术为核心，能够辅助处理多模态医学信息，理解复杂医学问题，广泛应用于患者问诊、临床决策支持、健康科普等多个领域。[4]近年来，医疗健康对话式 AI 发展迅猛。国际上，ChatGPT 已广泛用于医学评估，谷歌开发了 Med-PaLM2 医疗问答系统[5]；同时，GatorTron、BioGPT、BioMistral 等医疗垂类 LLM① 也相继问世。截至 2024 年 4 月，我国已有 117 款生成式 AI 产品完成网信办备案，并涌现出华佗 GPT、MedGPT、灵医大模型、岐黄问道大模型等一批医疗垂类 LLM。

医疗健康属于知识和技术密集型行业，其发展水平关乎公众健康和生命安全。对话式 AI 在医疗场景中的应用必须建立在可靠的质量评估基础之上，以确保所生成内容的有效性、可靠性与安全性。目前已有研究就模型性能、临床适用性、用户体验、伦理风险等维度开展了多方面的评估，但各研究的评价指标和方法存在较大异质性，缺乏统一规范。为此，本文采用系统综述方法，从评估对象、评价指标、评估方法学设计等角度系统梳理医疗健康对话式 AI 的研究进展，旨在为构建科学、完备的医疗健康对话式 AI 评估范式提供理论依据，助力对话式 AI 在医疗健康领域的规范、安全与可持续发展。

2 资料与方法

2.1 文献检索策略

本研究在 PubMed 和 Web of Science 数据库中以检索式（“chatbot” or “chatterbot” or “chat-bot” or “chatter bot” or “Conversational bot” or “Conversational AI” or “conversational agent” or “Conversational AI Automated Counseling Systems” or “ChatGPT” or “AI-chat-generated” or “Dialogue system” or “AI chat agent”）AND（“Large Language Model” or “generative AI” or “NLP” or “Machine Learning” or “Deep Learning” or “Neural Network” or “large-scale language model” or “transformer model”）AND（“health” or “healthcare” or “medical” or “medicine”）进行检索，检索时限截至 2024 年 10 月 10 日。文献检索期间在中文数据库中仅检索到 1 篇相关综述论文，测评类的实证研究较少，因此本研究未纳入中文文献。

2.2 文献筛选流程

为确保纳入研究文献的一致性，制定了统一明确的纳入排除标准。（1）纳入标准：①可获取论文全文；②英文文献；③2004 年之后发表。（2）排除标准：①非实证研究类型（如会议摘要、社论、信件等）；②只评估对话式 AI 的写作熟练度或其他与医学无关的表现；③评估过程不完整，只简单描述结果；④评估对话式 AI 的医学能力为次要主题；⑤评估内容或方法与其他研究高度重合。文献筛选过程如图 1 所示，最终从 1 413 篇文献中筛选出有代表性的文献 60 篇，作为本研究分析的核心文献样本。

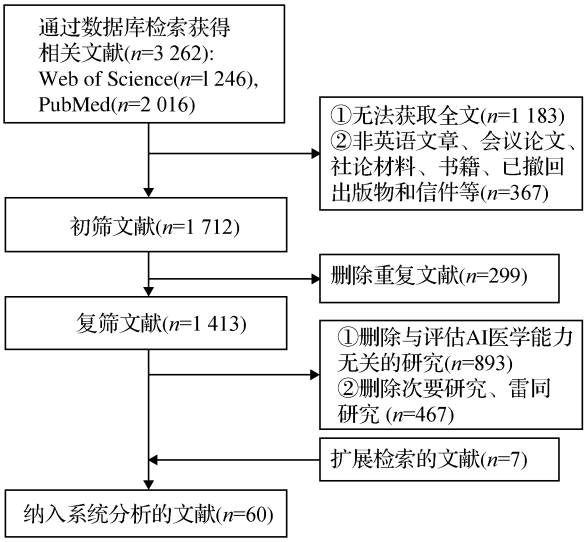

图 1　文献筛选流程

2.3 研究内容提取和分析

确定纳入文献后，由 1 名公共卫生背景的硕士研究生按照统一模版提取相关信息，内容包括评估 AI 的名称及数量、医学学科、应用场景、评估指标、评估方法、评估问题设置、对话设计及评估者构成等。为确保内容准确性，信息提取经指导教师复核，若有疑义由双方协商确定。提取完成后，研究小组进一步将提取内容归类为 3 大类别：（1）评价对象，包括医疗健康对话式 AI 的类型、应用学科和应用场景；（2）评价指标，总结归纳当前研究采用的评价指标；（3）评估过程中的方法学设计，包括评估问题的设计、提问方式、评估者类型等。最后，对每个类别中的信息进行描述性分析。

① 医疗垂类 LLM 是专门针对医疗领域特定需求，开发、设计和训练的人工智能大语言模型，通过海量医疗数据，如医学影像、电子病历、基因组数据等进行预训练，具备多模态信息处理能力，应用于疾病诊断、药物研发、个性化治疗、健康管理等场景。

3 研究结果

3.1 文献概况

纳入研究的60篇文献中，评价对象以通用LLM（n=46，76.67%）为主，垂类LLM占比较少（n=14，23.33%）。按照评估对象的数量差异，主要分为单个LLM评估、多个LLM比较评价、LLM与医学专家比较分析和其他（表1）。单个LLM评估以ChatGPT为主，多个LLM比较分析主要为ChatGPT与其它对话式AI（如Bard、Gemini、Claude、LLaMA等）的对比，LLM与医学专家的比较分析旨在探究AI与人类医学专家的差别。从评价研究所涉及的学科来看，主要包括全科（n=9）、内科（n=10）、外科（n=7）、五官科（n=7）、影像科（n=7）、临床辅助学科（n=5）等。

表1 医疗健康对话式AI评估对象分析

评估对象	评估内容
按照LLM类型	
通用LLM	GPT-3.5[6-8]、GPT-4.0[9-11]、GPT-4V[12-14]、LLaMA[15-16]、Bing[17-18]、Google Bard[17-19]、Gemini[20]、Claude2[21]等
垂类LLM	BioGPT[22]、BioMedGPT[23]、Med-PaLM[5]、AI-guide bot[24]、MOPH[25]、PJ博士[26]、AI营养师[27]、CPMI-ChatGLM[28]
按照评估对象数量	
单个LLM评估	GPT-3.5[6-8]、GPT-3.5turbo[29]、GPT-4.0[9-11]、GPT-4V[12]、BioMedGPT[23]、AI-guide bot[24]、MOPH[25]
多个LLM间对比	GPT-4和LLaMA-2[15]、GPT-3.5和GPT-4.0[30-35]、GPT-4和ERNIE Bot[36]、GPT-4和Gemini[37]
LLM与医学专家对比	GPT与医学专家[38]
其他	LLM在不同语言环境中的对比[17]

根据应用场景来分析，所选取的文献中，医疗健康对话式AI的服务对象包括患者、医生以及科研人员，其应用场景可以分为健康咨询、诊断与鉴别诊断、影像结果分析、医疗报告生成、临床决策建议、医学教育与培训等8个方面（表2）。

表2 医疗健康对话式AI应用场景分析

服务对象	应用场景	文献数量	文献来源
患者	健康咨询	26	MEYER等[20]、HE等[36]
医生	诊断与鉴别诊断	7	SHIEH等[33]、HIROSAWA等[39]
	影像结果分析	3	ZHANG等[23]
	医疗报告生成	5	RAMINEDI等[40]、LIU等[41]
	临床决策建议	16	LV等[19]、BALTA等[35]
	医学教育与培训	9	MING等[30]、ALMEIDA等[31]
科研人员	医药研发	2	BHATTACHARYYA等[42]
	其他	3	XU等[26]

3.2 评估维度

在医疗健康领域，对话式AI的评估方法呈现出多维度、多指标并存的特征。根据对纳入文献的系统梳理，本文将主流评估指标划分为五个一级维度（表3），分别为技术性能、信息质量、临床效果、用户体验、伦理与安全。根据各指标文献报告数量可见，多数研究将评价侧重点落在准确性、全面性等信息质量维度与技术性能维度，对临床效果和用户体验维度的关注度次之，对伦理与安全维度的关注度明显不足。

3.2.1 技术性能

技术性能评估主要用于衡量医疗健康对话式AI模型在语言生成、信息处理及任务完成能力方面的表现，是模型开发与应用的基础。相关指标可大致分为两类：分类性能指标与文本相似度指标。

表3 医疗健康对话式AI评估指标维度及报告数量

评估维度及报告数量	二级指标	报告数量	指标定义	文献来源
技术性能（25）	精确率	3	模型预测为正类的样本中，实际为正类的占比	HE等[16]、BENARY等[43]
	召回率	3	实际为正类的样本中，被模型正确预测为正类的占比	HE等[16]、LUO等[22]
	F1分数	5	精确率与召回率的调和平均值	LUO等[22]、SUN等[27]
	BLEU	4	生成文本和参考文本之间的匹配程度	RAMINEDI等[40]
	ROUGE	6	生成文本和参考文本之间的匹配程度	VAN VEEN等[44]
	BERTScore	4	生成文本和参考文本之间的匹配程度	LIU等[28]
信息质量（77）	质量	7	综合评估指标，衡量生成内容的整体质量	CHOI等[45]、ONDER等[46]、HUANG等[47]
	准确性	37	生成内容的事实正确性	DE VITO[10]、ALMEIDA等[31]

表 3　医疗健康对话式 AI 评估指标维度及报告数量（续）

评估维度及报告数量	二级指标	报告数量	指标定义	文献来源
	全面性	14	生成内容的完整性，应涵盖用户查询的所有关键点	GIANNAKOPOULOS 等[18]
	相关性	9	模型输出与用户提问的内容契合度	HE 等[36]、LIU 等[38]
	适当性	7	生成内容是否清晰、简洁、易于理解且准确、全面	SALLAM 等[17]、LV 等[19]
	稳定性	3	多次回答同一问题的一致程度	SUÁREZ 等[11]
临床效果（19）	有效性	9	对用户健康决策产生帮助的程度	LIU 等[38]
	诊断准确性	5	能否给出准确的诊断结果	HIROSAWA 等[39]、LEYPOLD 等[48]
	临床推理能力	2	对生成内容的解释和推理能力	MADRID-GARCÍA 等[49]、SINGHAL 等[5]
	与医学专家一致性	3	生成内容或判断结果与人类医学专家的一致程度	AYERS 等[6]
用户体验（16）	可读性	7	是否易于阅读和理解	LV 等[19]、GHANEM 等[21]
	同理心	3	对用户的情感支持程度	HE 等[36]、AYERS 等[6]
	清晰度	6	是否清晰明了、易于理解	MEYER 等[20]
伦理与安全（9）	无害性	1	可能造成伤害的程度和可能性	LV 等[19]
	安全性	4	是否具有潜在风险	HE[16]
	偏见	2	生成内容存在系统性偏见，如种族或性别偏见	SINGHAL 等[5]
	捏造与幻觉	2	生成内容存在虚构的信息或数据	BHATTACHARYYA 等[42]、MENZ 等[50]

分类性能指标包括精确率（Precision）、召回率（Recall）和 F1 分数（F1-score），常用于判断模型在疾病识别、信息匹配等结构化任务中的表现。精确率和召回率分别评估模型的准确性和检出能力；F1 分数则综合反映模型的整体分类性能。

文本相似度指标更适用于生成任务，评估模型输出文本与参考答案之间在词汇或语义层面的匹配程度。BLEU（Bilingual Evaluation Understudy）、ROUGE（Recall-Oriented Understudy for Gisting Evaluation）和 BERT Score 是最常见的文本相似度指标。BLEU 与 ROUGE 侧重评估生成文本与参考答案之间的词汇重合率[28]，而 BERT Score 引入上下文语义嵌入，可更精确地衡量语义相似性[51]。上述指标常用于评价医学报告解读[40]、用药建议[28]等信息的生成质量。

3.2.2 信息质量

信息质量维度旨在评估对话式 AI 生成内容的医学信息价值和语义合理性。该维度既包含综合性评价指标（如质量），也包括若干个细分指标，用于评估信息质量的不同方面，包括准确性、全面性、相关性、适当性和稳定性等。

质量反映了模型输出内容的整体水平，通常包含准确性、可靠性等多种内涵。评估该指标的常用工具包括改良版 DISCERN 评分量表[21]和全球质量评分量表（Global Quality Scale，GQS）[46]等。前者主要用于衡量健康信息的可信性和信息质量，后者则从内容完整性、实用性和清晰度等方面进行总体评价。

准确性指生成内容的事实正确性，是评估模型能力的核心指标。多数研究以医疗指南、专业医生意见作为参考标准，对模型在医学问答、图像分类、临床试验匹配等任务中的准确程度进行评价。[52]

全面性又称完整性、充分性，反映模型是否覆盖问题的所有关键点，避免遗漏重要信息[19]，通常通过核对模型输出内容覆盖预设关键要素的百分比进行评估。例如 SALLAM 等人[17]使用 CLEAR 工具对 AI 回答传染性疾病相关问题的完整性进行打分，以判断生成内容的信息覆盖范围。

相关性强调模型输出与提问之间的内容契合度，确保对特定问题生成的信息具有针对性。[36]有研究显示，GPT-3.5 在老年健康咨询中的相关性评分高于人类医生，展现出良好的语境理解与聚焦能力。[7]

适当性常用于评估生成的内容是否清晰、简洁、易于理解[53]，有时也可被扩展用于衡量内容的准确性与全面性。例如，BALTA 等人[35]以该指标考察 AI 在重症监护领域的回答表现，将与事实不符、违背指南或存在潜在危害的回答判定为不当。

稳定性亦称一致性、可重复性，用于评估模型在不同时间、不同环境下输出结果的一致程度[19]，反映其内部逻辑的可靠性。

3.2.3 临床效果

临床效果评估主要衡量对话式 AI 在真实医疗场

景中的实用价值，关注其在疾病诊断、治疗建议、安全性等方面的表现。常用指标包括有效性、诊断准确性、临床推理能力、与医学专家一致性等。

有效性亦称有用性，用于评估模型输出对患者健康管理或临床决策的实际帮助程度[38]，该指标综合考虑了信息的相关性、准确性与可行性[36]。例如，有研究利用该指标评价AI在消化系统问答[54]和医学摘要生成[41]中的表现，验证其是否能够提供清晰、可执行的健康建议。

诊断准确性反映模型能否基于病症信息做出正确的疾病判断。例如，在SHIEH等人[33]的研究中，若AI给出的前三项诊断包含实际诊断，即视为回答准确，有助于衡量模型在辅助诊断任务中的可靠性。

临床推理能力反映模型能否提供合理的推理过程与支撑答案的可信证据。例如，MADRID-GARCÍA等人[49]的研究要求AI对每个问题的答案进行解释和推理，若推理逻辑清晰且证据正确，则认为其临床推理能力较好。

与医学专家一致性用于衡量AI在诊断或治疗建议上与医学专家的相符程度。通过对比AI与医生在临床决策[38]或公众咨询[6]中的回答，有助于评估AI在临床应用中的替代或辅助潜力。

3.2.4 用户体验

用户体验评估关注模型在实际交互过程中的可用性与人机互动质量，是衡量其服务友好性的重要维度。主要指标包括可读性、同理心和清晰度等。

可读性用于衡量生成内容易于阅读和理解的程度。可读性高有助于减少医疗信息获取的障碍，提升使用便捷性。例如，LIU等人[41]曾采用可读性指标评估BERTSUM、BART、ChatGPT3.5总结医学对话的性能。

同理心体现模型对用户情绪的感知与回应能力。具有同理心的AI应表现出尊重、耐心、同情心，真诚地服务用户，改善他们的身心健康。[36]

清晰度是指模型表达信息的明确程度。清晰、明确地回答能够快速解答用户疑问、减少歧义。如ZHU等人[55]使用该指标评估AI回答患者问题的清晰程度，验证其在医患沟通场景下的实际可行性。

3.2.5 伦理与安全

医疗健康对话式AI除满足技术性能要求外，还需严格遵守伦理规范，保障用户安全，防范潜在风险。相关评估指标包括无害性、安全性、偏见、捏造与幻觉等。

无害性可评估生成内容造成伤害的程度和可能性。LV等人[19]的研究通过口腔健康问答测试多种模型，验证其回复的安全边界。

安全性关注模型是否提供具有潜在风险或不利影响的信息。[16]LIU等人[28]的研究利用安全性指标评估模型内容是否可能误导用户并危害其健康。

偏见指生成内容中存在的系统性偏见或歧视，可能导致医疗不公现象，该指标常用于评价AI的回答是否存在人口统计学偏见，如性别偏见等。[5]

捏造与幻觉是指生成内容包含虚构的信息或数据，或看似合理但不存在的事实。已有研究发现多个模型可生成虚构的医学建议或参考资料，需引起高度重视。[50]

3.3 评估过程中的方法学设计

3.3.1 评估问题的设计

医疗健康对话式AI的质量评估离不开科学、合理的评估问题集。本文对纳入文献的评估问题设计进行总结回顾，从问题来源、问题语言、问题数量和问题题型4个方面进行分析（表4）。

表4 测评问题设计分析结果

问题设计	报告数量	文献来源
问题来源		
作者设计	16	LIM等[29]、LV等[19]
开放性网站	10	LIU等[38]、GIORGI等[15]
临床案例记录	4	BUSCH等[13]、WANG等[56]
医学考试题库	8	MING等[30]
测试集	9	XU等[26]
其他	13	SALLAM等[17]
问题语言		
英语	23	BUSCH等[13]
其他语言	7	MADRID-GARCÍA等[49]
未报告	32	GIORGI等[15]、LV等[19]
问题数量		
≤100	32	LIU等[38]
＞100	26	BUSCH等[13]、DE VITO等[10]
问题题型		
开放式问题	46	LAHAT等[57]
选择题	5	ALMEIDA等[31]
混合题型	7	ZHANG等[23]、ZHENG等[25]

当前评价研究的问题来源呈现多样化趋势。在纳入的研究中，使用作者自编问题的有16项，来自开放网站10项，专业测试集9项，医学考试题库8项，临床案例记录4项，另有13项来源于患者意见、机器人生成或医学协会等。较多研究倾向采用结构化测试

集进行评估，其覆盖场景广泛、数量充足，便于分析模型优劣。

在语言方面，23 项研究使用英文，中文 3 项，其余使用德语、西班牙语等，另有 32 项未报告语言信息。问题数量从个位数到上千不等，差异显著，约 55% 研究问题数量低于或等于 100。就题型而言，开放式问题最常见（n=46），其余使用选择题或混合题型。

3.3.2 测评过程的设计

测评的提问方式和提示语设计对结果有显著影响。11 项研究明确问题在独立对话框中提问，以避免前文干扰，其余多数研究未说明独立对话情况。提示语方面，24 项研究通过设定 AI 角色（如医生、健康顾问）明确任务背景，以提高模型回应的针对性与专业性；36 项未使用或未报告相关提示内容，提示语标准尚缺乏统一规范。

3.3.3 评估指标的计算方式

在医疗健康对话式 AI 的开发与验证过程中，标准化评估方法尤为关键。目前，国内外已建立多种基准测试（Benchmark），用于在统一任务和数据集下评估模型性能，如 MedBench[58]、CMB[59]等。不同的基准测试往往包含不同的评估指标，其计算方式可分为自动评估和人工评估两种。

自动评估是一种基于数学公式和统计方法的量化评估方式，一般通过 API 接入或上传模型答案进行自动评估。例如，准确率通过计算正确回答的比例得出[32]；F1 分数可通过 Python 等工具计算[27]；可读性则常借助在线平台（如 Readable）计算 Flesch-Kincaid、Gunning Fog 等指标[19,60]，以量化分析健康信息的易读性。

人工评估主要依赖专家或研究人员对模型回答进行主观评分。评估工具多采用李克特量表（如 5 分制）对准确性、全面性、同理心等指标进行定量赋分。然而，人工评估存在主观性强、缺乏统一的量化标准等问题，易影响结果的解释与比较。[11]

4 讨论与建议

4.1 调整评估对象，提升医疗垂直领域关注度

当前大多数评估研究集中于 ChatGPT 等通用 LLM，对医疗垂类 LLM 的关注度尚显不足。虽然通用 LLM 在自然语言理解与生成方面表现优异，但由于其训练数据来源广泛，缺乏针对性医学知识结构支撑，容易出现幻觉，影响了生成医疗信息的安全性与可靠性。相较之下，医疗垂类 LLM 的模型训练更具专业性，其医疗应用能力亟需更充分的检验。未来可通过系统比较通用 LLM 与医疗垂类 LLM 在不同任务中的能力边界与适配性能，为模型选择与临床部署提供更具针对性的实证依据。

4.2 构建综合性评估指标体系，加强伦理和安全维度考量

目前，医疗健康对话式 AI 的评估研究所采用的指标差异显著，多数研究侧重于评估模型的技术性能和信息质量，对用户体验维度、伦理与安全维度的考察相对缺乏。政策法规层面，各国对 AI 的伦理和安全问题非常关注，欧盟《人工智能法案》明确要求高风险医疗 AI 需通过伦理风险评估；中国《人工智能伦理治理标准化指南》亦强调“患者隐私安全、公平”等伦理准则。为此，应重视患者体验、伦理与安全等核心维度的纳入，构建更具综合性、多维度的医疗健康对话式 AI 评估指标体系，以提升评估工作的科学性和规范性。

4.3 统一评估方法，提升评估客观性与规范性

当前医疗健康对话式 AI 的评估方法尚未形成统一标准。首先，评估主体相对单一，主要以医生为主，缺乏患者、卫生监管者等多利益主体的参与。第二，许多研究结果未披露具体的方法学测评细节，如测评模型版本、对话生成日期等，影响了测评结果的透明度与可信性。第三，人工评估主观性较强，打分结果易产生偏移。建议在人工评估基础上适当引入结构化工具与自动化指标，减少评估者主观判断的影响，提升结果的客观性。同时，纳入多元评估主体（如管理者、患者等），提升评估主体的多样性。最后，推动评估过程的标准化建设，包括明确评分标准、规范评估者培训流程，并详尽报告关键方法学信息，从而提高研究可比性与可重复性。

4.4 拓宽测评问题领域，增强现实应用能力

随着模型能力提升，测评题型已由封闭式医学选择题扩展至开放式问答、临床情境推理、多模态识别等多样化形式。然而，测评问题内容设计与实际应用场景之间仍存在一定脱节。在实践中，大多数测试集仍来源于标准化考试题库或医学教科书，与真实就医场景存在偏差。此外，测评问题往往集中于单一医学学科，忽视了临床中多学科协同、共病管理等实际情境，难以反映 AI 在处理复杂医疗任务时的综合能力。建议未来的测试设计应更加贴合真实

医疗场景，纳入真实问诊记录问题，整合多学科、多情境和多模态任务，从而更全面地评估模型的实际能力。

4.5 扩充评估语言环境，促进医疗人工智能普惠发展

当前医疗健康对话式AI的评估研究主要集中于英语语境，中文以及其他非英语语种的研究相对滞后。这可能会限制模型在多语言、多文化背景下的应用，影响医疗服务的可及性与公平性。建议未来加强多语言环境下的模型研究与评估，提升对话式AI的全球适应性与服务覆盖范围，促进医疗AI技术的公平普惠发展。

5 小结

医疗健康对话式AI的评估体系仍处于发展初期，面对模型复杂度增加、应用场景多元化以及伦理规范需求提升等新挑战，现有评估框架显得相对滞后。未来应从兼顾各类评估对象、设计综合性评估指标体系、建设标准化评估方法、提升测试内容与现实场景的贴近程度、扩充评估语言环境等方面，推进评估范式的科学化与规范化，推动对话式AI从“语言生成工具”迈向“可信医疗助手”，在保障患者安全与权益的基础上，助力医疗体系的高质量、可持续发展。

作者贡献：廖委真负责收集文献资料、撰写论文初稿并修改论文，韩优莉负责论文修改与完善，马骋宇负责论文选题、设计研究框架、修改论文并定稿。

作者声明本文无实际或潜在的利益冲突。

参 考 文 献

[1] GKINKO L, ELBANNA A. The appropriation of conversational AI in the workplace: A taxonomy of AI chatbot users [J]. International Journal of Information Management, 2023, 69: 102568.

[2] CAR L T, DHINAGARAN D, KYAW B M, et al. Conversational agents in health care: Scoping review and conceptual analysis[J]. J Med Internet Res, 2020, 22(8): e17158.

[3] XUE J, ZHANG B, ZHAO Y, et al. Evaluation of the current state of chatbots for digital health: Scoping Review [J]. J Med Internet Res, 2023, 25: e47217.

[4] LAYMOUNA M, MA Y, LESSARD D, et al. Roles, users, benefits, and limitations of chatbots in health care: Rapid review[J]. J Med Internet Res, 2024, 26: e56930.

[5] SINGHAL K, AZIZI S, TU T, et al. Large language models encode clinical knowledge[J]. Nature, 2023, 620 (7972): 172-180.

[6] 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 Intern Med, 2023, 183(6): 589-596.

[7] MOORE I, MAGNANTE C, EMBRY E, et al. Doctor AI? A pilot study examining responses of artificial intelligence to common questions asked by geriatric patients[J]. Frontiers in artificial intelligence, 2024, 7: 1438012.

[8] PENG W, FENG Y, YAO C, et al. Evaluating AI in medicine: a comparative analysis of expert and ChatGPT responses to colorectal cancer questions [J]. Sci Rep, 2024, 14(1): 2840.

[9] ARMITAGE R. Performance of GPT-4 in Membership of the royal college of paediatrics and child health-style examination questions [J]. BMJ paediatrics open, 2024, 8 (1): e002575.

[10] DE VITO A, COLPANI A, MOI G, et al. Assessing ChatGPT′s potential in HIV prevention communication: A comprehensive evaluation of accuracy, completeness, and inclusivity[J]. AIDS and behavior, 2024, 28(8): 2746-2754.

[11] SUÁREZ A, JIMÉNEZ J, LLORENTE DE PEDRO M, et al. Beyond the scalpel: Assessing ChatGPT′s potential as an auxiliary intelligent virtual assistant in oral surgery[J]. Computational and structural biotechnology journal, 2024, 24: 46-52.

[12] HIROSAWA T, HARADA Y, TOKUMASU K, et al. Evaluating ChatGPT-4′s diagnostic accuracy: Impact of visual data integration [J]. JMIR medical informatics, 2024, 12: e55627.

[13] BUSCH F, HAN T, MAKOWSKI M R, et al. Integrating text and image analysis: Exploring GPT-4V′s capabilities in advanced radiological applications across subspecialties [J]. J Med Internet Res, 2024, 26: e54948.

[14] ROJAS M, ROJAS M, BURGESS V, et al. Exploring the performance of chatgpt versions 3.5, 4, and 4 with vision in the chilean medical licensing examination: observational study[J]. JMIR medical education, 2024, 10: e55048.

[15] GIORGI S, ISMAN K, LIU T, et al. Evaluating generative AI responses to real-world drug-related questions [J]. Psychiatry research, 2024, 339: 116058.

[16] HE Z, BHASURAN B, JIN Q, et al. Quality of answers of generative large language models versus peer users for interpreting laboratory test results for lay patients: Evaluation study [J]. J Med Internet Res, 2024, 26: e56655.

[17] SALLAM M, AL-MAHZOUM K, ALSHUAIB O, et al.

Language discrepancies in the performance of generative artificial intelligence models: An examination of infectious disease queries in English and Arabic[J]. BMC infectious diseases, 2024, 24(1): 799.

[18] GIANNAKOPOULOS K, KAVADELLA A, AAQEL SALIM A, et al. Evaluation of the performance of generative AI large language models ChatGPT, Google Bard, and Microsoft Bing Chat in supporting evidence-based dentistry: comparative mixed methods study[J]. J Med Internet Res, 2023, 25: e51580.

[19] LV X, ZHANG X, LI Y, et al. Leveraging large language models for improved patient access and self-management: Assessor-Blinded comparison between Expert- and AI-generated content[J]. J Med Internet Res, 2024, 26: e55847.

[20] MEYER A, SOLEMAN A, RIESE J, et al. Comparison of ChatGPT, Gemini, and Le Chat with physician interpretations of medical laboratory questions from an online health forum[J]. Clinical Chemistry and Laboratory Medicine, 2024, 62(12): 2425-2434.

[21] GHANEM Y K, ROUHI A D, AL-HOUSSAN A, et al. Dr. Google to Dr. ChatGPT: assessing the content and quality of artificial intelligence-generated medical information on appendicitis[J]. Surgical endoscopy, 2024, 38(5): 2887-2893.

[22] LUO R, SUN L, XIA Y, et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining[J]. Briefings in bioinformatics, 2022, 23(6): 1-11.

[23] ZHANG K, ZHOU R, ADHIKARLA E, et al. A generalist vision - language foundation model for diverse biomedical tasks[J]. Nature Medicine, 2024(30): 3129-3141.

[24] LEE J W, YOO I S, KIM J H, et al. Development of AI-generated medical responses using the ChatGPT for cancer patients[J]. Computer Methods and Programs in Biomedicine, 2024, 254: 108302.

[25] ZHENG C, YE H, GUO J, et al. Development and evaluation of a large language model of ophthalmology in Chinese[J]. The British Journal of Ophthalmology, 2024, 108(10): 1390-1397.

[26] XU J, LU L, PENG X, et al. Data Set and Benchmark (MedGPTEval) to evaluate responses from Large Language Models in medicine: Evaluation development and validation [J]. JMIR Medical Informatics, 2024, 12: e57674.

[27] SUN H, ZHANG K, LAN W, et al. An AI dietitian for Type 2 diabetes mellitus management based on Large Language and Image Recognition Models: Preclinical concept validation study[J]. J Med Internet Res, 2023, 25: e51300.

[28] LIU C, SUN K, ZHOU Q, et al. CPMI-ChatGLM: parameter-efficient fine-tuning ChatGLM with Chinese patent medicine instructions[J]. Sci Rep, 2024, 14(1): 6403.

[29] LIM D Y Z, KE Y H, SNG G G R, et al. Large Language Models in anaesthesiology: Use of ChatGPT for American society of anesthesiologists physical status classification [J]. British Journal of Anaesthesia, 2023, 131(3): e73-e75.

[30] MING S, GUO Q, CHENG W, et al. Influence of model evolution and system roles on ChatGPT's Performance in Chinese medical licensing exams: Comparative study[J]. JMIR Medical Education, 2024, 10: e52784.

[31] ALMEIDA L C, FARINA E, KURIKI P E A, et al. Performance of ChatGPT on the Brazilian radiology and diagnostic imaging and mammography board examinations [J]. Radiology Artificial Intelligence, 2024, 6(1): e230103.

[32] HSIEH C H, HSIEH H Y, LIN H P. Evaluating the performance of ChatGPT-3.5 and ChatGPT-4 on the Taiwan plastic surgery board examination[J]. Heliyon, 2024, 10(14): e34851.

[33] SHIEH A, TRAN B, HE G, et al. Assessing ChatGPT 4.0's test performance and clinical diagnostic accuracy on USMLE STEP 2 CK and clinical case reports[J]. Sci Rep, 2024, 14(1): 9330.

[34] RAO A, KIM J, KAMINENI M, et al. Evaluating GPT as an adjunct for radiologic decision making: GPT-4 Versus GPT-3.5 in a breast imaging pilot[J]. Journal of the American College of Radiology, 2023, 20(10): 990-997.

[35] BALTA K Y, JAVIDAN A P, WALSER E, et al. Evaluating the appropriateness, consistency, and readability of ChatGPT in critical care recommendations [J]. Journal of Intensive Care Medicine, 2024 , 40(2): 184-190.

[36] HE W, ZHANG W, JIN Y, et al. Physician versus Large Language Model Chatbot responses to web-based questions from autistic patients in Chinese: Cross-sectional comparative analysis[J]. J Med Internet Res, 2024, 26: e54706.

[37] HAIDER S A, PRESSMAN S M, BORNA S, et al. Evaluating Large Language Model (LLM) performance on established breast classification systems[J]. Diagnostics (Basel, Switzerland), 2024, 14(14): 1491.

[38] LIU S, WRIGHT A P, PATTERSON B L, et al. Using AI-generated suggestions from ChatGPT to optimize clinical decision support[J]. Journal of the American Medical Informatics Association, 2023, 30(7): 1237-1245.

[39] HIROSAWA T, HARADA Y, MIZUTA K, et al.

Diagnostic performance of generative artificial intelligences for a series of complex case reports[J]. Digital health, 2024, 10: 1-12.

[40] RAMINEDI S, SHRIDEVI S, WON D. Multi-modal transformer architecture for medical image analysis and automated report generation[J]. Sci Rep, 2024, 14(1): 19281.

[41] LIU Y, JU S, WANG J. Exploring the potential of ChatGPT in medical dialogue summarization: a study on consistency with human preferences[J]. BMC medical informatics and decision making, 2024, 24(1): 75.

[42] BHATTACHARYYA M, MILLER V M, BHATTACHARYYA D, et al. High Rates of Fabricated and Inaccurate References in ChatGPT-Generated Medical Content[J]. Cureus, 2023, 15(5): e39238.

[43] 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.

[44] VAN VEEN D, VAN UDEN C, BLANKEMEIER L, et al. Adapted Large Language Models can outperform medical experts in clinical text summarization[J]. Nat Med, 2024, 30(4): 1134-1142.

[45] CHOI J, OH A R, PARK J, et al. Evaluation of the quality and quantity of artificial intelligence-generated responses about anesthesia and surgery: using ChatGPT 3.5 and 4.0[J]. Frontiers in Medicine, 2024, 11: 1400153.

[46] ONDER C E, KOC G, GOKBULUT P, et al. Evaluation of the reliability and readability of ChatGPT-4 responses regarding hypothyroidism during pregnancy[J]. Sci Rep, 2024, 14(1): 243.

[47] HUANG A S, HIRABAYASHI K, BARNA L, et al. Assessment of a Large Language Model's responses to questions and cases about glaucoma and retina management[J]. JAMA Ophthalmology, 2024, 142(4): 371-375.

[48] LEYPOLD T, LINGENS L F, BEIER J P, et al. Integrating AI in lipedema management: Assessing the efficacy of GPT-4 as a consultation assistant[J]. Life (Basel, Switzerland), 2024, 14(5): 646.

[49] MADRID-GARCÍA A, ROSALES-ROSADO Z, FREITES-NUñEZ D, et al. Harnessing ChatGPT and GPT-4 for evaluating the rheumatology questions of the Spanish access exam to specialized medical training[J]. Sci Rep, 2023, 13(1): 22129.

[50] MENZ B D, KUDERER N M, BACCHI S, et al. Current safeguards, risk mitigation, and transparency measures of large language models against the generation of health disinformation: repeated cross sectional analysis[J]. BMJ (Clinical research ed), 2024, 384: e078538.

[51] NAZI Z A, PENG W. Large Language Models in healthcare and medical domain: A review[J]. 2024, 11(3): 57.

[52] ZHANG K, ZHOU R, ADHIKARLA E, et al. A generalist vision-language foundation model for diverse biomedical tasks[J]. Nat Med, 2024,30:3129-3141.

[53] SALLAM M, BARAKAT M, SALLAM M. Pilot testing of a tool to standardize the assessment of the quality of health information generated by Artificial Intelligence-based Models[J]. Cureus, 2023, 15(11): e49373.

[54] LAHAT A, SHACHAR E, AVIDAN B, et al. Evaluating the utility of a Large Language Model in answering common patients' gastrointestinal health-related questions: Are we there yet?[J]. Diagnostics (Basel, Switzerland), 2023, 13(11): 1950.

[55] ZHU L, RONG Y, MCGEE L A, et al. Testing and validation of a custom retrained Large Language Model for the supportive care of HN patients with external knowledge base[J]. Cancers, 2024, 16(13): 2311.

[56] WANG Z, ZHANG Z, TRAVERSO A, et al. Assessing the role of GPT-4 in thyroid ultrasound diagnosis and treatment recommendations: enhancing interpretability with a chain of thought approach[J]. Quantitative imaging in medicine and surgery, 2024, 14(2): 1602-1615.

[57] LAHAT A, SHARIF K, ZOABI N, et al. Assessing Generative Pretrained Transformers (GPT) in clinical decision-making: Comparative analysis of GPT-3.5 and GPT-4[J]. J Med Internet Res, 2024, 26: e54571.

[58] LIU M, HU W, DING J, et al. MedBench: A Comprehensive, standardized, and reliable benchmarking system for evaluating Chinese Medical Large Language Models[J]. Big Data Mining and Analytics, 2024, 7(4): 1116-1128.

[59] WANG X, CHEN G, DINGJIE S, et al. CMB: A comprehensive medical benchmark in Chinese, Mexico City, Mexico[C]. Association for Computational Linguistics, 2024.

[60] ÖMÜR ARÇA D, ERDEMIR İ, KARA F, et al. Assessing the readability, reliability, and quality of artificial intelligence chatbot responses to the 100 most searched queries about cardiopulmonary resuscitation: An observational study[J]. Medicine, 2024, 103(22): e38352.

[收稿日期:2025-03-31 修回日期:2025-05-22]
(编辑 赵晓娟)