中文大语言模型在乳腺癌相关淋巴水肿护理问答中的性能比较

沈傲梅[1]，李温荣[2]，武佩佩[1]，文翠菊[3]，刘飞[4]，张紫娟[5]，弓政[5]，汪秦羽[5]，杨盼[5]，胡倩[5]，强万敏[1]，路潜[5]
（1.天津医科大学肿瘤医院 护理部，天津 300000；2.首都医科大学附属北京世纪坛医院 淋巴外科，北京 100038；
3.北京大学肿瘤医院暨北京市肿瘤防治研究所 乳腺癌预防治疗中心，北京 100142；
4.北京大学第一医院 护理部，北京 100034；5.北京大学 护理学院，北京 100191）

【摘要】 目的 评价国内 5 种中文大语言模型（large language models，LLMs）在乳腺癌相关淋巴水肿常见问题问答中的综合表现，为其应用及优化提供依据。**方法** 基于 LLMs、小组讨论和专家意见确定 100 个乳腺癌相关淋巴水肿的常见问题，分别由 3 名护理硕士生将问题输入 5 种 LLMs 模拟咨询，邀请 5 位专家从整体质量、准确性、全面性方面评估模型表现，以字符数评价应答的简洁性，分析模型的性能表现。采用组内相关系数（intraclass correlation coefficient，ICC）评价专家间一致性。**结果** 5 位专家评价者间一致性中等（$ICC=0.594$）。5 种 LLMs 综合表现均较好，“豆包”的整体质量和准确性评分均高于其他模型，差异有统计学意义（均 $P<0.05$）；“豆包”与“通义千问”的全面性评分差异无统计学意义（$P>0.05$）；二者评分均高于其他模型，差异有统计学意义（均 $P<0.05$）；“DeepSeek”和“文心一言”的字符数低于其他模型，差异均有统计学意义（均 $P<0.05$）。**结论** 以“豆包”为代表的 LLMs 在乳腺癌患者淋巴水肿相关护理问答的模拟咨询场景中显示出应用潜力，可进一步评价其在乳腺癌相关淋巴水肿预防管理中的应用效果。

【关键词】 大语言模型；乳腺癌；淋巴水肿；整体质量；全面性；准确性

DOI：10.3969/j.issn.2097-1826.2026.02.005

【中图分类号】 R473.73　**【文献标识码】** A　**【文章编号】** 2097-1826(2026)02-0020-04

Performance Comparison of Chinese Large Language Models in Answering Nursing Questions Related to Breast Cancer-associated Lymphedema

SHEN Aomei[1]，LI Wenrong[2]，WU Peipei[1]，WEN Cuiju[3]，LIU Fei[4]，ZHANG Zijuan[5]，GONG Zheng[5]，WANG Qinyu[5]，YANG Pan[5]，HU Qian[5]，QIANG Wanmin[1]，LU Qian[5]（1.Department of Nursing，Tianjin Medical University Cancer Institute and Hospital，Tianjin 300000，China；2.Department of Lymphatic Surgery，Beijing Shijitan Hospital，Capital Medical University，Beijing 100038，China；3.Breast Cancer Prevention and Treatment Center，Peking University Cancer Hospital & Institute，Beijing 100142，China；4.Department of Nursing，Peking University First Hospital，Beijing 100034，China；5.School of Nursing，Peking University，Beijing 100191，China）
Corresponding author：LU Qian，Tel：010-82805283

【Abstract】 Objective To evaluate the comprehensive performance of five domestic Chinese Large Language Models (LLMs) in answering frequently asked questions (FAQs) regarding breast cancer-associated lymphedema (BCRL)，and to provide a basis for their application and optimization.**Methods** Based on LLMs，group discussions，and expert opinions，100 FAQs about BCRL were identified.Three students with Master Degree in nursing input these questions into the five models for simulated consultations.Five lymphedema nursing experts were invited to evaluate the model performance in terms of information quality，accuracy，and comprehensiveness.The conciseness of responses was evaluated by word character count.The differences in performance among the models were analyzed.The intraclass correlation coefficient (ICC) was used to evaluate inter-expert consistency.**Results** The consistency among the five expert evaluators was moderate ($ICC=0.594$).The comprehensive performance of the five models was generally good.Doubao received significantly higher scores than the other models in terms of information quality and accuracy ($P<0.05$).There was no statistically significant difference in comprehensiveness scores between Doubao and Tongyi Qianwen($P>0.05$)，but both scored significantly higher than the other models ($P<0.05$).DeepSeek and Wenxin Yiyan had significantly lower character counts than the other models (all $P<0.05$).**Conclusions** LLMs represented by Doubao show application potential in simulated consultation scenarios for answering nursing questions related to BCRL.Future research should further evaluate their clinical application effects in the preventive management of BCRL.

【Key words】 large language model；breast cancer；lymphedema；information quality；comprehensiveness；accuracy

[Mil Nurs，2026，43(02)：20-23]

【收稿日期】 2025-06-30　**【修回日期】** 2026-01-16
【基金项目】 国家自然科学基金面上项目（72174011）；天津市医学重点学科（专科）建设项目（TJYXZDXK 011A）
【作者简介】 沈傲梅，博士，主管护师，电话：022-23524903
【通信作者】 路潜，电话：010-82805283

乳腺癌相关淋巴水肿（breast cancer-related lymphedema，BCRL）是乳腺癌治疗后的常见慢性并发症，具有风险终身存在、难以治愈等特点，早期识

别、持续管理尤为关键[1]。相关指南[2]建议,在乳腺癌患者长期随访中持续开展 BCRL 健康教育。但传统健康教育依赖医护人力,难以长期、及时满足患者的健康信息需求。近年来,人工智能技术迅速发展,大语言模型(large language models,LLMs)在自然语言生成、知识问答等任务中表现出潜力[3]。国内相继推出“文心一言”“通义千问”“智谱清言”“豆包”“DeepSeek”等 LLMs,并逐步拓展至医学健康场景。然而,LLMs 存在“幻觉”问题,其生成内容的真实性、全面性和专业性仍存在较大变异,尤其在处理医学相关问题时更为显著[4]。目前,针对国内 LLMs 的系统性评估相对不足[5]。本研究旨在围绕 BCRL 护理常见问题,系统比较 5 种国内主流 LLMs 在整体质量、全面性、准确性及简洁性方面的表现,为其在患者健康教育中的应用及优化提供依据。

1 资料与方法

1.1 问题设计 选取“ChatGPT-4o”(国际)和“DeepSeek”(国内)各生成 100 个 BCRL 常见问题,以保证问题的全面性。采用统一提示词“乳腺癌患者关于淋巴水肿的 100 个常见问题是什么”进行提问,提示词设计参考既往使用 LLMs 开展医疗问答研究的常用做法[6];并结合临床实践中总结的患者常见问题,由 1 名具有 BCRL 实践和研究经验的研究者人工审阅基于语义的一致性:合并语义相同问题,保留表述更具体者,剔除明显偏离 BCRL 护理范畴的问题,形成问题清单初稿。问题分类依据其主要信息或护理目标,遵循“单一主导内容优先”原则,即每个问题仅归入其核心关注内容最突出的一个类别。然后,邀请 5 位 BCRL 领域专家对问题的科学性、合理性、全面性、分类准确性、问题表述的清晰性进行独立审阅和修订。汇总专家意见并经研究组讨论修改(补充12 条、修改 21 条、删除 12 条),最终形成 100 个常见问题清单,包括疾病基本信息(16 个)、病因与风险因素(15 个)、症状与诊断(12 个)、预防(20 个)、治疗与预后(11 个)、康复锻炼(8 个)、日常生活护理(18 个)。

1.2 模型测试 基于公众可及性、医疗领域应用情况及代表性,经系统比较选取“DeepSeek”“通义千问”“智谱清言”“豆包”及“文心一言”5 个主流 LLMs,见表 1。由 3 名护理研究生按统一标准化提示词模板进行问题检索。提示词格式固定为:“我有一个关于乳腺癌相关淋巴水肿的问题:[具体问题]”,其中“具体问题”为专家审阅修订后的问题清单条目。所有模型均使用相同的提示词模板与问题内容输入,并统一设为“深度思考、联网搜索”模式,以控制输入表述差异对输出的影响。查询均在 2025 年 5 月 9—12 日完成,期间所使用模型未发生公开版本更新。每个问题仅查询 1 次,查询后重置对话以避免记忆偏倚,最终将所有问题答案整理汇总于预设结果表中。

表 1 5 种 LLMs 的基本情况

模型名称	研发机构	发布时间	开源情况	版本号
DeepSeek	深度求索	2024 年 1 月 5 日	部分开源	DeepSeek V3
通义千问	阿里云	2023 年 4 月 11 日	全面开源	Qwen3
智谱清言	清华/智谱	2023 年 8 月 31 日	部分开源	ChatGLM4
豆包	字节跳动	2024 年 5 月 15 日	未开源	Doubao-1.5-thinking
文心一言	百度	2023 年 3 月 16 日	未开源	X1 Turbo

1.3 专家评价 邀请 5 位具有 BCRL 临床实践及研究经验的护理专家(1 名博士、4 名硕士;2 名副主任护师、3 名主管护师;相关工作年限≥8 年)对问答结果进行独立评价。向专家提供统一的评分标准及评价表,评价前对模型匿名化处理,并对全部问答文本进行预处理,去除可能暴露模型来源的特征信息。评价指标如下:(1)整体质量。采用 Bernard 等[7] 2007 年开发的整体质量分数(global quality score,GQS)进行评价。从“质量差逻辑混乱,大部分信息缺失,对患者完全没有帮助”至“质量和逻辑性极佳,对患者非常有帮助”依次计 1~5 分。(2)准确性。采用 Likert 5 级评分,从“准确性非常差,回答不准确,可能会误导和伤害患者” 至“准确性非常好”依次计 1~5 分[8]。(3)全面性。采用 Likert 5 级评分,从“信息非常不全面”至“非常全面”依次计 1~5 分[9]。(4)简洁性。通过计算各问题答案的字符数评价问答结果的简洁程度[10]。

1.4 统计学处理 采用 R 语言(R 4.4.1)进行统计分析,连续变量以 $\bar{x}\pm s$ 表示。不同模型评价结果差异采用单因素方差分析比较,方差齐性时行 Tukey 事后比较,方差不齐时采用 Welch 校正及 Games-Howell 法。所有检验均为双侧检验,以 $P<0.05$ 或 $P<0.01$ 为差异有统计学意义。采用组内相关系数(intraclass correlation coefficient,ICC)评价评分者一致性,$ICC<0.5$ 为一致性较差、0.5~0.75 为中等、≥0.75 为良好。

2 结果

2.1 整体质量比较 5 种模型 GQS 评分差异有统计学意义($F=141.77$,$P<0.001$),“豆包”评分最高。“豆包”在预防及治疗与预后问题得分最高,“通义千问”在疾病基本信息、病因与风险因素、症状与诊断问题得分最高,“DeepSeek”在康复锻炼问题得分最高,见表 2。

2.2 全面性比较 5 个模型全面性评分差异有统计学意义($F=76.60$,$P<0.001$),“豆包”得分较高。“豆包”在预防、治疗与预后、日常生活护理类问题的全面性得分最高,“通义千问”在疾病基本信息、病因与风险因素、症状与诊断类问题的全面性得分最高,见表 2。

2.3　准确性比较　5 个模型准确性评分差异有统计学意义（$F=84.32$，$P<0.001$），“豆包”得分最高。“豆包”在病因与风险因素、预防、治疗与预后及日常生活护理类问题的准确性得分最高，“通义千问”对基本概述方面得分最高，“DeepSeek”在症状与诊断类问题中得分最高，“文心一言”在康复锻炼领域得分最高出，见表 2。

2.4　简洁性比较　5 个模型字符数差异有统计学意义（$F=2448.17$，$P<0.001$），其中“DeepSeek”字符数最少；“智谱清言”字符数高于其他模型，差异均有统计学意义（均 $P<0.05$）。“DeepSeek”在各类问题上的回答均最简洁，“智谱清言”冗长程度最高，见表 2。

表 2　5 种 LLMs 在不同类别的乳腺癌相关淋巴水肿问答任务中的表现

评估指标	模型	总体	疾病基本信息（$n=16$）	病因与风险因素（$n=15$）	症状与诊断（$n=12$）	预防（$n=20$）	治疗与预后（$n=11$）	康复锻炼（$n=8$）	日常生活护理（$n=18$）
整体质量	DeepSeek	4.53±0.63	4.51±0.66	4.55±0.55	4.55±0.65	4.52±0.54	4.47±0.66	4.65±0.48	4.49±0.75
	通义千问	4.58±0.59	4.71±0.51^{a}	4.79±0.44^{a}	4.55±0.59	4.54±0.58	4.25±0.64	4.60±0.55	4.53±0.67
	智谱清言	3.82±0.77^{ab}	3.69±0.94^{ab}	3.91±0.82^{ab}	3.62±0.72^{ab}	3.84±0.75^{ab}	3.95±0.68^{ab}	3.95±0.71^{ab}	3.82±0.66^{ab}
	豆包	4.68±0.56^{abc}	4.60±0.54^{c}	4.73±0.53^{ac}	4.50±0.62^{c}	4.75±0.56^{abce}	4.65±0.58^{bce}	4.50±0.68^{c}	4.83±0.43^{abc}
	文心一言	4.37±0.64^{abcd}	4.21±0.69^{abcd}	4.36±0.65^{bcd}	4.32±0.72^{c}	4.30±0.63^{abcd}	4.35±0.64^{cd}	4.53±0.55^{c}	4.56±0.54^{cd}
全面性	DeepSeek	4.43±0.67	4.48±0.66	4.55±0.62	4.43±0.65	4.30±0.66	4.36±0.68	4.43±0.71	4.46±0.71
	通义千问	4.59±0.59	4.68±0.55^{a}	4.76±0.46^{a}	4.58±0.59	4.51±0.61^{a}	4.42±0.60	4.48±0.68	4.6±0.60
	智谱清言	4.01±0.71^{ab}	3.98±0.78^{ab}	3.97±0.73^{ab}	3.83±0.78^{ab}	3.90±0.64^{ab}	4.05±0.62^{ab}	4.20±0.65	4.20±0.66^{ab}
	豆包	4.66±0.59^{abc}	4.55±0.63^{c}	4.76±0.46^{ac}	4.55±0.59^{c}	4.69±0.61^{abc}	4.58±0.69^{c}	4.53±0.72^{c}	4.84±0.39^{abc}
	文心一言	4.45±0.68^{bcd}	4.16±0.79^{abd}	4.36±0.69^{bcd}	4.35±0.71^{c}	4.47±0.64^{cd}	4.4±0.66^{c}	4.6±0.59^{c}	4.79±0.44^{abc}
准确性	DeepSeek	4.57±0.60	4.51±0.67	4.63±0.54	4.68±0.62	4.56±0.59	4.36±0.73	4.63±0.49	4.61±0.49
	通义千问	4.58±0.61	4.65±0.62	4.75±0.50	4.52±0.65	4.58±0.59	4.35±0.70	4.55±0.55	4.58±0.62
	智谱清言	4.04±0.67^{ab}	3.93±0.74^{ab}	4.03±0.75^{ab}	3.98±0.72^{ab}	4.03±0.74^{ab}	3.95±0.56^{ab}	4.10±0.50^{ab}	4.24±0.53^{ab}
	豆包	4.69±0.56 ^{abc}	4.56±0.73^{c}	4.77±0.45^{c}	4.57±0.59^{c}	4.75±0.56^{abc}	4.71±0.60^{abc}	4.60±0.55^{c}	4.78±0.42^{abc}
	文心一言	4.48±0.65 ^{abcd}	4.26±0.71^{abcd}	4.40±0.68^{abcd}	4.42±0.70^{ac}	4.49±0.69^{cd}	4.45±0.63^{cd}	4.65±0.48^{c}	4.69±0.51^{c}
简洁性	DeepSeek	838.84±163.84	787.38±138.25	809.67±98.72	779.67±159.44	879.20±159.98	975.09±163.25	938.13±166.75	776.11±152.94
	通义千问	1096.76±215.68^{a}	1034.13±212.25^{a}	1055.8±146.12^{a}	1077.42±299.19^{a}	1132.60±111.58^{a}	1248.82±299.64^{a}	1138.63±180.3^{a}	1048.11±186.98^{a}
	智谱清言	4619.2±1579.20^{ab}	3618.63±1418.76^{ab}	4831.8±1184.24^{ab}	3732.67±1615.82^{ab}	4449.35±1535.15^{ab}	5238.64±2069.27^{ab}	5137.25±836.29^{ab}	5502.39±1007.01^{ab}
	豆包	1115.68±281.00^{ac}	1024.63±271.63^{ac}	1141.13±275.68^{abc}	1062.0±306.74^{ac}	1204.35±213.78^{abc}	1220.45±248.99^{ac}	969.50±366.34^{bc}	1113.61±256.3^{ac}
	文心一言	1020.24±190.66^{c}	979.81±154.56^{ac}	1068.8±165.28^{ac}	1042.83±180.6^{ac}	1012.6±173.65^{abcd}	1004.55±235.57^{bcd}	1055.75±199.09^{ac}	1002.94±217.19^{acd}

a：$P<0.05$，与第 1 层比较；b：$P<0.05$，与第 2 层比较；c：$P<0.05$，与第 3 层比较；d：$P<0.05$，与第 4 层比较

2.5　评价者间一致性　5 位评价者组内相关系数 *ICC* 为 0.594[95%*CI*(0.561，0.626)，$F=2.465$，$P<0.001$]，一致性中等。在整体质量、全面性、准确性方面，*ICC* 分别为 0.665、0.509、0.593，一致性均为中等。

3　讨论

3.1　5 种 LLMs 在 BCRL 问答任务中的整体表现与差异分析　本研究表明，5 种 LLMs 在 BCRL 常见问题问答任务中整体表现较好。综合 4 个评价指标，“豆包”表现最佳，其次为“通义千问”和“DeepSeek”，“文心一言”与“智谱清言”表现相对较差。各模型在不同维度的表现有所差异。部分模型虽语言流畅、逻辑清晰，但存在事实性错误或遗漏关键信息，例如“根据患肢周径进行淋巴水肿分级”，可能导致患者延误就医或采取不当护理措施，表明“幻觉”问题仍然是医学问答的重要挑战[11]。此外，“DeepSeek”回答虽较为简洁性，但平均字符数仍超过 800，可能会增加患者阅读负担；而“智谱清言”字符数最多，却未能体现信息增量，冗长的表述反而影响了整体评价结果。模型间差异可能与其架构范式、训练策略及产品定位有关[12]。“豆包”表现最佳的可能原因包括：一是大规模中文对话数据与应用场景迭代，在理解患者口语化表述及生成结构化回复方面具有明显优势；二是官方产品能力描述显示其支持上下文理解与长期记忆等机制；三是模型在输出医学信息的同时更强调对话中的情绪理解与支持性表达[13]。研究[14]表明，“通义千问”在前列腺癌围术期健康教育问答中优于“文心一言”及“智谱清言”。“DeepSeek”和“通义千问”采用混合专家架构，有助于聚焦高信息密度内容并减少计算冗余[12]。相比之下，“文心一言”与“智谱清言”在本研究中表现相对较弱，可能与其医学知识覆盖及针对临床问答任务的优化程度不足有关[15]。各模型的设计目标与应用定位差异，决定了其在医学问答场景中的适用程度。

3.2　5 种 LLMs 在不同类别 BCRL 问答任务中的表现分析　本研究发现，5 种 LLMs 在日常生活护理、病因与风险因素、康复锻炼类问题中的表现较好，而在症状识别、诊断、治疗及预防类问题中的准确性和全面性明显下降，提示模型知识库覆盖差异及不同类别医学知识在语言建模中的可表达性和复杂度存在区别[16]。日常护理和康复锻炼类问题多为流程性、规范化内容，模型易于从语料中总结模式并生成结构化回复。而诊断、治疗与预防问题涉及复杂决策与个体化差异，依赖专业训练与临床经验，公开语料覆盖有限，模型生成难度显著增加[17]。不同模型在任务表现上各具特点，“豆包”在预防、治疗与预后、日常护理等应用性问题中表现优异；“通义千问”在疾病基本信息、病因与风险因素、症状与诊断类问题方面表现较好；“DeepSeek”则在康复锻炼类问题

具有优势。“DeepSeek”侧重推理范式与链式逻辑，适于结构化问题，但在病因与预防类问题中信息覆盖相对不足[12]。“豆包”与“通义千问”采用生成范式，前者经医疗任务微调，结构清晰、表述规范，后者强调通识知识体系，在解释型问题中表现稳定[14]。因此，LLMs 在复杂医疗决策支持中的适用性仍有限，未来需加强专业微调、专家知识注入与风险提示机制建设[16]。临床实践中可根据不同模型在特定问题类型上的表现进行有针对性的选择与辅助应用。

3.3 LLMs 在患者科普和健康教育中的应用前景及挑战 本研究显示，部分 LLMs（如“豆包”“通义千问”）在 BCRL 常见问题的回答中具备较高的整体质量与可读性，提示其在患者健康教育中的应用具备应用潜力。在医学资源或随访管理不足的情境下，将 LLMs 作为辅助性健康信息工具，有助于提升患者信息获取效率与自我管理能力[11]。需要注意的是，LLMs 尚不可替代专业医护人员[4]，其仍存在准确性不足、知识更新滞后及缺乏个体化互动等局限，临床应用中应引入人工审核与反馈机制[18]。因此，未来部署应优先考虑“专家＋AI”协同模式，在风险控制与知识可信度之间取得平衡；同时，可基于综合表现较优的模型开展针对特定疾病的医学微调，构建可验证、可解释、可控的临床语言生成系统，推动其在健康教育与慢病随访中的规范化应用[19]。

3.4 研究的局限性及展望 （1）常见问题的整合与分类由单一研究者完成，可能引入主观性偏倚，未来应进行独立重复编码与一致性检验；（2）本采用单次提问策略，未对模型输出进行重复采样，可能低估模型稳定性并且引入随机波动偏倚，后续研究将采用多次（≥3 次）重复提问并汇总评分以提高稳健性；（3）评分者间一致性处于中等水平，部分指标具有主观性，不同专家在理解与权重分配上可能存在差异，提示需通过标准化培训进一步提高评分一致性；（4）简洁性仅以字符数衡量，未反映可读性、信息密度及临床效用，未来可结合关键要点覆盖率、可读性指标及专家评分进行综合评价；（5）采用联网搜索模式，模型输出可能受外部信息更新影响，限制结果在不同时间点的可重复性，后续研究需控制。

4 小结

本研究比较了 5 种中文 LLMs 在 BCRL 护理问答场景中的表现，结果显示各模型整体表现较好，其中“豆包”综合性能最佳。本研究为模型在该类护理问答场景中的性能评估提供了基础证据，但其对患者行为改变及临床结局的实际影响仍需研究验证。

【参考文献】

[1] PAPPALARDO M, STARNONI M, FRANCESCHINI G, et al. Breast cancer-related lymphedema: recent updates on diagnosis, severity and available treatments[J/OL]. [2025-05-01]. https://www.mdpi.com/2075-4426/11/5/402. DOI:10.3390/jpm11050402.

[2] DAVIES C, LEVENHAGEN K, RYANS K, et al. Interventions for breast cancer-related lymphedema: clinical practice guideline from the Academy of Oncologic Physical Therapy of APTA[J]. Phys Ther, 2020, 100(7): 1163-1179.

[3] 潘辉.全球大语言模型研究进展：基于知识图谱的研究主题识别[J].造纸装备及材料，2025，54(2)：76-81.

[4] 刘泽垣，王鹏江，宋晓斌，等.大语言模型的幻觉问题研究综述[J].软件学报，2025，36(3)：1152-1185.

[5] 邢倩，何达.医疗大语言模型的评价现状及思考[J].健康发展与政策研究，2025，28(1)：65-72，79.

[6] HANCI V, ERGUN B, GUL S, et al. Assessment of readability, reliability, and quality of ChatGPT®, BARD®, Gemini®, Copilot®, Perplexity® responses on palliative care[J/OL]. [2025-05-01]. https://journals.lww.com/md-journal/fulltext/2024/08160/assessment_of_readability,_reliability,_and.61.aspx. DOI: 10.1097/MD.0000000000039305.

[7] BERNARD A, LANGILLE M, HUGHES S, et al. A systematic review of patient inflammatory bowel disease information resources on the World Wide Web[J]. Am J Gastroenterol, 2007, 102(9): 2070-2077.

[8] BALCI A S, ÇAKMAK S. Evaluating the accuracy and readability of ChatGPT-4o's responses to patient-based questions about keratoconus[J]. Ophthalmic Epidemiol, 2025, 32(6): 698-703.

[9] ZHANG Q, WU Z, SONG J, et al. Comprehensiveness of large language models in patient queries on gingival and endodontic health[J]. Int Dent J, 2025, 75(1): 151-157.

[10] YALAMANCHILI A, SENGUPTA B, SONG J, et al. Quality of large language model responses to radiation oncology patient care questions[J/OL]. [2025-05-01]. https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2816884. DOI:10.1001/jamanetworkopen.2024.4630.

[11] 汤志杰，孙国珍，李芸霞，等.大语言模型在护理领域应用的机遇与挑战[J].中国护理管理，2024，24(6)：929-933.

[12] 张静，江永丰，黄俊杰.国产大语言模型在健康科普文案生成中的对比研究——以数字人播报为应用场景[J].无线互联科技，2025，22(19)：80-85.

[13] 梁筱.字节跳动：AI 落地，实力派玩家入局[J].国际品牌观察，2024(18)：46-53.

[14] 谭晓文，陈文芳，王娜娜，等.国内不同大型语言模型对前列腺癌围术期护理与健康教育相关问题的查询响应与效果评价[J].中华男科学杂志，2024，30(2)：151-156.

[15] 赵朋伟，李干，薛紫阳，等.以 ChatGPT-4o 和 DeepSeek-V3 为代表的生成式人工智能在肠造口患者教育支持中的对比研究[J].军事护理，2025，42(12)：71-74.

[16] 李源，罗碧如，FU M R，等.大语言模型在临床护理实践的潜在应用及障碍分析[J].护理学报，2024，31(21)：44-48.

[17] 邢倩，何达.医疗大语言模型的评价现状及思考[J].健康发展与政策研究，2025，28(1)：65-72，79.

[18] SHOOL S, ADIMI S, SABOORI AMLESHI R, et al. A systematic review of large language model (LLM) evaluations in clinical medicine[J/OL]. [2025-05-30]. https://link.springer.com/article/10.1186/s12911-025-02954-4. DOI:10.1186/s12911-025-02954-4.

[19] 吴春志，赵玉龙，刘鑫，等.大语言模型微调方法研究综述[J].中文信息学报，2025，39(2)：1-26.

（本文编辑：郁晓路）