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

赵朋伟[1],李干[1],薛紫阳[2],王龙强[1],乔莉娜[1],李徐奇[1]
(1.西安交通大学第一附属医院 普通外科,陕西 西安 710061;
2.西北大学附属医院暨西安市第三医院 肿瘤外科,陕西 西安 710018)

【摘要】 **目的** 对比分析以 ChatGPT-4o 和 DeepSeek-V3 为代表的生成式人工智能(generative artificial intelligence,GenAI)在肠造口患者教育方面生成信息的质量,为护理领域健康教育的智能化改革提供参考。**方法** 将临床常见的 34 个肠造口护理问题分为 3 类并编写为结构化指令,依次向 2 种 GenAI 进行交互,由 6 位国内肠造口护理专家对模型输出结果进行盲法评价;再使用改良的确保患者信息质量(ensuring quality information for patients,EQIP)评估工具对 2 种 GenAI 生成的信息进行系统评价,分数≥19 分为高分模型。**结果** ChatGPT-4o 和 DeepSeek-V3 生成肠造口护理信息的准确率分别为 82.4%(28/34)、67.6%(23/34),EQIP 评分分别为 22、24 分。**结论** 2 种 GenAI 均可作为肠造口患者初步了解医疗信息的辅助工具,涉及决策问题时仍需结合专科医护团队的经验和判断。垂直领域大语言模型的开发和应用是肠造口患者教育智能化改革的重要方向。

【关键词】 生成式人工智能;ChatGPT;DeepSeek;肠造口护理;健康教育
doi:10.3969/j.issn.2097-1826.2025.12.017
【中图分类号】 R473.6 【文献标识码】 A 【文章编号】 2097-1826(2025)12-0071-04

A Comparative Study of Generative Artificial Intelligence Represented by ChatGPT-4o and DeepSeek-V3 in Educational Support for Patients with Enterostomy

ZHAO Pengwei[1],LI Gan[1],XUE Ziyang[2],WANG Longqiang[1],QIAO Lina[1],LI Xuqi[1](1.Department of General Surgery,The First Affiliated Hospital of Xi'an Jiaotong University,Xi'an 710061,Shaanxi Province,China;2.Department of Oncology Surgery,The Affiliated Hospital of Northwest University/Xi'an No.3 Hospital,Xi'an 710018,Shaanxi Province,China)
Corresponding author:LI Xuqi,Tel:029-85324343

【**Abstract**】 **Objective** To compare and analyze the quality of information generated by generative artificial intelligence (GenAI) represented by ChatGPT-4o and DeepSeek-V3 in patient education for enterostomy,providing reference for the intelligent reform of health education in the nursing field.**Methods** A total of 34 common questions of clinical enterostomy care were categorized into 3 types and formulated as structured prompts.These prompts were sequentially input into the 2 GenAI models for interaction.The outputs were then blindly evaluated by 6 domestic experts in enterostomy care.Additionally,a modified Ensuring Quality Information for Patients (EQIP) tool was used to systematically evaluate the information generated by the 2 GenAI models,with a score ≥19 indicating a high-quality model.**Results** The accuracy rates of information on enterostomy care generated by ChatGPT-4o and DeepSeek-V3 were 82.4% (28/34) and 67.6% (23/34),respectively.Their EQIP scores were 22 and 24 points,respectively.**Conclusions** Both GenAI models can serve as auxiliary tools for enterostomy patients to gain preliminary understanding of medical information.However,it is still necessary to integrate the experience and judgment of specialized healthcare teams in terms of decisions.The development and application of large language models in vertical domains represent an important direction for the intelligent reform of patient education in enterostomy care.

【**Key words**】 generative artificial intelligence;ChatGPT;DeepSeek;enterostomy care;health education
[Mil Nurs,2025,42(12):71-74]

生成式人工智能(generative artificial intelligence,GenAI)的发展对医学领域产生了深远影响,以 ChatGPT 和 DeepSeek 为代表的主流 GenAI 展示了强大的自然语言处理能力。GenAI 凭借即时响应能力,能够突破时空限制解答患者疑问[1-2]。在加速康复外科的背景下,肠造口患者在院期间学习造口护理必备知识和技能的时间有限,院外护理阶段可能面临诸多问题,且护理需求随时间动态变化[3]。传统的健康教育受限于医疗资源分配不均、专业人员短缺等现实困境,难以满足患者持续性、个性化的

【收稿日期】 2025-08-19 【修回日期】 2025-11-01
【基金项目】 西安交通大学基本科研业务团队和人才支持-学科交叉团队项目(xtr062025009)
【作者简介】 赵朋伟,本科,主管护师,电话:029-85324343
【通信作者】 李徐奇,电话:029-85324343

知识需求[4]。GenAI 为解决这一矛盾提供了创新路径。在肠造口护理健康教育场景中,专业知识的精确性、操作指导的规范性对 GenAI 生成信息的质量提出了更高的要求。然而,尚未有研究系统地对比 ChatGPT 和 DeepSeek 等 GenAI 在护理健康教育中的使用情况。鉴于此,本研究以 ChatGPT-4o 和 DeepSeek-V3 为代表的 GenAI 为例,聚焦肠造口护理健康教育,系统性分析 2 种模型在生成信息质量方面的性能差异,揭示现有 GenAI 在护理领域的性能边界,为筛选可靠的教育支持工具提供科学依据。

1　资料与方法

1.1　*研究资料*　基于前期研究[5],针对在我院普通外科接受肠造口手术,且出院后在造口门诊就诊的 14 例患者进行深入访谈,贯穿患者在院内、院外护理的全过程,充分了解并总结了肠造口患者在护理过程中的真实感受和所遇到的问题。回顾既往研究[5]中 14 例患者的访谈资料和研究结果,再结合造口患者在微信群中提出的常见护理问题;依据中华护理学会成人肠造口护理团体标准(标准号:T/CNAS07-2019)主题框架进行分类,经双人背对背提炼、汇总出 3 类、34 个肠造口护理问题,分别为:基础定义类(16 问)、生活指导类(3 问)、措施指导类(15 问)。护理问题的分类依据为:(1)基础定义类,即需要解释何为肠造口、造口定位等类型的问题。例如:什么是肠造口?(2)生活指导类,即需要指导肠造口如何洗澡、饮食等类型的问题。例如:造口术后可以洗澡吗?(3)措施指导类,即需要指导如何进行某项护理操作,或出现并发症时的处理等类型的问题。例如,如何进行造口灌洗?最后,经小组会议讨论并确定。

1.2　*方法*

1.2.1　*评价 2 种 GenAI 生成肠造口护理信息的准确性*

1.2.1.1　*编写结构化指令*　将 34 个肠造口护理问题编写为结构化指令。经预实验验证、动态优化以及小组会议讨论后确定,主要包括角色、任务、目标人群、字数等。如:“您是 1 名在肠造口领域深耕多年和经验丰富的专家,请针对(患者的造口护理问题),做出护理回答,字数在 200 字以内”。

1.2.1.2　*输入指令并自主生成信息*　在开始输入指令之前,将所有浏览数据完全删除,创建单独的账户与 2 种 GenAI 进行交互。将肠造口护理问题结构化指令依次输入到 ChatGPT-4o (https://openai.com/index/gpt-4/)、DeepSeek-V3 (https://www.deepseek.com/),由 2 种模型进行应答,并以 word 形式保存。

1.2.1.3　*专家函询评价生成肠造口护理信息的准确性*　以中华护理学会成人肠造口护理团体标准为统一评价标准,由 6 位国内肠造口护理领域的专家进行独立、匿名评价。结果由双人核对并录入,如果 GenAI 生成的信息准确,则标记答案为“是”;如果不准确,则标记为“否”。当评价结果存在冲突时,召开专家小组讨论会确定最终结果。

1.2.2　*系统性评价 2 种 GenAI 生成信息的质量*

1.2.2.1　*评价工具*　确保患者信息质量(ensuring quality information for patients,EQIP)评估工具是由 Moult 等[6]于 2004 年在英国编制的一种用于评估医疗信息质量的标准化评估工具。Charvet-Berard 等[7]在 2008 年对 EQIP 工具进行修订和扩增,包含内容数据(18 个条目)、识别数据(6 个条目)和结构数据(12 个条目)3 个维度。改良的 EQIP 工具已被证明能够在多个医疗领域生成信息的精准性、可读性和逻辑性进行可靠评估[8-9]。该工具不仅可以分析信息呈现的形式,还能够有效识别信息资源的优缺点。本研究使用改良的 EQIP 工具对 2 种 GenAI 生成信息的质量进行全面分析。

1.2.2.2　*编写指令并完成问答*　参照 EQIP 工具的使用说明[6],将改良的 EQIP 工具内容数据中的每个条目重写为指令,如内容数据中“医疗问题、治疗或程序的描述”,改写为“请描述肠造口患者常见的问题”。所有指令由 2 名研究者进行核对,经小组会议讨论并确定。在开始输入指令之前,将所有浏览数据完全删除并创建单独的账户,将指令分别输入 2 种 GenAI 进行应答,结果以 word 形式进行保存。

1.2.2.3　*信息质量评价*　由 2 名研究者独立评价生成信息质量。如答案正确完整,则计为“1”;若不正确、不完整或矛盾,则计为“0”,如不涉及则标记为“NA”。错误或不切实际的答案记录为“人工智能幻觉”。如果存在矛盾,经过协商形成一致意见。EQIP 得分≥19 分为高分模型[10]。

1.2.3　*质量控制*　本研究采取了以下质量控制措施以保证研究的严谨性:(1)所有研究者均经过系统的科研培训,具有较高的理论和科研实践基础;(2)所有信息资料均由 2 名研究者背对背完成收集、汇总和核对,以减少主观偏倚。

1.2.4　*统计学处理*　采用 SPSS 25.0 统计学软件进行数据分析。生成信息的准确率采用例数、百分比进行描述性描述。

2　结果

2.1　*2 种 GenAI 生成肠造口护理信息的准确率比较*　6 名国内肠造口护理专家应答率为 100.0%。由专家独立评价 ChatGPT-4o、DeepSeek-V3 生成信息的准确率分别为 82.4%(28/34)、67.6%(23/34),两者对 3 类肠造口护理问题得分,见表 1。

表 1　2 种 GenAI 生成肠造口护理信息的准确率[N=34,n(%)]

项　目	基础定义类（n=16）	生活指导类（n=3）	措施指导类（n=15）	准确率
ChatGPT-4o	14(87.5)	3(100.0)	11(73.3)	28(82.4)
DeepSeek-V3	12(75.0)	3(100.0)	8(53.3)	23(67.6)

2.2　2 种 GenAI 生成信息的质量评估　改良的 EQIP 工具最高分数为 36 分，内容数据、识别数据、结构数据 3 个维度的最高分数分别为 18、6 和 12 分。使用 ChatGPT-4o、DeepSeek-V3 的生成信息的 EQIP 评分分别为 22 分和 24 分，见表 2。在内容数据维度，使用 2 种 GenAI 生成信息的质量评分均为 15 分。因未提供任何有关来源的信息，包括参与信息生成的发布机构、人员。在识别数据维度，ChatGPT-4o 评分为 1 分。尽管 ChatGPT-4o 能够提供包含分析条件信息的文档链接，但其提供的信息无法追溯到原始数据。而 DeepSeek-V3 提供了可获取肠造口护理的具体书目、在线资源信息。在结构数据维度，2 种 GenAI 生成的信息结构包括简短的介绍、项目列表和结论，评分均为 6 分。由于句子长度经常超过 15 个字，该项没有得分。此外，在“解决医疗干预费用和保险问题”项目中提供的数据及信息模糊，因此“信息明确”和“清晰且相关的数字或图表”项目没有得分。

3　讨论

3.1　GenAI 生成肠造口护理信息质量的对比分析　本研究结果显示 2 种 GenAI 生成信息的 EQIP 评分均高于 19 分，说明两者生成信息的质量较高。由国内肠造口护理专家评价 2 种模型生成信息的准确性结果发现 ChatGPT-4o 生成信息的准确率更高。2 种模型对于基础定义类和生活指导类问题的回答整体可被接受，但在措施指导类问题准确率略低，ChatGPT-4o 生成信息的准确率高于 DeepSeek-V3（73.3% *vs*. 53.3%），表明 GenAI 在自动化总结医疗证据时，生成的结果可能缺乏稳定性和一致性，其与李鹏等[11]研究结果基本一致。总体而言，2 种 GenAI 均表现出较强的理论知识掌握能力，可以作为初步了解肠造口护理信息的检索手段。但涉及整合动态知识并生成可操作的序列化指令时，存在决策机械化倾向，需要与临床专家协作加以解决。

3.2　GenAI 在结构化指令编写中的设计　本研究编写了常见肠造口护理问题的结构化交互指令，限定了使用者身份，并明确了生成信息的范围和字数。通过结构化提示设计及标准化输出框架，以优化 GenAI 生成内容的相关性、可读性和实用性，从而提升患者教育效果并降低因信息模糊或误导性陈述导致的健康风险。GenAI 在处理复杂的隐含关系时，能够通过自然语言处理技术和机器学习算法，模拟人类的思维过程，整合上下文信息，更好地理解问题的背景，有效解决一对多映射问题，从而生成更具参考性的健康信息[12]。因此，提出问题的方式和背景信息的提供对于患者获取高质量的信息至关重要。

表 2　2 种 GenAI 生成肠造口护理信息的 EQIP 评估结果（分）

项　目	ChatGPT-4o	DeepSeek-V3
内容数据		
涵盖哪些主题的初步定义	0	0
先前定义的主题覆盖范围[a]	NA	NA
医疗问题、治疗或程序的描述	1	1
干预目的的定义	1	1
治疗方案描述（保守治疗）	1	1
干预顺序和外科手术的描述	1	1
描述对患者的定性益处	1	1
描述对患者的定量益处	1	1
定性风险和并发症的描述	1	1
定量风险和并发症的描述	1	1
解决生活质量的问题	1	1
描述如何解决并发症	1	1
患者可能采取的预防措施描述	1	1
提及患者可能检测到的警觉信号	1	1
解决医疗干预费用和保险问题	1	1
医院服务的具体联系方式[b]	NA	NA
其他可靠信息/支持来源的具体细节	1	1
主要所有相关问题的覆盖范围（所有内容标准的摘要项）	1	1
维度得分	15	15
识别数据		
发布或修订数据	1	1
发行机构的标志	NA	NA
生成信息的人员或实体的名称	0	0
为信息提供经济支持的个体或实体的名称	0	0
信息中使用循证数据的简短参考书目	0	1
关于患者是否以及如何参与/咨询的声明	0	1
维度得分	1	3
结构数据		
使用日常语言和解释复杂单词或专业用语	1	1
所有药物或商品均使用通用名称[c]	NA	NA
使用短句（平均少于 15 个字）	0	0
给读者个人地址	0	0
尊重的语气	1	1
信息清晰（无歧义或矛盾）	0	0
关于风险和益处的平衡信息	1	1
按逻辑顺序呈现信息	1	1
令人满意的设计和布局（不包括图形或图表）	1	1
清晰且相关的数字或图表[d]	0	0
为读者的注释或问题提供命名空间	1	1
包含与建议相反的知情同意书[e]	NA	NA
维度得分	6	6
总分	22	24

a：如第 1 项答案为“否”，则为 NA；b：如不是医院，则为 NA；c：如未描述药物，则为 NA；d：如没有，则为 NA；e：如不是来自医院，则为 NA

3.3　GenAI 在肠造口患者教育中的应用潜力　本研究中，2 种模型均能将专业的医学术语转化为易于患者理解的语言。对于接受肠造口手术的患者而言，GenAI 可以作为辅助工具，为患者提供造口护理知识等信息支持，帮助其更有效地进行自我护理管理。GenAI 可以通过促进信息交换和弥合临床和临床前环境之间的差距，使患者更容易获取关于其病情和治疗方案的更多信息，有助于患者在与医生的

沟通中更有针对性，从而在住院或门诊期间获得更及时的医疗支持[13]。虽然 GenAI 可以整合对话信息给出特定的回答，但是在提供个性化建议方面往往力有未逮，即使是准确的结果也可能是不完整的，无法完全满足患者的具体需求[14]。在复杂的决策过程中，临床专家的经验和判断仍是确保患者得到高质量优质护理服务的关键。

3.4　GenAI 在肠造口患者教育中的挑战与突破

本研究中，ChatGPT-4o 提供的包含分析条件信息的文档链接无法追溯到原始数据；而 DeepSeek-V3 提供了可获取肠造口护理相关的具体书目、在线资源信息，这与 Kacer 等[15]的研究一致，提示 ChatGPT 在信息来源的透明度方面不足。这不仅影响了模型的可解释性，也对其信任度和接受度产生负面影响。此外，本研究中，2 种 GenAI 在回答肠造口患者常见护理问题时均出现了“人工智能幻觉”。例如，DeepSeek 错误的认为造口周围皮炎需要引流处理。GenAI 在回答特定专业领域的问题时，往往因缺乏该领域内的专业知识而受到限制，“人工智能幻觉”就会发生[16]。一旦 GenAI 出现幻觉问题，将严重影响模型的可靠性和实用性，这一现象在需要高度专业知识和精准信息的医学领域表现尤为突出[17]。在应对“人工智能幻觉”问题上，垂直领域大语言模型通过领域适应和定制化开发[18]，融合结构化知识图谱并结合微调策略[19]，能够克服现有通用大语言模型存在的缺陷与问题。

4　小结

GenAI 在改善患者教育和自我管理方面具有一定的潜力，可以作为肠造口患者初步获取医疗信息的辅助工具。开发和应用垂直领域大语言模型是肠造口护理健康教育智能化改革的重要方向。通过结合领域特定知识和大语言模型技术，可以为患者提供个性化、易于理解和基于证据的教育信息，从而改善患者的健康管理和生活质量。

致谢：感谢 6 位国内肠造口护理领域的专家为 2 种 GenAI 生成肠造口护理信息的准确性做出盲法评价：徐洪莲（海军军医大学第一附属医院）、田丽（首都医科大学附属北京友谊医院）、朱卉（天津市人民医院）、吕琳（甘肃省人民医院）、孙佳男（吉林大学第一医院）、吕颀（北京大学第三医院）。感谢郑雪梅（西安交通大学第一附属医院）在论文修改过程中给予的指导。

【参考文献】

[1] STAFIE C S，SUFARU I G，GHICIUC C M，et al.Exploring the intersection of artificial intelligence and clinical healthcare：a multidisciplinary review[J/OL].[2025-08-01].https://pmc.ncbi.nlm.nih.gov/articles/PMC10297646/.DOI：10.3390/diagnostics13121995.

[2] ZHENG Y，SUN X，KANG K，et al.Breast cancer in the era of generative artificial intelligence：assistant tools for clinical doctors based on ChatGPT[J].Int J Surg，2024，110(8)：5304-5305.

[3] 胡娇娇，王建宁，詹梦梅，等.肠造口患者居家护理需求质性研究的 Meta 整合[J].中华护理杂志，2023，58(3)：357-365.

[4] 宁美，陈晨，张洲一，等.基于 Kano 模型的肠造口患者健康教育需求的调查研究[J].中国临床护理，2024，16(12)：744-748，753.

[5] 温绣蔺，李小妹，辛霞，等.首次肠造口患者术后早期自我照护经历的质性研究[J].现代临床护理，2019，18(12)：42-47.

[6] MOULT B，FRANCK L S，BRADY H.Ensuring quality information for patients：development and preliminary validation of a new instrument to improve the quality of written health care information[J].Health Expect，2004，7(2)：165-175.

[7] CHARVET-BERARD A I，CHOPARD P，PERNEGER T V.Measuring quality of patient information documents with an expanded EQIP scale[J].Patient Educ Couns，2008，70(3)：407-411.

[8] CHIEN S，CUNNINGHAM D，KHAN K S.Inguinal hernia repair：a systematic analysis of online patient information using the modified ensuring quality information for patients tool[J].Ann R Coll Surg Engl，2022，104(4)：242-248.

[9] MARCASCIANO M，VITTORI E，CIRIACO A G，et al.A systematic quality assessment of online resources on eyelid ptosis using the modified ensuring quality information for patients (mEQIP) tool[J].Aesthetic Plast Surg，2024，48(9)：1688-1697.

[10] RAPTIS D A，SINANYAN M，GHANI S，et al.Quality assessment of patient information on the management of gallstone disease in the internet：a systematic analysis using the modified ensuring quality information for patients tool[J].HPB (Oxford)，2019，21(12)：1632-1640.

[11] 李鹏，张源慧，唐龙，等.ChatGPT 在护理问题和护理措施自主回答中的应用性研究[J].中华护理教育，2025，22(4)：398-402.

[12] SHEN K，WU L，TANG S，et al.Ask questions with double hints：visual question generation with answer-awareness and region-reference[J].IEEE Trans Pattern Anal Mach Intell，2024，46(12)：9648-9660.

[13] GRAVINA A G，PELLEGRINO R，CIPULLO M，et al.May ChatGPT be a tool producing medical information for common inflammatory bowel disease patients' questions? An evidence-controlled analysis[J].World J Gastroenterol，2024，30(1)：17-33.

[14] DE VITO A，GEREMIA N，MARINO A，et al.Assessing ChatGPT's theoretical knowledge and prescriptive accuracy in bacterial infections：a comparative study with infectious diseases residents and specialists[J].Infection，2025，53(3)：873-881.

[15] KACER E O，IPEKTEN F.Can ChatGPT provide quality information about fever in children? [J].J Paediatr Child Health，2025，61(1)：60-65.

[16] FELDMAN K，NEHME F.Beyond clinical accuracy：considerations for the use of generative artificial intelligence models in gastrointestinal care[J].Gastroenterology，2023，165(2)：336-338.

[17] WANG Z，WANG J，LU Z，et al.Large language modeling of hallucinatory problem mitigation based on the wheel of emotions[J/OL].[2025-08-01].https://www.sciencedirect.com/science/article/pii/S0893608025008779?via%3Dihub.DOI：10.1016/j.neunet.2025.107996.

[18] WANG X，HUANG L，XU S，et al.How does a generative large language model perform on domain-specific information extraction? A comparison between GPT-4 and a rule-based method on band gap extraction[J].J Chem Inf Model，2024，64(20)：7895-7904.

[19] 陈静，曹智勋.对抗幻觉：垂直领域中大语言模型的应用策略探讨——以中医知识问答领域为例[J].数据分析与知识发现，2025，9(4)：1-13.

（本文编辑：王园园）