[DOI]10.12016/j.issn.2096-1456.202440370

·防治实践·

大语言模型在儿童口腔预防医学领域问答的准确性比较

管伯颜[1]，　许明鹤[1]，　张惠淇[1]，　马舒蕾[1]，　张珊珊[2]，　赵俊峰[3,4]
1. 北京大学口腔医学院，北京（100081）；　2. 北京大学口腔医院口腔预防保健科，北京（100081）；　3. 北京大学计算机学院，北京（100871）；　4. 高可信软件技术教育部重点实验室，北京（100871）

【摘要】　目的　探讨国内大语言模型代表ChatGLM-6B与国外大语言模型代表ChatGPT3.5在儿童口腔预防医学领域问题回答的准确性差异，为国内大语言模型在口腔医学领域的研发提供思路。**方法**　由儿童口腔预防专家从基础（n = 35）、进阶（n = 35）、深入（n = 30）三个层次，提供了不同难度的共计100个常见儿童口腔预防医学领域问题，由2名医生分别输入到ChatGPT3.5和ChatGLM-6B中，并收集问题答案。由16名口腔医生按照预定义的3点Likert量表对ChatGLM-6B和ChatGPT3.5生成的答案进行评分，计算评分的平均分作为答案得分，答案得分高于2.8接受其为正确答案；答案得分低于1.4接受其为不正确答案；答案得分介于1.4～2.8，接受其为部分正确答案。比较2组生成答案的正确率及评分结果；对口腔医生评分进行一致性分析。**结果**　ChatGPT3.5与ChatGLM-6B对100个儿童口腔预防医学领域问题的回答正确率相似：ChatGPT3.5回答正确率为68%，部分正确率为30%，不正确率为2%；ChatGLM-6B回答正确率为67%，部分正确率为31%，不正确率为2%，无统计学差异（P > 0.05）；ChatGPT3.5与ChatGLM-6B回答不同难度（基础、进阶、深入）问题的准确性均无统计学差异（P > 0.05）。ChatGPT3.5与ChatGLM-6B回答所有问题的整体平均得分均为2.65，无统计学差异（P > 0.05）；ChatGPT3.5与ChatGLM-6B不同难度问题的得分：基础问题ChatGPT3.5平均得分2.66，ChatGLM-6B平均得分2.70；进阶问题ChatGPT3.5平均得分2.63，ChatGLM-6B平均得分2.64；深入问题ChatGPT3.5平均得分2.68，ChatGLM-6B平均得分2.61，均无统计学差异（P > 0.05）。口腔医生评分具有一致性，评价范围为一般至中等。**结论**　ChatGLM-6B与ChatGPT3.5在回答儿童口腔预防医学领域问题方面均具有潜力。ChatGLM-6B在回答儿童口腔预防医学领域问题方面取得了与ChatGPT3.5相似的表现，但二者正确率均未达到预期，不能应用于临床。未来需要进一步提升大语言模型提供医疗信息的准确性和一致性，并研发适用于口腔医学领域的医疗大模型。

【关键词】　大语言模型；　儿童口腔医学；　口腔预防医学；　口腔医学；　ChatGPT；　人工智能；　聊天机器人；　医学
【中图分类号】　R78　**【文献标志码】**　A　**【文章编号】**　2096-1456（2025）04-0313-07
【引用著录格式】　管伯颜，许明鹤，张惠淇，等. 大语言模型在儿童口腔预防医学领域问答的准确性比较[J]. 口腔疾病防治，2025，33(4)：313-319. doi:10.12016/j.issn.2096-1456.202440370.

微信公众号

Accuracy of large language models for answering pediatric preventive dentistry questions

GUAN Boyan[1], XU Minghe[1], ZHANG Huiqi[1], MA Shulei[1], ZHANG Shanshan[2], ZHAO Junfeng[3,4].　1. Peking University School of Stomatology, Beijing 100081, China;　2. Peking University Hospital of Stomatology, Beijing 100081, China;　3. School of Computer Science, Peking University, Beijing 100871, China;　4. Key Laboratory of High Confidence Software Technologies, Beijing 100871, China
Corresponding author: ZHANG Shanshan, Email: zhangshanshan@hsc.pku.edu.cn, Tel: 86-10-82195593; ZHAO Jun-

【收稿日期】2024-09-22；　**【修回日期】**2025-01-02
【基金项目】首都卫生发展科研专项项目（2024-2G-4106）；2023年度人卫创新发展研究项目（RWCY23DⅡ007007）；北京大学口腔医学院教育教学研究项目（2024-ZC-07）
【作者简介】管伯颜，在读本科生，Email：13810364792@163.com
【通信作者】张珊珊，副教授，博士，Email：zhangshanshan@hsc.pku.edu.cn，Tel：86-10-82195593；赵俊峰，研究员，博士，Email：zhaojf@pku.edu.cn，Tel：86-13601389906

feng, Email: zhaojf@pku.edu.cn, Tel: 86-13601389906

【**Abstract**】 **Objective** To evaluate and compare the accuracy of responses to pediatric preventive dentistry-related questions between the domestic large language model, ChatGLM-6B, and the international large language model, ChatGPT-3.5, in order to provide insights for further research and development of domestic language models in the field of oral medicine. **Methods** A total of 100 common pediatric preventive dentistry questions of varying difficulty levels [basic (n = 35), intermediate (n = 35), and advanced (n = 30)] were provided by pediatric preventive dentistry experts. Two doctors independently registered these questions with ChatGPT-3.5 and ChatGLM-6B and collected the answers. A cohort of 16 dentists assessed responses generated by ChatGLM-6B and ChatGPT-3.5 using a predefined 3-point Likert scale. The average score of the ratings from 16 doctors was taken as the answer score. If the answer score was higher than 2.8, it was accepted as a accurate answer; if the score was lower than 1.4, it was accepted as an inaccurate answer; if the score was between 1.4 and 2.8, it was accepted as a partially accurate answer. Comparative analysis was conducted on the accuracy rates and evaluation outcomes between the two groups. Consistency analysis of the ratings was conducted. **Results** The answer accuracy rates of ChatGPT-3.5 and ChatGLM-6B for 100 pediatric preventive dentistry questions were comparable: ChatGPT-3.5 demonstrated 68% accurate, 30% partially accurate, and 2% inaccurate responses, while ChatGLM-6B showed 67% accurate, 31% partially accurate, and 2% inaccurate responses, with no statistically significant differences ($P > 0.05$). Both models exhibited equivalent accuracy across questions of varying difficulty levels (basic, intermediate, advanced), showing no statistical differences ($P > 0.05$). The overall average scores for ChatGPT3.5 and ChatGLM-6B in answering all questions were both 2.65, with no statistically significant difference ($P > 0.05$). For questions of different difficulty levels, ChatGPT3.5 had an average score of 2.66 for basic questions while ChatGLM-6B had an average score of 2.70. For intermediate questions, ChatGPT3.5 had an average score of 2.63 and ChatGLM-6B had an average score of 2.64. For advanced questions, ChatGPT3.5 had an average score of 2.68, and ChatGLM-6B had an average score of 2.61. No statistically significant differences were observed across any difficulty category ($P > 0.05$). The consistency of the experts' grading ranged from fair to moderate. **Conclusion** This study demonstrates the potential of both ChatGLM-6B and ChatGPT-3.5 in answering pediatric preventive dentistry questions. ChatGLM-6B performed similarly to ChatGPT-3.5 in this field, but the accuracy rates of both models fell short of expectations and are not suitable for clinical use. Future efforts should focus on improving the accuracy and consistency of large language models in providing medical information, as well as developing specialized medical models for the field of oral medicine.

【**Key words**】 large language model; pediatric stomatology; preventive dentistry; stomatology; ChatGPT; artificial intelligence; Chatbot; medicine

J Prev Treat Stomatol Dis, 2025, 33(4): 313-319.

【**Competing interests**】 The authors declare no competing interests.

This study was supported by the grants from Capital's Funds for Health Improvement and Research (No. 2024-2G-4106); The Project Sponsored by the Innovative Development Research of People's Medical Publishing House (No. RWCY23DⅡ007007); Education Research Project of Peking University School and Hospital of Stomatology in 2024 (No. 2024-ZC-07).

人工智能(artificial intelligence, AI)是研究开发用于模拟、延伸和扩展人类智能的理论、方法、技术及应用的一门新兴的计算机科学技术。近年来,以大语言模型(large language model, LLM)为代表的人工智能技术在医学领域有多种交叉应用,如辅助教育、支持临床决策等[1],成为医信交叉研究的热点方向。语言数据与许多其他数据类型不同,语言数据储存的信息不只是客观的数字,还包含思维、逻辑、知识、行为模式等丰富的人文要素[2]。因此,基于大语言模型的交互式人工智能超越了传统聊天机器人的局限性,产生了越来越人性化的对话功能。该技术展示了非凡的能力,例如理解上下文、生成连贯的文本以及适应各种自然语言处理(natural language processing, NLP)任务,包括但不限于语言翻译、回答问题和文本生成[3]。ChatGPT3.5是由美国公司OpenAI基于预训练生成式通用大语言模型创建的聊天机器人,其在没有任何专门培训的情况下可以达到或接近美国医生执照考试(United States medical licensing exam, USMLE)的及格门槛,表明了其在医学教育和临床决策支持方面的巨大潜力[4-5]。近年来,我国的大语言模型也在不断推进。2023年7月,智谱

AI发布了大语言模型ChatGLM-6B，但其在医学领域的应用仍还有限。随着技术的进步，人们越来越倾向于使用自然语言对话系统获取医疗知识，以便迅速、高效地获取基本医疗信息。目前，ChatGPT在整形外科、泌尿科、肝脏病理学等领域都有研究，其报告的准确性和全面性也各有不同[6-8]。在口腔医学领域，ChatGPT也能为患者提供即时反馈，满足患者特定的口腔健康信息需求[9]，并为患者提供情感支持[10]。虽然聊天机器人为复杂的医疗问题提供了对话式的、看似很权威的回答，但这些结果的准确性和一致性有待进一步确认和提升[11-14]。本研究旨在评估和比较ChatGLM-6B与ChatGPT3.5在儿童口腔预防医学领域问题回答的准确性，以期能对国内医疗大语言模型，特别是口腔医疗大语言模型的研究提供帮助。

1　资料和方法

1.1　问题设计

由儿童口腔预防专家提取在临床诊疗和科普宣教中患者关心的100个口腔医学问题，根据其回答难易程度分为基础（n = 35）、进阶（n = 35）、深入（n = 30）三个层次。其中，医学概念性问题与常识性问题被划分为基础层次，需要结合部分专业知识及临床资料进行解释和判断的问题被划分为进阶层次，综合专业知识与临床资料进行系统性陈述的问题被划分为深入层次。

1.2　大语言模型生成答案

由2名医生将100个问题分别输入到ChatGPT3.5和ChatGLM-6B中，收集大语言模型初次问答生成答案。

1.3　答案评估

由16名口腔医生使用预定义的3点Likert量表（1分为不正确，表示医学相关内容全部不正确或答非所问；2分为部分正确，表示医学相关内容部分不正确或待改进；3分为正确，表示医学相关内容全部正确，符合医学常识）分别对ChatGLM-6B和ChatGPT3.5生成的100个答案进行准确性评分。取16名医生评分的平均分作为答案得分，答案得分高于2.8接受其为正确答案；答案得分低于1.4接受其为不正确答案；答案得分介于1.4～2.8，接受其为部分正确答案。计算不同难度问题不正确、部分正确和正确的比例。计算不同难度问题答案得分的平均值，进行组间比较（表1、图1）。

表1　儿童口腔预防医学领域不同难度问题大语言模型生成答案及口腔医生评分（以3位评分者为例）示例

Table 1　Examples of pediatric preventive dentistry questions with various levels of difficulty and answers generated by LLM and experts' grading (3 experts as an example)

Category	Example question	Answer generated by LLM	Expert 1 score	Expert 2 score	Expert 3 score	Average score	Evaluation
Basic	Can children undergo crown placement?	Undertaking crown placement might potentially damage the teeth, which could affect their proper development. Therefore, children do not usually require dental crowns	1	1	1	1	Inaccurate
Intermediate	Why do children experience tooth replacement?	Children undergo tooth replacement because their oral cavity and craniofacial bones continue to change during growth and development. The primary teeth are eventually replaced by permanent teeth that grow in. Typically, children start losing baby teeth around the age of 6, beginning with the front incisors and followed by the molars on the sides. The process usually lasts until around 12 to 13 years old. It is a normal physiological phenomenon and is beneficial for the oral and dental health of children	2	2	3	2.3	Partially accurate
Advanced	What should be done if a child has clicking and pain in the joint in front of the ear when opening the mouth?	When a child has clicking and pain in the temporomandibular joint in front of the ear, it may be a sign of temporomandibular joint disorder syndrome (TMJDS), which requires prompt medical attention. Treatment options include oral therapy, medication, dental orthodontics, and more, with the specific method chosen based on the patient's condition and individual differences. Additionally, it is important to maintain good oral hygiene habits, avoid excessive mouth opening, clenching of the teeth, and maintain normal chewing functions. If symptoms are severe and affect the quality of life, it is recommended to seek treatment from a dental specialist in a timely manner for personalized treatment plans	3	3	3	3	Accurate

Inaccurate: score < 1.4; Partially accurate: 1.4-2.8; Accurate: > 2.8; LLM: large language model

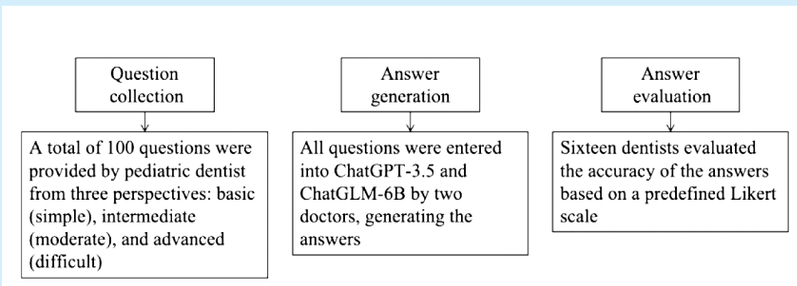

Figure 1　Procedure of experiment for evaluating the accuracy of large language models in answering pediatric preventive dentistry questions

图1　大语言模型回答儿童口腔预防医药问题准确性的试验步骤

1.4　数据分析

描述性列出对ChatGPT3.5和ChatGLM-6B生成答案的评分结果，并使用卡方检验和U检验（SPSS Statistics 26，GraphPad Prism版本9）进行组间比较。使用Kendall协同系数对16名口腔医生的答案评分进行一致性检验（SPSS Statistics 26）。

2　结　果

2.1　ChatGPT3.5与ChatGLM-6B对100个儿童口腔预防医学领域问题的整体回答情况

ChatGPT3.5与ChatGLM-6B对100个儿童口腔预防医学领域问题的回答正确率相似：ChatGPT3.5回答正确率为68%，部分正确率为30%，不正确率为2%；ChatGLM-6B回答正确率为67%，部分正确率为31%，不正确率为2%，无统计学差异（$P>0.05$）（表2）。

表2　ChatGPT3.5与ChatGLM-6B对100个儿童口腔预防医学领域问题的回答准确性

Table 2　The accuracy of ChatGPT 3.5 and ChatGLM-6B for 100 pediatric preventive dentistry questions　n (%)

LLM	Inaccurate	Partially accurate	Accurate	χ^2	P
ChatGPT3.5	2 (2)	30 (30)	68 (68)	0.024	0.988
ChatGLM-6B	2 (2)	31 (31)	67 (67)		

LLM: large language model. Inaccurate: score < 1.4; Partially accurate: score 1.4-2.8; Accurate: score > 2.8

2.2　ChatGPT3.5与ChatGLM-6B对不同难度儿童口腔预防医学领域问题回答情况

ChatGPT3.5与ChatGLM-6B对不同难度问题答案的准确性评价见表3。ChatGPT3.5与ChatGLM-6B回答不同难度（基础、进阶、深入）问题的准确性均无统计学差异（$P>0.05$）。

表3　ChatGPT3.5与ChatGLM-6B对不同难度儿童口腔预防医学领域问题的回答准确性

Table 3　The accuracy of ChatGPT 3.5 and ChatGLM-6B for pediatric preventive dentistry questions with various difficulty levels　n (%)

Category	ChatGPT3.5 (n = 100)			ChatGLM-6B (n = 100)			χ^2	P
	Inaccurate	Partially accurate	Accurate	Inaccurate	Partially accurate	Accurate		
Basic	0 (0)	10 (29)	25 (71)	0 (0)	10 (29)	25 (71)	0.000	>0.999
Intermediate	2 (6)	10 (28)	23 (66)	1 (3)	11 (31)	23 (66)	0.660	0.719
Advanced	0 (0)	10 (33)	20 (67)	1 (3)	10 (33)	19 (64)	0.767	0.682

Inaccurate: score < 1.4; Partial accurate: score 1.4-2.8; Accurate: score > 2.8

2.3　ChatGPT3.5与ChatGLM-6B回答不同难度儿童口腔预防医学领域问题得分情况

ChatGPT3.5与ChatGLM-6B对不同难度问题的平均得分见表4。基础问题ChatGPT3.5平均得分2.66，ChatGLM-6B平均得分2.70；进阶问题ChatGPT3.5平均得分2.63，ChatGLM-6B平均得分2.64；深入问题ChatGPT3.5平均得分2.68，ChatGLM-6B平均得分2.61，均无统计学差异（$P>0.05$）。

2.4　16名专家评价一致性情况

16名专家分别对ChatGPT3.5和ChatGLM-6B生成的100个答案评价结果具有一致性（表5）。

表4　ChatGPT3.5与ChatGLM-6B回答不同难度儿童口腔预防医学领域问题得分情况

Table 4　Scores of ChatGPT3.5 and ChatGLM-6B in answering pediatric preventive dentistry questions with varous difficulty levels

Category	ChatGPT3.5 (n=100)		ChatGLM-6B (n = 100)		U	P
	Mean ± SD	95%CI	Mean ± SD	95%CI		
Basic	2.66 ± 0.24	2.59, 2.75	2.70 ± 0.27	2.61, 2.79	517.0	0.346
Intermediate	2.63 ± 0.33	2.51, 2.74	2.64 ± 0.29	2.54, 2.73	606.5	0.944
Advanced	2.68 ± 0.17	2.61, 2.74	2.61 ± 0.26	2.52, 2.70	401.5	0.470
Total	2.65 ± 0.26	2.60, 2.71	2.65 ± 0.27	2.60, 2.70	4 901.5	0.809

SD: standard deviation; CI: confidence interval

表5　ChatGPT3.5与ChatGLM-6B对100个儿童口腔预防医学领域问题生成答案的专家评价一致性

Table 5　Consistency in the experts' assessment of 100 pediatric preventive dentistry answers generated by ChatGPT3.5 and ChatGLM-6B

	ChatGPT3.5			ChatGLM-6B		
Category	Kendall's W	*P*	Benchmark scale	Kendall's W	*P*	Benchmark scale
Basic	0.370	< 0.001	Fair	0.356	< 0.001	Fair
Intermediate	0.334	< 0.001	Fair	0.452	< 0.001	Moderate
Advanced	0.233	< 0.001	Fair	0.336	< 0.001	Fair

Kendall's W: Kendall's coefficient of concordance. Benchmark scale: poor < 0.001, slight 0.001-0.200, fair 0.200-0.400, moderate 0.400-0.600, substantial 0.600-0.800, almost perfect 0.800-1.000

3　讨　论

近年来，人工智能的飞速发展已经对包括医疗保健在内的许多行业产生了深远的影响[15-16]。人工智能在医疗保健中的一个重要应用是开发对话式人工智能模型，如ChatGPT3.5，这些模型有可能帮助医疗专业人员提供准确和及时的信息[17]。研究人员也发现有更多的患者倾向于选择与AI聊天机器人进行交流和参与症状评估[18]，形成了患者主动学习的模式，大量节省医护人员的人力和临床实践，有助于提高医疗服务利用率和服务质量。大部分儿童口腔疾病可以在早期通过科学的口腔卫生观念和行为指导进行预防，因此科学地评估并改进大语言模型回答儿童口腔预防保健问题的质量，有助于减少儿童口腔疾病发病率，提升儿童生活质量，减少其所在家庭及社区的社会经济成本及心理负担。本研究旨在评估和比较ChatGLM-6B与ChatGPT3.5在儿童口腔预防医学领域问题回答的准确性，以期能对国内口腔医疗大语言模型的研究提供帮助。

近年来，大语言模型在妇科、眼科等其他医学专业领域的类似研究层出不穷。Sütcüoğlu等[19]对25个与卵巢功能不全有关的答案进行了分析，其结果显示出很高的正确率（76%）。Mihalache等[20]对ChatGPT回答眼科董事会认证的考试题目答案进行分析，其结果显示ChatGPT对125个多项选择题中的73个（58%）和78个单项选择题中的42个（54%）提供了正确答案。但大语言模型在口腔医学领域中的应用潜力尚未被充分研究，有限的研究集中在口腔颌面外科和牙体牙髓病专业。Suárez等[21]对30个口腔手术问题提供的答案进行了分析，其结果也显示出较高的正确率（71.7%），本研究结果（68%）与其相似。Balel[22]和Ayers等[18]的研究也证实了这一点，但其认为大语言模型目前仅限于提供理论知识和客观事实，是否能够根据现实病例提供临床建议与指导还需进一步调试和验证。Mohammad-Rahimi等[23]发现大语言模型对于牙体牙髓病学问题的回答基本准确，但少部分回答违背科学事实，多次提问生成答案不一致，存在误导非专业使用者的可能。

在儿童口腔医学领域，Rokhshad等[24]对两名儿童专科口腔医生提出的30个儿童口腔领域问题应用ChatGPT3.5进行回答，结果表明其正确率为78% ± 3%，比本研究的正确率更高，且其他公开访问的聊天机器人软件（Google Bard、ChatGPT4、Llama、Sage、Claude 2 100k、Claude-instant、Claude-instant-100k和Google Palm）与ChatGPT3.5无组间差异。Gugnani等[25]采用Likert量表判断ChatGPT是否可以准确回答父母提出的儿童口腔相关问题，认为ChatGPT能够比较完整、清晰、合乎逻辑地回答实际问题。

本研究采用多名医生对大模型生成答案评分的形式，旨在模拟和计算模型输出的科普形式与专业人员认知之间的差异及其局限性。本研究进行统一提问并选用模型初次生成的答案进行评判，模拟非人工智能专业用户使用习惯，还原真实使用场景，并避免不同提问方式和答案筛选对答案评价的影响。参与答案评分的16名口腔医生均具有5年以上口腔专业训练，故试验中未预先形成针对100个儿童口腔预防问题的标准答案，避免其对口腔医生的评分产生影响。16名口腔医生分别对大语言模型生成的100个答案进行评分，对既往同类研究[21,24]的样本量进行扩充；其评分通过一致性检验，但仍具有提升空间。16名评估者均为受过标准专业训练的口腔医生，但对于评估标准的理解差异难免存在，未来可以选择儿童口腔专业人员进行一致性培训和评价，或开始前对评估者进行医学标准答案知识的相关培训，提高评分一致性，使大语言模型答案质量评估的有效性与可靠性得到保障。

总体而言，ChatGPT3.5和ChatGLM-6B在回答

儿童口腔预防医学领域问题类型和难度方面正确率相似，均近似达到70%。ChatGPT3.5和ChatGLM-6B回答程度较复杂专业问题的准确性似乎低于程度较简单的专业问题，这表明大模型在处理复杂的医疗问题方面可能仍存在局限性，尽管实验数据上并未显示统计学差异。本研究中，ChatGLM-6B与ChatGPT3.5的统计学结果大致相似，二者之间未看到显著的差异，这意味着ChatGLM-6B可能可以像ChatGPT3.5那样对不同难度的开放式问题提供广泛的适用性。

本研究存在一些局限性。首先，口腔医生的评估具有主观性。通过扩大评价人数的方式可减小主观性造成的影响，但这依然可能会导致调查结果存在偏差。其次，对大语言模型生成随机性的考量较局限。课题组为模拟大语言模型真实使用场景而选取其初次生成答案进行评价，相对忽略了大语言模型的随机性问题。由于模型输出受到温度、top-k采样等超参数的显著影响，单次生成的结果可能无法完全反映模型的整体性能。再次，问题集的广度和代表性有限。研究中选用100个儿童口腔预防医学领域问题的主要目的是评价大语言模型在该领域问答的正确性，而非形成用于评价大模型问答能力的标准化量表。采用的数量和问题类型可能难以全面代表临床实际中的复杂情况。未来可对问卷进一步调整和验证，筛选合适数量和质量的问题，揭示模型在更复杂和多样化情境下的表现，进一步提升研究结果的外部适用性。最后，评分体系存在主观性局限。Likert量表评分体系依赖于评审者的主观判断，异质性相对较大[8, 26-27]，难以完全避免个人偏见的影响。与既往大模型相关研究类似[28-29]，本研究选择3点Likert量表，以提供更直接的答案评估。在未来的研究中可以采用5点或7点量表，以更细致地划分答案的准确性等级，并减少评分者之间的主观差异。但医学领域对准确性要求极高，模型性能评估真实性的影响难以完全避免。有文献报道为评估大语言模型在临床环境中的表现而设计了校准量表，但这些量表主要面向需要对病史、症状、检查、诊断和治疗计划进行全面评估的复杂病例，而非针对患者就特定疑虑进行咨询[30]。所以应谨慎解读这项研究的结果。

今后有以下方向待改进。①扩大模型选择范围，未来研究将纳入ChatGPT-4等已在多个领域表现出更强推理能力和更高准确性的大语言模型，将其纳入对比能更全面地揭示当前模型在医学问题回答中的表现差异，并更好地为未来的发展方向提供依据。②增加模型直接对比，设定“win、lose、tie”，直接比较大模型答案，计算模型胜率。③对生成策略超参数调优进行深入探讨。对温度、top-k采样等超参数的设置对模型输出的影响进行详细讨论，充分发掘模型潜力。

综上，本研究展示了ChatGLM-6B与ChatGPT3.5在回答儿童口腔预防医学问题方面的潜力。尽管ChatGLM-6B回答儿童口腔预防医学领域问题取得了与ChatGPT3.5相似的表现，但二者正确率均未达到预期，不能应用于临床。在未来，需要进一步研究提升大语言模型提供医疗信息的准确性和一致性，并研发适用于口腔医学领域的医疗大模型。

【Author contributions】 Guan BY reviewed previous studies, collected, analyzed the data, wrote and revised the article. Xu MH, Zhang HQ collected, analyzed the data, and wrote the first draft. Ma SL collected and analyzed the data. Zhang SS designed the study, guided and critically reviewed the article structures. Zhao JF guided and gave critical advice towards the study. All authors read and approved the final manuscript as submitted.

参考文献

[1] Murphy Lonergan R, Curry J, Dhas K, et al. Stratified evaluation of GPT's question answering in surgery reveals artificial intelligence (AI) knowledge gaps[J]. Cureus, 2023, 15(11): e48788. doi: 10.7759/cureus.48788.

[2] Miao H, Li C, Wang J. A future of smarter digital health empowered by generative pretrained transformer[J]. J Med Internet Res, 2023, 25: e49963. doi: 10.2196/49963.

[3] Gurrapu S, Kulkarni A, Huang L, et al. Rationalization for explainable NLP: a survey[J]. Front Artif Intell, 2023, 6: 1225093. doi: 10.3389/frai.2023.1225093.

[4] Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models[J]. PLoS Digit Health, 2023, 2(2): e0000198. doi: 10.1371/journal.pdig.0000198.

[5] Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States medical licensing examination (USMLE)? The implications of large language models for medical education and knowledge assessment[J]. JMIR Med Educ, 2023, 9: e45312. doi: 10.2196/45312.

[6] Seth I, Cox A, Xie Y, et al. Evaluating chatbot efficacy for answering frequently asked questions in plastic surgery: a ChatGPT case study focused on breast augmentation [J]. Aesthet Surg J, 2023, 43 (10): 1126-35.doi: 10.1093/asj/sjad140.

[7] Whiles BB, Bird VG, Canales BK, et al. Caution! AI bot has entered the patient chat: ChatGPT has limitations in providing accurate urologic healthcare advice [J]. Urology, 2023, 180: 278-84.

doi: 10.1016/j.urology.2023.07.010.

[8] Yeo YH, Samaan JS, Ng WH, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma [J]. Clin Mol Hepatol, 2023, 29(3): 721-32.doi: 10.3350/cmh.2023.0089.

[9] Alhaidry HM, Fatani B, Alrayes JO, et al. ChatGPT in dentistry: a comprehensive review[J]. Cureus, 2023, 15(4): e38317. doi: 10.7759/cureus.38317.

[10] 袁瑞, 司敏敏, 张印, 等. ChatGPT在口腔正畸教育和临床中的应用前景[J]. 口腔疾病防, 2024, 32(6): 478-484. doi: 10.12016/j.issn.2096-1456.2024.06.011
Yuan R, Si MM, Zhang Y, et al. Prospects of ChatGPT in orthodontic education and clinical practice[J]. J Prev Treat Stomatol Dis, 2024, 32(6): 478-484. doi: 10.12016/j.issn.2096-1456.2024.06.011.

[11] Thorp HH. ChatGPT is fun, but not an author[J]. Science, 2023, 379(6630): 313. doi: 10.1126/science.adg7879.

[12] Stokel-Walker C. ChatGPT listed as author on research papers: many scientists disapprove[J]. Nature, 2023, 613(7945): 620-621. doi: 10.1038/d41586-023-00107-z.

[13] Goodman RS, Patrinely JR Jr , Osterman T, et al. On the cusp: considering the impact of artificial intelligence language models in healthcare[J]. Med, 2023, 4(3): 139 - 140. doi: 10.1016/j.medj.2023.02.008.

[14] Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns[J]. Healthcare(Basel), 2023, 11(6): 887. doi: 10.3390/healthcare11060887.

[15] Sarker IH. AI-based modeling: techniques, applications and research issues towards automation, intelligent and smart systems[J]. SN Comput Sci, 2022, 3(2): 158. doi: 10.1007/s42979-022-01043-x.

[16] Korteling JEH, van de Boer-Visschedijk GC, Blankendaal RAM, et al. Human- versus artificial intelligence[J]. Front Artif Intell, 2021, 4: 622364. doi: 10.3389/frai.2021.622364.

[17] Moshirfar M, Altaf AW, Stoakes IM, et al. Artificial intelligence in ophthalmology: a comparative analysis of GPT-3.5, GPT-4, and human expertise in answering statpearls questions[J]. Cureus, 2023, 15(6): e40822. doi: 10.7759/cureus.40822.

[18] Ayers JW, 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. doi: 10.1001/jamainternmed.2023.1838.

[19] Sütcüoğlu BM, Güler M. Appropriateness of premature ovarian insufficiency recommendations provided by ChatGPT[J]. Menopause, 2023, 30(10): 1033-1037. doi: 10.1097/GME.0000000000002246.

[20] Mihalache A, Popovic MM, Muni RH. Performance of an artificial intelligence chatbot in ophthalmic knowledge assessment[J]. JAMA Ophthalmol, 2023, 141(6): 589-597. doi: 10.1001/jamaophthalmol.2023.1144.

[21] 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]. Comput Struct Biotechnol J, 2024, 24: 46-52. doi: 10.1016/j.csbj.2023.11.058.

[22] Balel Y. Can ChatGPT be used in oral and maxillofacial surgery? [J]. J Stomatol Oral Maxillofac Surg, 2023, 124(5): 101471. doi: 10.1016/j.jormas.2023.101471.

[23] Mohammad-Rahimi H, Ourang SA, Pourhoseingholi MA, et al. Validity and reliability of artificial intelligence chatbots as public sources of information on endodontics[J]. Int Endod J, 2024, 57(3): 305-314. doi: 10.1111/iej.14014.

[24] Rokhshad R, Zhang P, Mohammad-Rahimi H, et al. Accuracy and consistency of chatbots versus clinicians for answering pediatric dentistry questions: a pilot study[J]. J Dent, 2024, 144: 104938. doi: 10.1016/j.jdent.2024.104938.

[25] Gugnani N, Pandit IK, Gupta M, et al. Parental concerns about oral health of children: is ChatGPT helpful in finding appropriate answers?[J]. J Indian Soc Pedod Prev Dent, 2024, 42(2): 104-111. doi: 10.4103/jisppd.jisppd_110_24.

[26] Mago J, Sharma M. The potential usefulness of ChatGPT in oral and maxillofacial radiology[J]. Cureus, 2023, 15(7): e42133. doi: 10.7759/cureus.42133.

[27] 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), 2023, 13(11): 1950. doi: 10.3390/diagnostics13111950.

[28] Luykx JJ, Gerritse F, Habets PC, et al. The performance of ChatGPT in generating answers to clinical questions in psychiatry: a two-layer assessment[J]. World Psychiatry, 2023, 22(3): 479-480. doi: 10.1002/wps.21145.

[29] Ali MJ. ChatGPT and lacrimal drainage disorders: performance and scope of improvement[J]. Ophthalmic Plast Reconstr Surg, 2023, 39(3): 221-225. doi: 10.1097/IOP.0000000000002418.

[30] Lechien JR, Maniaci A, Gengler I, et al. Validity and reliability of an instrument evaluating the performance of intelligent chatbot: the artificial intelligence performance instrument (AIPI)[J]. Eur Arch Otorhinolaryngol, 2024, 281(4): 2063-2079. doi: 10.1007/s00405-023-08219-y.

（编辑　张琳）

官网