Journal of Multidisciplinary Healthcare

Dovepress
open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

Comparison of the Performance of ChatGPT, Claude and Bard in Support of Myopia Prevention and Control

Yan Wang[1], Lihua Liang[2,*], Ran Li[2,*], Yihua Wang[3], Changfu Hao[1]

[1]Department of Child and Adolescent Health, School of Public Health, Zhengzhou University, Zhengzhou, Henan, People's Republic of China; [2]Primary and Secondary School Health Center, Zhengzhou Education Science Planning and Evaluation Center, Zhengzhou Municipal Education Bureau, Zhengzhou, Henan, People's Republic of China; [3]Institute of Science and Technology Information, Zhengzhou University, Zhengzhou, Henan, People's Republic of China

*These authors contributed equally to this work

Correspondence: Yihua Wang, Institute of Science and Technology Information, Zhengzhou University, Science Avenue, Zhengzhou, Henan Province, 450001, People's Republic of China, Email wangyh2015@icloud.com; Changfu Hao, Department of Child and Adolescent Health, School of Public Health, Zhengzhou University, Science Avenue, Zhengzhou, Henan Province, 450001, People's Republic of China, Email haochangfu@126.com

Purpose: Chatbots, which are based on large language models, are increasingly being used in public health. However, the effectiveness of chatbot responses has been debated, and their performance in myopia prevention and control has not been fully explored. This study aimed to evaluate the effectiveness of three well-known chatbots—ChatGPT, Claude, and Bard—in responding to public health questions about myopia.

Methods: Nineteen public health questions about myopia (including three topics of policy, basics and measures) were responded individually by three chatbots. After shuffling the order, each chatbot response was independently rated by 4 raters for comprehensiveness, accuracy and relevance.

Results: The study's questions have undergone reliable testing. There was a significant difference among the word count responses of all 3 chatbots. From most to least, the order was ChatGPT, Bard, and Claude. All 3 chatbots had a composite score above 4 out of 5. ChatGPT scored the highest in all aspects of the assessment. However, all chatbots exhibit shortcomings, such as giving fabricated responses.

Conclusion: Chatbots have shown great potential in public health, with ChatGPT being the best. The future use of chatbots as a public health tool will require rapid development of standards for their use and monitoring, as well as continued research, evaluation and improvement of chatbots.

Keywords: chatbot, large language model, public health, myopia

Introduction

Myopia has become a major public health problem, affecting physical and mental health, learning and employment. It is the leading cause of visual impairment worldwide.[1] The prevalence of myopia was 39.6% in Korea,[2] 39.1% in France,[3] 41% in Canada,[4] and 65.48% among junior high school students in China.[5] It has been predicted that nearly 50% of the world's population will be myopic by 2050 (approximately 4.758 billion people).[6] Myopia usually occurs in childhood and adolescence and is irreversible once it occurs.[7] Delaying the age of onset of myopia reduces the risk of high myopia and even blindness later in life.[8,9] It is therefore very important to make progress in the early prevention and control of myopia.

Unfortunately, there is a general lack of public awareness of myopia.[10,11] Disease prevention and control authorities need to respond globally by improving spectacle coverage and related health education efforts.[12] In addition, public health resources vary from place to place. In underdeveloped regions, there is a shortage of public health personnel to provide systematic services. These personnel receive limited education and face challenges in updating their knowledge

Received: 13 April 2024
Accepted: 25 July 2024
Published: 13 August 2024

base.[13–15] The need for prevent and control myopia places enormous demands on the healthcare system. Effective use of available resources, especially artificial intelligence, is therefore crucial.

Chatbots can provide highly targeted and accessible responses, which can help public health workers improve efficiency, increase health response capacity and reduce workload.[16–19] Chatbots, which are based on Large Language Models, can draw on vast amounts of data to provide responses, support multiple languages, and interact in a human-like manner.[20] Notable chatbots include OpenAI's ChatGPT, Google's Bard and Anthropic's Claude.

However, chatbots do have their drawbacks and errors may be made in their responses.[21] In the field of ophthalmology, the quality of information provided by chatbots has already been evaluated. The evaluation has covered a range of ocular conditions, including myopia, lacrimal drainage disorders, macular degeneration, and other conditions.[22,23] The majority of this research has been conducted from a clinical diagnostic and care perspective. Nevertheless, the quality of the chatbot's responses to ophthalmological queries from a public health perspective, particularly with regard to myopia prevention and control, remains uncertain. Furthermore, there has been limited research conducted on Claude in comparison to ChatGPT and Bard. Developed by former OpenAI employees, Claude has similar use cases to those of ChatGPT and Bard, and it deserves to be evaluated.[24] To drive future research, understand the potential harms of chatbots and mitigate negative societal impacts, it is important to clearly evaluate the effectiveness of chatbots on specific problems.

The objective of this study was to evaluate and compare the effectiveness of three chatbots —ChatGPT, Claude and Bard— in responding to questions about myopia prevention and control with regard to word count, comprehensiveness, accuracy and relevance. This study aims to scientifically and accurately assess the ability of chatbots to assist public health workers. This study can provide a theoretical basis for chatbot developers and researchers to carry out subsequent optimization and application.

Methods

Ethics

As the data were publicly available and no patients were involved in the study, ethics committee approval was not required.

Questions Selection

The 19 questions for this study were selected from an article entitled "NCDC Releases | Technical Expert Answers on Public Health Interventions for Prevention and Control of Myopia in Children and Adolescents" by the Institute of Child and Adolescent Health of Peking University (https://mp.weixin.qq.com/s/c22KmC-XqJKlm2BuTLUt-Q). This institute is a technical support unit on child and adolescent health and school health for the Ministry of Education of the People's Republic of China, the National Health Commission of the People's Republic of China and others.[25] It has taken on the tasks of policy and regulation drafting, scientific research, technical guidance, management consultancy and training of professional and technical staff in China's national child and adolescent health and school health programs, and has a high degree of authority.[25] To understand the effectiveness of chatbot responses on different topics, the 19 questions are grouped into three topics: policy (8, 12, 14, 16, 17, 19), basics (1, 4, 5, 9, 11, 15, 18), and measures (2, 3, 6, 7, 10, 13).

Generating Responses

ChatGPT (version GPT-4.0, OpenAI, California, USA, https://chat.openai.com/), Claude (version Claude 2, Anthropic, California, USA, https://claude.ai/) and Bard (Google, California, USA, https://bard.google.com/) were used to respond to the questions from 20 October, 2023, to 22 October, 2023.

Prompt engineering is the method by which a chatbot' response is guided by the explicit definition of the desired style of output. This approach can enhance the quality and relevance of the response.[26,27] To standardize the responses, the same prompt (Appendix 1) was entered into the dialogue box before each question, which could place the chatbots in the role of a "public health expert". After the bots responded, the 19 questions were manually entered into the input boxes of the three chatbots in turn. In total, 57 responses were collected (Appendix 2). In order to avoid historical conversation interference, the chat box was reopened for each new response.

Content Evaluation

The responses relating to the prevention and control of myopia were evaluated by four raters (L.L., R.L., Y.W., C.H). who possessed a profound understanding of the subject matter. The raters collectively have an average of over 20 years' experience working in school health or academia, which provides them with a considerable reservoir of expertise and insight. To reduce bias, information about the chatbot characters was removed from the responses. Then, in the text given to the raters, the three responses from different chatbots for each question were sorted in random order. The chatbots' responses are scored independently by the raters for comprehensiveness, accuracy, and relevance. Each aspect is rated on a scale of 1–5, with higher scores indicating better. (Comprehensiveness: ① 1 point: incomplete, very little detail; ② 2 points: somewhat comprehensive, with basic detail; ③ 3 points: moderately comprehensive, with some detail; ④ 4 points: relatively comprehensive, with a fair amount of detail; ⑤ 5 points: very comprehensive, with exhaustive detail. Accuracy: ① 1 point: completely inaccurate; ② 2 points: more inaccurate than accurate; ③ 3 points: accurate and inaccurate are about equal; ④ 4 points: more accurate than inaccurate; ⑤ 5 points: completely accurate. Relevance: ① 1 point: very irrelevant; ② 2 points: irrelevant; ③ 3 points: somewhat relevant; ④ 4 points: fairly relevant; ⑤ 5 points: very relevant).

Statistical Analysis

Statistical analysis was performed using SPSS 21.0 (SPSS, Chicago, Illinois, USA). The mean score of the four raters was calculated for each question. The composite score is the mean of the comprehensiveness, accuracy and relevance scores. Rating results were described using indices of central tendency (mean and median) and indices of variability (minimum, maximum, standard deviation, standard error and coefficient of variance). Cronbach's α and intraclass correlation coefficient (ICC) were calculated to assess the reliability of the questions. To assess the responses, the scores were compared using a one-way ANOVA or the Kruskal–Wallis test, as appropriate. Bonferroni corrections for multiple comparisons were performed. A p-value of less than 0.05 was considered statistically significant.

Results

Reliability of Questions

The internal consistency reliability coefficients Cronbach's α, and ICC for the question as a whole and the three topics are shown in Table 1. All Cronbach's α values exceeded 0.7, and all ICCs were statistically significant. This indicates a high level of reliability for the questions as a whole and three topics of policy, basics and measures.

Word Count in Response

Figure 1 shows the word count for each response generated by the three chatbots. Table 2 presents descriptive statistics on the word count in the responses of the three chatbots. The Kruskal–Wallis test was used to find statistically significant differences in the word count in the responses of the three chatbots. Significant differences among the word count of all three chatbot responses were found in post-hoc analyses using Bonferroni adjustment. ChatGPT produced the highest word count (median = 479), followed by Bard (median = 334), with Claude producing the lowest (median = 205) (Table 3).

Table 1 Reliability of Questions

	Number of questions	ICC single	p value	ICC average	p value	Cronbach's α
All topics	19	0.406	<0.001***	0.928	<0.001***	0.940
Policy	6	0.512	<0.001***	0.863	<0.001***	0.867
Basics	7	0.280	<0.001***	0.731	<0.001***	0.799
Measures	6	0.476	<0.001***	0.845	<0.001***	0.842

Notes: ***p<0.001.

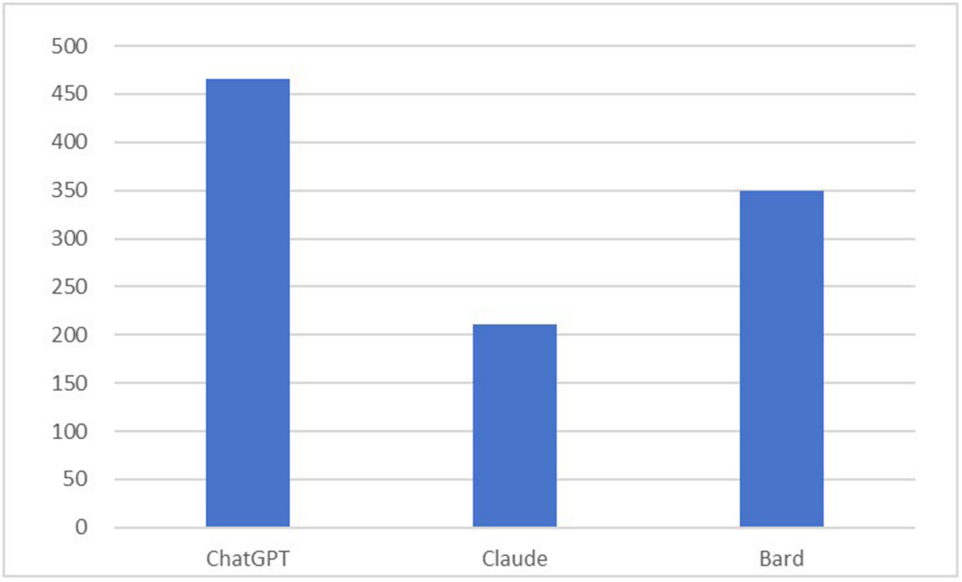

Figure 1 Word count in response from the three chatbots.

Scores

Scores of Different Aspects

Appendix 3 lists the scores given by the four raters for the three chatbot responses. Figure 2 shows the comprehensiveness, accuracy, relevance, and composite scores of the three chatbots. The mean scores of all three chatbots were above 4

Table 2 Word Count in Response from the Three Chatbots

	ChatGPT	Claude	Bard
Minimum	327	152	215
Median	479	205	334
Maximum	544	273	564
Mean	465.421	211.053	348.895
SD	60.475	30.325	84.173
SE	13.874	6.957	19.311
Coefficient of Variance	0.130	0.144	0.241
Rank mean	45.53	10.74	30.74
H	42.054		
p value	<0.001***		

Notes: ***p<0.001.
Abbreviations: SD, standard deviation; SE, standard error;

Table 3 Post-Hoc Analyses of the Word Count of the Responses for the Three Chatbots

	t	SE	Adjusted *p* value
ChatGPT vs Claude	34.789	5.385	<0.001***
ChatGPT vs Bard	14.789	5.385	0.018*
Claude vs Bard	−20.000	5.385	0.001**

Notes: *p<0.05; **p<0.01; ***p<0.001.

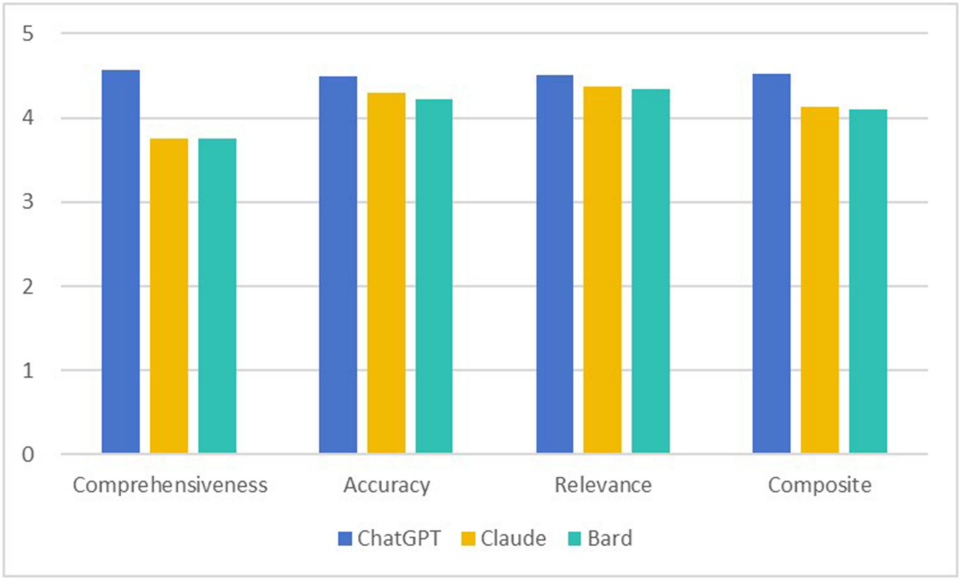

Figure 2 Different aspect scores of three chatbots' responses.

for accuracy, relevance and composite scores. ChatGPT had the highest mean scores for comprehensiveness, accuracy, relevance and composite, which were 4.566, 4.487, 4.513, and 4.522, respectively.

A one-way ANOVA revealed statistically significant differences in the composite and comprehensiveness scores among the three chatbots (Table 4). Post-hoc analyses using the Bonferroni adjustment showed that ChatGPT was significantly higher than Claude and Bard in the composite and comprehensiveness scores, while there was no difference between Claude and Bard. There was no difference among the three chatbots for accuracy and relevance scores (Table 5).

Scores of Different Topics

Figure 3 shows the composite scores of the three chatbots in the policy, basics and measures topics. The mean scores of all three chatbots were above 4 in the topics of policy and measures. ChatGPT had the highest mean scores in the topics of policy, basics and measures with 4.542, 4.452, and 4.583, respectively.

A one-way ANOVA showed statistically significant differences among the three chatbots in the policy and measures topics. There were no differences among the three chatbots on the basics topic (Table 6). Post-hoc analyses using the Bonferroni adjustment showed that ChatGPT's scored significantly higher than Claude's in the policy topic. Meanwhile,

Table 4 Scores of the Three Chatbots

	Comprehensiveness			Accuracy			Relevance			Composite		
	ChatGPT	Claude	Bard	ChatGPT	Claude	Bard	ChatGPT	Claude	Bard	ChatGPT	Claude	Bard
Minimum	3.5	3	3	3.25	3.5	2.5	3.5	3.5	3.5	3.417	3.417	3.083
Median	4.5	3.75	3.75	4.5	4.5	4.25	4.5	4.5	4.5	4.583	4.167	4.167
Maximum	5	4.5	4.25	4.75	4.75	4.5	4.75	4.75	5	4.833	4.500	4.417
Mean	4.566	3.750	3.750	4.487	4.289	4.224	4.513	4.368	4.342	4.522	4.136	4.105
SD	0.380	0.373	0.373	0.395	0.346	0.478	0.294	0.305	0.346	0.305	0.274	0.327
SE	0.087	0.085	0.085	0.091	0.079	0.110	0.068	0.070	0.079	0.070	0.063	0.075
Coefficient of Variance	0.083	0.099	0.099	0.088	0.081	0.113	0.065	0.070	0.080	0.068	0.066	0.080
F	29.927			2.120			1.617			11.161		
p value	<0.001***			0.130			0.208			<0.001***		

Notes: ***, p<0.001.
Abbreviations: SD, standard deviation; SE, standard error.

Table 5 Post-Hoc Analyses of the Comprehensiveness and Composite Scores of the Three Chatbots

	Difference in means	SE	*p* value
Comprehensiveness			
ChatGPT vs Claude	0.386	0.098	0.001**
ChatGPT vs Bard	0.417	0.098	<0.001***
Bard vs Claude	0.031	0.098	1.000
Composite			
ChatGPT vs Claude	0.816	0.122	<0.001***
ChatGPT vs Bard	0.816	0.122	<0.001***
Bard vs Claude	<0.001	0.122	1.000

Notes: **p<0.01; ***p<0.001.

ChatGPT scored significantly higher than Claude and Bard on the topic of measures, while there was no difference between Claude and Bard (Table 7).

Discussion

Performance of Chatbot Responses to Myopia Prevention and Control Questions

Chatbots can respond to questions from different perspectives. For example, the responses to question 16 were informative, with ChatGPT dividing the focus group into five categories, Claude dividing the focus group into six categories, and Bard dividing the focus group into three categories for suggestions. In addition, the chatbots, especially ChatGPT, often provided additional content related to the question. This addition increased the comprehensiveness of the responses. Ethical considerations were mentioned in some of the chatbots' responses. For example, ChatGPT mentions "respect individual and cultural differences" in question 10 and ethical considerations in questions 7, 8, 14, 15 and 19; Claude mentions "respect student privacy" in question 14, also with ethical considerations.

The accuracy of chatbots' responses is as follows. To illustrate, in question 3, " If the duration of outdoor activities fails to meet the recommended time, will it still be effective in preventing myopia?" A substantial body of evidence from scientific studies indicates that engagement in outdoor activities exerts a protective influence against the development of

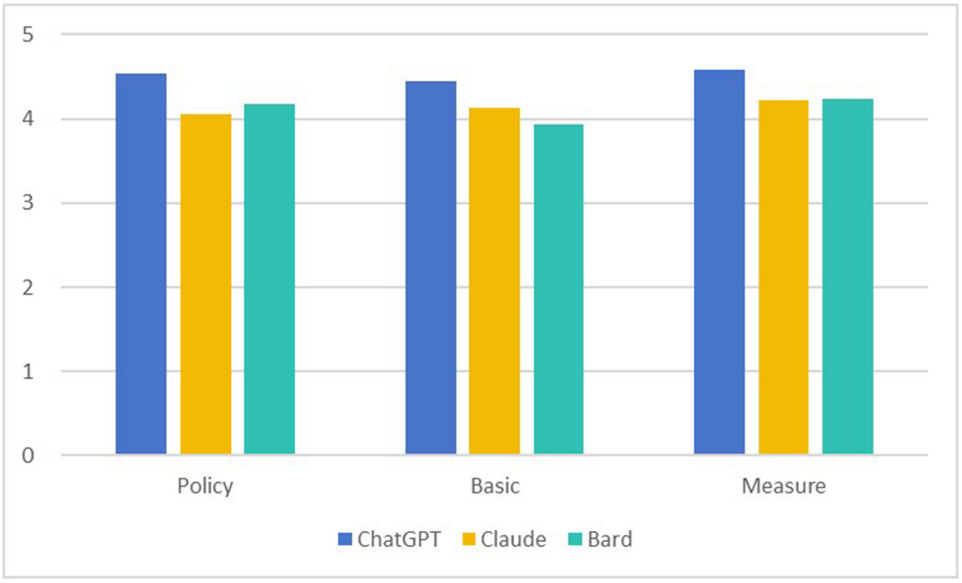

Figure 3 Different topic scores of three chatbots' responses.

Table 6 Composite Scores of the Three Chatbots on Different Topics

	Policy			Basics			Measures		
	ChatGPT	Claude	Bard	ChatGPT	Claude	Bard	ChatGPT	Claude	Bard
Minimum	4.250	3.417	4.083	3.417	3.583	3.083	4.500	4.167	4.000
Median	4.583	4.167	4.125	4.583	4.167	4.167	4.542	4.167	4.250
Maximum	4.750	4.333	4.417	4.750	4.500	4.333	4.833	4.417	4.417
Mean	4.542	4.056	4.181	4.452	4.131	3.929	4.583	4.222	4.236
SD	0.209	0.348	0.134	0.469	0.319	0.480	0.129	0.101	0.144
SE	0.085	0.142	0.055	0.177	0.120	0.181	0.053	0.041	0.059
Coefficient of Variance	0.046	0.086	0.032	0.105	0.077	0.122	0.028	0.024	0.034
F	6.272			2.658			15.878		
p	0.01*			0.097			<0.001***		

Notes: *p<0.05; ***p<0.001.
Abbreviations: SD, standard deviation; SE, standard error.

Table 7 Post-Hoc Analyses of the Three Chatbots in the Topics of Policy and Measures

	Difference in means (I-J)	SE	*p* value
Policy			
ChatGPT vs Claude	0.486	0.143	0.012*
ChatGPT vs Bard	0.361	0.143	0.069
Bard vs Claude	−0.125	0.143	1.000
Measures			
ChatGPT vs Claude	0.361	0.073	0.001**
ChatGPT vs Bard	0.347	0.073	0.001**
Bard vs Claude	−0.014	0.073	1.000

Notes: *p<0.05; **p<0.01.

myopia in children.[28,29] This effect might be attributed to the distinctive properties of natural light and the release of retinal dopamine.[30] All three chatbots indicated that participation in outdoor activities can contribute to the prevention of myopia, even if the time spent outdoors does not reach the recommended threshold. Furthermore, ChatGPT highlighted the rationale, indicating that exposure to outdoor natural light stimulates the retina to release dopamine. These responses are in accordance with the results of scientific studies.[28–30]

In regards to question 18, certain companies have made claims on the internet that anti-blue light glasses can prevent myopia. However, these claims are misleading. Some studies have indicated that blue light blocking glasses may alleviate eye fatigue.[31,32] Nevertheless, no studies have documented the protective impact of blue light blocking glasses on myopia progression in school-aged children.[33] Fortunately, all three chatbots responded objectively. ChatGPT stated that the effectiveness of anti-blue light glasses in preventing myopia is not well-established in the scientific literature. Claude replied that the evidence is still inconclusive overall. Bard suggested proposed that more research is needed to determine if blue light-blocking glasses are effective in preventing myopia. The chatbots also warned about the possibility of inaccuracies in their responses. For questions 3, 4 and 18, ChatGPT recommended that the questioner consult scientific literature or guidelines. Bard recommended that the ophthalmologist should be informed, as stated in the responses to question 18.

The responses provided by the chatbots are frequently pertinent and directly address the posed questions.

While chatbots offer several advantages, they also present certain limitations. Specifically, the responses provided by Claude and Bard were frequently less comprehensive in scope and lacked the same degree of detail as those generated by

ChatGPT. This suggests that they are deficient in providing detailed information in comparison to ChatGPT. In addition, all three chatbots occasionally produced incorrect content. For example, in question 4, ChatGPT and Bard misinterpreted the term "hyperopia reserve insufficiency". Although Claude understood this term correctly, it incorrectly defined "premyopia" as "between −0.50 to −0.75 diopters" (actually "-0.50 to +0.75 diopters").[34] Chatbot's responses are sometimes not relevant enough. For example, in Claude's response to question 17, the statement "institutional infection control and safety practices" refers to infections but not to myopia, and the other sentences do not refer specifically to vision or myopia. Sometimes the chatbots responded in general terms.

The study by Sayantan and others[23] also addressed issues related to myopia. The results demonstrated that over 70% of the responses from the chatbot were rated as either good or very good. In particular, ChatGPT achieved median scores of 3.0 and 4.0 (out of 5) for the prevention and treatment of myopia, respectively. However, the median ChatGPT score on basics topics (mainly prevention and treatment related basics) in this study was 4.583. The possible reasons for the differences in scores may be attributed to the following factors. Firstly, there are discrepancies in the methodology employed in the calculation of the scores. In Sayantan's study, the score was the median of the five raters' scores for a single question on either the prevention or treatment of myopia. In this study, however, the score was the median of the composite scores for all questions within the base topic. In the study conducted by Sayantan, the score represented each rater's assessment of the content of the question as a whole. In contrast, in the present study, the composite score was derived by calculating the mean of the four raters' scores for each aspect of each question, and subsequently calculating the mean of the scores for these three aspects. Secondly, there is a distinction to be made in terms of the scope of the questions. In the study conducted by Sayantan, the questions were of a broader nature, for example, " How to prevent myopia?" In contrast, the questions posed in the present study were more specific in nature, for example, " What is the principle behind using outdoor activities to prevent myopia?"

Differences Among the Three Chatbots

ChatGPT, Claude and Bard have different strengths. ChatGPT is a chatbot publicly released by OpenAI in November 2022 that uses the Transformer architecture.[24] ChatGPT-3 contains 175 billion parameters.[35] In March 2023, ChatGPT-4.0 was released with improved linguistic generation and multi-turn dialogue.[36] In this study, ChatGPT provided the most complete and comprehensive responses.

Claude was Anthropic's March 2023 chatbot release, which also uses the Transformer architecture.[21] Claude2 was released in July 2023, and it performs well in coding, mathematics and reasoning.[37] In the study by Alkuraya and others[38] dealing with complex genetic problems, once did Claude appeared to respond completely correctly. In addition, Claude2's context window (the amount of information that can be processed in a single chat) is up to 100K, which means that it is capable of processing about 75,000 words in a single command (about 272 pages of a book).[37] Standing out from ChatGPT-4.0 (8K or less) and Bard (2K or less), Claude2 can ingest entire books for health professional and process long-term disease data.[37]

Bard, a chatbot released by Google in March 2023, is powered by LaMDA and trained on 540 billion parameters.[39] It is trained in dialogue and can distinguish between open and closed discussions.[40] Bard can keep up with the latest information on the Internet and provide real-time, up-to-date content in a way that ChatGPT and Claude cannot.

In this study, ChatGPT was the most effective at responding to public health questions related to myopia and had the highest mean scores for comprehensiveness, accuracy and relevance. A review article showed that compared to other chatbots ChatGPT performed optimally in answering ophthalmology questions.[22] In other chatbot comparison studies, ChatGPT was found to be more effective than Bard in medical examinations,[41–43] myopia care[44] and neurosurgery.[45] ChatGPT outperformed Bard and Claude on the Licensing Medical Examination[46] and in Medical Arthropodology learning objectives.[47] ChatGPT and Claude are equally optimal on candidate gene selection[48] and rheumatic disease identification.[37] Bard and ChatGPT were found to have higher response accuracy, whereas Claude performed better in empathy and expression.[49]

In conclusion, each of the three chatbots possesses distinctive strengths, as evidenced by their respective performance in diverse scenarios. In the case of ChatGPT, in particular, its excellence has been widely acknowledged by the academic community, with numerous studies confirming its efficacy and impact in a range of application areas.[37,41–49]

Role of Chatbot

Myopia prevention is urgent. The three chatbots in this study provided comprehensive, accurate and relevant responses about myopia prevention policy, basics and measures.[1] Chatbots can interact with the public and provide convenient, easy-to-understand, and targeted information to improve the public's health literacy and ophthalmic knowledge. Chatbots can provide myopia health education, explaining the rationale and offering advice to different populations. Chatbots can also assist public health personnel by providing myopia prevention strategies, planning activities, improving work efficiency, and initiating various myopia control measures. On the economic side, the direct cost of myopia correction in the United States alone is at least $3.8 billion per year, which could be reduced through the use of chatbots.[50]

Similarly, other studies have shown that chatbots play an important role in public health. Ayers and others[51] conducted a study in which they evaluated the responses of doctors and chatbots to patient questions taken from a forum. The study found that chatbots provided 3.6 times the number of high-quality responses and 9.8 times the number of empathetic responses, outperforming doctors in 78.6% of responses. Ayers and others[52] also found that 91% of ChatGPT's responses to public health questions were identified as evidence-based. In addition, 22% of responses referred to relevant government-recommended resources such as helplines. In their study, Duong and others[53] found that ChatGPT's accuracy (68.2%) in responding to 85 multiple-choice questions about human genetics was not statistically different from that of humans (66.6%), with the chatbot's accuracy slightly higher.

Chatbots can be a valuable tool for disease prevention, health advice and public health decision-making. However, it is important to note that chatbots should not be relied upon as the sole source of public health advice.

Shortcomings of Chatbots

Content Errors

The most significant issue identified in this study was the inclusion of fabricated information in the chatbots' responses. This issue, known as "hallucination", involves generating text that, while semantically or syntactically correct, is factually incorrect or meaningless.[54] Similar to this study, McGowan and others[55] found that ChatGPT-3.5 generated citations with only 6% accuracy, and Bard 2.0 generated citations with all errors. Only 33.0% of the responses generated by ChatGPT were correct on the French medical school entrance examination.[56] What's more, the study found that chatbots would analyze questions correctly but draw the wrong conclusions. For example, Alkuraya and others[38] found that when responding to a question about the simultaneous risk of two diseases, ChatGPT, Claude, and Bard all inaccurately concluded the probability of having a healthy child, despite correctly identifying the inheritance patterns of diseases.

Possible reasons for the hallucinations are listed below. (1) Text generation method: Illusion is inherent in chatbots.[57] This is because the chatbot's output is probabilistic.[55] It decomposes the incoming text into words and subwords, uses an attention mechanism to weigh the importance of different parts of the input sequence, and then iteratively searches for words that occur at the same time as those words until it generates a coherent output sequence.[58] (2) Cannot be updated in real-time: Some chatbots are trained on datasets from a specific period and cannot refer to new research or guidelines in real-time.[59] (3) Lack of specific expertise: Chatbots are trained on general text data and do not have access to professional databases.[60] (4) Difficulty dealing with complex situations: Medical issues often require reasoning based on multiple variables. Chatbots may have difficulty understanding and accurately explaining the complexity of certain studies.[61] (5) Sensitive wording of questions: If questions are not worded precisely enough, chatbots will not be able to understand them correctly.[62]

Others

Claude and Bard's responses are not as comprehensive as ChatGPT's. This may be due to the fact that ChatGPT was trained on a larger dataset with a more extensive and diverse data.

Three chatbots sometimes responded in generalities, and similar situations have occurred in other studies. For example, ambiguous advice was given by 4 chatbots in clinical medicine in the study by Wilhelm and others.[63] In the study by Gao and others[64] the evaluators distinguished between abstracts generated by the chatbot and those generated by humans. They found that the chatbot-generated abstracts were superficial, ambiguous and occasionally overly detailed.

Chatbots also have some hidden dangers. (1) Dangers in input data: If a user enters a malicious message, the chatbot will display an error message to other users. This is because the information the user enters into the chatbot can be incorporated into the model.[65] (2) Privacy issues: Similarly, the private information entered by the user will be included in the training set, which will then be available to other questioners, resulting in a breach of privacy.[66] (3) Language limitations: Chatbots have limited ability to respond to questions in less commonly spoken languages.[67]

Prospects

Chatbots, as a tangible manifestation of artificial intelligence, enable interaction with users through the utilization of technologies such as natural language processing (NLP) and machine learning (ML).[68] Currently, they are being used in a wide range of fields, including healthcare,[67] finance,[69] management,[70] sociology[71] and education.[72] They bring new ideas and challenges to the field of public health, because of their ability to quickly retrieve information and make decisions. The utilization of AI-powered chatbots has the potential to reduce the workload of public health professionals, particularly in the context of myopia prevention and control. However, it is important to recognize that such technology cannot replace the expertise of human professionals.[73] Critical thinking is required when using them.

Considering the current circumstances, potential avenues for further research directions can be identified in the following aspects. The initial direction is to explore further standardization of the application of future chatbots in public health. (1) Developing a standardized and rigorous data management and validation process for chatbots is crucial, along with establishing guidelines for verifying the authenticity of responses. This will make it easier to assess the accuracy of the information provided. (2) To ensure greater accuracy and reliability, manual management should be included in content validation and fact-checking. (3) Collaboration between public health agencies and AI companies is necessary to ensure that chatbots meet the needs and requirements of the health field. It is important to involve experts in public health and related fields in the development and evaluation process of chatbots. Public health experts can suggest appropriate databases to be included in the training set to enhance the chatbots' ability to answer public health queries. (4) To ensure the protection of user privacy, it is important to establish clear guidelines and an ethical framework for the collection, storage and use of user input data. Sensitive information in question should be handled appropriately, and data anonymization should be ensured. Second, further research could explore how chatbots can utilize precise data to provide personalized prevention and control advice to individuals. For example, a chatbot could identify ocular behavior based on the duration of time a user's wearable device records the use of electronic devices, the length of time spent outdoors, and other factors, while suggesting rest periods in conjunction with the results of a vision assessment. Thirdly, further research is required to determine how professional medical resources could be integrated in order to enhance their role in public health management. For example, chatbots collaborate with ophthalmologists to facilitate remote consultations when required.

As performance and policy continue to improve, chatbots will undoubtedly play an increasingly important role in public health. Researchers predict that by 2030, AI will influence 14% of the world's products, half of which will increase productivity.[74] Chatbots can assist in health decision-making, provide personalized advice, conduct health surveillance, predict diseases, enable telemedicine, optimize intervention strategies, facilitate health education and more.[75,76] It has the potential to revolutionize public health practices and enhance the effectiveness of public health system management.

Strengths and Limitations

The advantages of this study are as follows. First, it explored the performance of chatbots on myopia-related public health issues, evaluating policy concerns in myopia prevention and control that have not been previously studied. Second, it simultaneously compared three chatbots, including Claude. Despite their recognition, few studies have evaluated the effectiveness of Claude's responses. Furthermore, in the research design, any characteristic information of the chatbots has been excluded from the responses, which have been randomly assigned. This ensures the study's reliability and reduces bias. Finally, this study evaluated the chatbots based on four key criteria: word count, comprehensiveness, accuracy, and relevance. The evaluation was detailed and meticulous, covering several key dimensions.

Nevertheless, the study has several limitations. First, the ratings are inconsistent in some responses and may contain subjective evaluations. Therefore, the mean scores of the four raters were used for analysis in this study. Secondly, only one

response was evaluated for each question in this study. However, chatbots often provide different responses to the same question, which could potentially affect the reliability of the results.[53] So, to guarantee the reliability of the results, a series of questions on each topic were posed to the chatbots. Finally, the results of this study can only represent the current situation. Chatbots continue to improve through iterations and database updates, and results may change in the future.

Conclusion

Overall, the study demonstrated that the three chatbots were effective at responding to public health questions about myopia. ChatGPT's responses outperformed those of Claude and Bard in terms of comprehensiveness, accuracy and relevance. Chatbots can play an important role in public health as powerful tools to improve efficiency. However, chatbots also have limitations, such as fabricated responses, and need to be used with caution. Moving forward, it is imperative to promptly develop standards for the use and management of chatbots. Further research, validation, and enhancement of chatbots for public health applications are essential.

Disclosure

The authors report no conflicts of interest in this work.

References

1. Bourne RRA, Stevens GA, White RA, et al. Causes of vision loss worldwide, 1990-2010: a systematic analysis. *Lancet Glob Health*. 2013;1(6):e339–e349. doi:10.1016/S2214-109X(13)70113-X
2. Rim TH, Kim SH, Lim KH, et al. Refractive errors in Koreans: the Korea national health and nutrition examination survey 2008-2012. *Korean J Ophthalmol*. 2016;30(3):214. doi:10.3341/kjo.2016.30.3.214
3. Matamoros E, Ingrand P, Pelen F, et al. Prevalence of myopia in France A cross-sectional analysis. *Medicine (Baltimore)*. 2015;94(45):e1976. doi:10.1097/MD.0000000000001976
4. Hrynchak PK, Mittelstaedt A, Machan CM, Bunn C, Irving EL. Increase in myopia prevalence in clinic-based populations across a century. *Optom Vis Sci*. 2013;90(11):1331–1341. doi:10.1097/OPX.0000000000000069
5. Li Y, Liu J, Qi P. The increasing prevalence of myopia in junior high school students in the Haidian district of Beijing, China: a 10-year population-based survey. *BMC Ophthalmol*. 2017;17(1):88. doi:10.1186/s12886-017-0483-6
6. Holden BA, Fricke TR, Wilson DA, et al. Global prevalence of myopia and high myopia and temporal trends from 2000 through 2050. *Ophthalmology*. 2016;123(5):1036–1042. doi:10.1016/j.ophtha.2016.01.006
7. Cumberland PM, Peckham CS, Rahi JS. Inferring myopia over the lifecourse from uncorrected distance visual acuity in childhood. *Br J Ophthalmol*. 2007;91(2):151–153. doi:10.1136/bjo.2006.102277
8. Hu Y, Ding X, Guo X, Chen Y, Zhang J, He M. Association of age at myopia onset with risk of high myopia in adulthood in a 12-year follow-up of a Chinese cohort. *JAMA Ophthalmol*. 2020;138(11):1129. doi:10.1001/jamaophthalmol.2020.3451
9. Baird PN, Saw SM, Lanca C, et al. Myopia. *Nat Rev Dis Primer*. 2020;6(1):99. doi:10.1038/s41572-020-00231-4
10. McCrann S, Flitcroft I, Lalor K, Butler J, Bush A, Loughman J. Parental attitudes to myopia: a key agent of change for myopia control? *Ophthalmic Physiol Opt*. 2018;38(3):298–308. doi:10.1111/opo.12455
11. Ang M, Flanagan JL, Wong CW, et al. Review: myopia control strategies recommendations from the 2018 WHO/IAPB/BHVI Meeting on myopia. *Br J Ophthalmol*. 2020;104(11):bjophthalmol–2019–315575. doi:10.1136/bjophthalmol-2019-315575
12. Li Q, Guo L, Zhang J, et al. Effect of school-based family health education via social media on children's myopia and parents' awareness a randomized clinical trial. *JAMA Ophthalmol*. 2021;139(11):1165. doi:10.1001/jamaophthalmol.2021.3695
13. Noknoy S, Kassai R, Sharma N, Nicodemus L, Canhota C, Goodyear-Smith F. Integrating public health and primary care: the response of six Asia–Pacific countries to the COVID-19 pandemic. *Br J Gen Pract*. 2021;71(708):326–329. doi:10.3399/bjgp21X716417
14. Irving G, Neves AL, Dambha-Miller H, et al. International variations in primary care physician consultation time: a systematic review of 67 countries. *BMJ Open*. 2017;7(10):e017902. doi:10.1136/bmjopen-2017-017902
15. Zhang T, Xu Y, Ren J, Sun L, Liu C. Inequality in the distribution of health resources and health services in China: hospitals versus primary care institutions. *Int J Equity Health*. 2017;16(1):42. doi:10.1186/s12939-017-0543-9
16. Kruk ME, Porignon D, Rockers PC, Van Lerberghe W. The contribution of primary care to health and health systems in low- and middle-income countries: a critical review of major primary care initiatives. *Soc Sci Med*. 2010;70(6):904–911. doi:10.1016/j.socscimed.2009.11.025
17. Amiri P, Karahanna E. Chatbot use cases in the Covid-19 public health response. *J Am Med Inform Assoc*. 2022;29(5):1000–1010. doi:10.1093/jamia/ocac014
18. Kowatsch T, Nißen M, Shih CHI, et al. *Text-Based Healthcare Chatbots Supporting Patient and Health Professional Teams: Preliminary Results of a Randomized Controlled Trial on Childhood Obesity*. ETH Zurich, Department of Management, Technology and Economics; 2017. doi:10.3929/ethz-b-000218776
19. Tudor Car L, Dhinagaran DA, Kyaw BM, et al. Conversational agents in health care: scoping review and conceptual analysis. *J Med Internet Res*. 2020;22(8):e17158. doi:10.2196/17158
20. Wang X, Sanders HM, Liu Y, et al. ChatGPT: promise and challenges for deployment in low- and middle-income countries. *Lancet Reg Health - West Pac*. 2023;41:100905. doi:10.1016/j.lanwpc.2023.100905
21. Coello CEA, Alimam MN, Kouatly R. Effectiveness of chatgpt in coding: a comparative analysis of popular large language models. *Digital*. 2024;4(1):114–125. doi:10.3390/digital4010005

22. Biswas S, Davies LN, Sheppard AL, Logan NS, Wolffsohn JS. Utility of artificial intelligence-based large language models in ophthalmic care. *Ophthalmic Physiol Opt*. 2024;44(3):641–671. doi:10.1111/opo.13284
23. Biswas S, Logan NS, Davies LN, Sheppard AL, Wolffsohn JS. Assessing the utility of ChatGPT as an artificial intelligence-based large language model for information to answer questions on myopia. *Ophthalmic Physiol Opt*. 2023;43(6):1562–1570. doi:10.1111/opo.13207
24. Brown TB, Mann B, Ryder N, et al. Language models are few-shot learners. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS'20. Curran Associates Inc.; 2020.
25. School of Public Health, Peking University. the Institute of Child and Adolescent Health, Peking University, 2012. Available from: https://sph.pku.edu.cn/info/1046/2971.htm. Accessed September 20, 2023.
26. Ekin S. Prompt engineering for ChatGPT: a quick guide to techniques, tips, and best practices. *Authorea Prepr*; 2023.
27. Bridgelall R. Unraveling the mysteries of AI chatbots. *Artif Intell Rev*. 2024;57(4):89. doi:10.1007/s10462-024-10720-7
28. Muralidharan AR, Lança C, Biswas S, et al. Light and myopia: from epidemiological studies to neurobiological mechanisms. *Ther Adv Ophthalmol*. 2021;13:251584142110592. doi:10.1177/25158414211059246
29. Karthikeyan SK, Ashwini D, Priyanka M, Nayak A, Biswas S. Physical activity, time spent outdoors, and near work in relation to myopia prevalence, incidence, and progression: an overview of systematic reviews and meta-analyses. *Indian J Ophthalmol*. 2022;70(3):728–739. doi:10.4103/ijo.IJO_1564_21
30. Biswas S, El Kareh A, Qureshi M, et al. The influence of the environment and lifestyle on myopia. *J Physiol Anthropol*. 2024;43(1):7. doi:10.1186/s40101-024-00354-7
31. Vagge A, Ferro Desideri L, Del Noce C, Di Mola I, Sindaco D, Traverso CE. Blue light filtering ophthalmic lenses: a systematic review. *Semin Ophthalmol*. 2021;36(7):541–548. doi:10.1080/08820538.2021.1900283
32. Lawrenson JG, Hull CC, Downie LE. The effect of blue-light blocking spectacle lenses on visual performance, macular health and the sleep-wake cycle: a systematic review of the literature. *Ophthalmic Physiol Opt*. 2017;37(6):644–654. doi:10.1111/opo.12406
33. Cougnard-Gregoire A, Merle BMJ, Aslam T, et al. Blue light exposure: ocular hazards and prevention-a narrative review. *Ophthalmol Ther*. 2023;12(2):755–788. doi:10.1007/s40123-023-00675-3
34. Flitcroft DI, He M, Jonas JB, et al. IMI – defining and classifying myopia: a proposed set of standards for clinical and epidemiologic studies. *Investig Opthalmology Vis Sci*. 2019;60(3):M20. doi:10.1167/iovs.18-25957
35. Stokel-Walker C. AI bot chatgpt writes smart essays—should professors worry? *Nature*; 2022. doi:10.1038/d41586-022-04397-7
36. Sanderson K. GPT-4 is here: what scientists think. *Nature*. 2023;615(7954):773. doi:10.1038/d41586-023-00816-5
37. Venerito V, Puttaswamy D, Iannone F, Gupta L. Large language models and rheumatology: a comparative evaluation. *Lancet Rheumatol*. 2023;5(10):e574–e578. doi:10.1016/S2665-9913(23)00216-3
38. Alkuraya IF. Is artificial intelligence getting too much credit in medical genetics? *Am J Med Genet C Semin Med Genet*. 2023;193(3):e32062. doi:10.1002/ajmg.c.32062
39. Chowdhery A, Narang S, Devlin J, et al. PaLM: scaling language modeling with pathways. *J Mach Learn Res*. 2022;24(240):1–113.
40. Gan RK, Ogbodo JC, Wee YZ, Gan AZ, González PA. Performance of google bard and chatgpt in mass casualty incidents triage. *Am J Emerg Med*. 2024;75:72–78. doi:10.1016/j.ajem.2023.10.034
41. Roos J, Kasapovic A, Jansen T, Kaczmarczyk R. Artificial intelligence in medical education: comparative analysis of chatgpt, bing, and medical students in Germany. *JMIR Med Educ*. 2023;9(e46482):e46482. doi:10.2196/46482
42. Raimondi R, Tzoumas N, Salisbury T, et al. Comparative analysis of large language models in the Royal College of Ophthalmologists fellowship exams. *Eye*. 2023;37(17):3530–3533. doi:10.1038/s41433-023-02563-3
43. Farhat F, Chaudry B, Nadeem M, Sohail S, Madsen D *Evaluating AI models for the national pre-medical exam in India: a head-to-head analysis of chatgpt-3.5, gpt-4, and bard (preprint)*.; 2023. doi:10.2196/preprints.51523.
44. Lim ZW, Pushpanathan K, Yew SME, et al. Benchmarking large language models' performances for myopia care: a comparative analysis of chatgpt-3.5, chatgpt-4.0, and google bard. *eBioMedicine*. 2023;95:104770. doi:10.1016/j.ebiom.2023.104770
45. Ali R, Tang OY, Connolly ID, et al. Performance of chatgpt, gpt-4, and google bard on a neurosurgery oral boards preparation question bank. *Neurosurgery*. 2023;93(5):1090–1098. doi:10.1227/neu.0000000000002551
46. Torres-Zegarra BC, Rios-Garcia W, Ñaña-Cordova AM, et al. Performance of chatgpt, bard, claude, and bing on the Peruvian national licensing medical examination: a cross-sectional study. *J Educ Eval Health Prof*. 2023;20:30. doi:10.3352/jeehp.2023.20.30
47. Lee H, Park S. Information amount, accuracy, and relevance of generative artificial intelligence platforms answers regarding learning objectives of medical arthropodology evaluated in English and Korean queries in December 2023: a descriptive study. *J Educ Eval Health Prof*. 2023;20:39. doi:10.3352/jeehp.2023.20.39
48. Toufiq M, Rinchai D, Bettacchioli E, et al. Harnessing large language models (LLMs) for candidate gene prioritization and selection. *J Transl Med*. 2023;21(1):728. doi:10.1186/s12967-023-04576-8
49. Alfredo Clerici C, Chopard S, Levi G. Rare disease in the age of artificial intelligence. *Recenti Prog Med*. 2024;115(2024Febbraio):67–75. doi:10.1701/4197.41839
50. Vitale S, Cotch MF, Sperduto R, Ellwein L. Costs of refractive correction of distance vision impairment in the United States, 1999–2002. *Ophthalmology*. 2006;113(12):2163–2170. doi:10.1016/j.ophtha.2006.06.033
51. 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. *JAMA Intern Med*. 2023;183(6):589. doi:10.1001/jamainternmed.2023.1838
52. Ayers JW, Zhu Z, Poliak A, et al. Evaluating artificial intelligence responses to public health questions. *JAMA Netw Open*. 2023;6(6):e2317517. doi:10.1001/jamanetworkopen.2023.17517
53. Duong D, Solomon BD. Analysis of large-language model versus human performance for genetics questions. *Eur J Hum Genet*. 2023. doi:10.1038/s41431-023-01396-8
54. Martino A, Iannelli M, Truong C, et al. Knowledge injection to counter large language model (LLM) hallucination. In: Pesquita C, Skaf-Molli H, Efthymiou V, editors. *The Semantic Web: ESWC 2023 Satellite Events*. Springer Nature Switzerland; 2023:182–185.
55. McGowan A, Gui Y, Dobbs M, et al. ChatGPT and Bard exhibit spontaneous citation fabrication during psychiatry literature search. *Psychiatry Res*. 2023;326:115334. doi:10.1016/j.psychres.2023.115334

56. Guigue P, Meyer R, Thivolle-Lioux G, Brezinov Y, Levin G. Performance of chatgpt in French language parcours d'accès spécifique santé test and in OBGYN. *Int J Gynecol Obstet*. 2024;164(3):959–963. doi:10.1002/ijgo.15083
57. Xu Z, Jain S, Kankanhalli M. Hallucination is inevitable: an innate limitation of large language models. *ArXiv E-Prints*. 2024. doi:10.48550/arXiv.2401.11817
58. Attawar A, Vora S, Narechania P, Sawant V, Vora H. NLSQL: generating and executing sql queries via natural language using large language models. In: *2023 International Conference on Advanced Computing Technologies and Applications (ICACTA)*.; 2023:1–6. doi:10.1109/ICACTA58201.2023.10392861.
59. Rahsepar AA, Tavakoli N, Kim GHJ, Hassani C, Abtin F, Bedayat A. How AI responds to common lung cancer questions: chatgpt vs google bard. *Radiology*. 2023;307(5):e230922. doi:10.1148/radiol.230922
60. Mago J, Sharma M. The potential usefulness of chatgpt in oral and maxillofacial radiology. *Cureus*. 2023;15(7):e42133. doi:10.7759/cureus.42133
61. Thapa S, ChatGPT AS. Bard, and large language models for biomedical research: opportunities and pitfalls. *Ann Biomed Eng*. 2023;51(12):2647–2651. doi:10.1007/s10439-023-03284-0
62. Vaishya R, Misra A, Vaish A. ChatGPT: is this version good for healthcare and research? *Diabetes Metab Syndr Clin Res Rev*. 2023;17(4):102744. doi:10.1016/j.dsx.2023.102744
63. Wilhelm TI, Roos J, Kaczmarczyk R. Large language models for therapy recommendations across 3 clinical specialties: comparative study. *J Med Internet Res*. 2023;25:e49324. doi:10.2196/49324
64. Gao CA, Howard FM, Markov NS, et al. Comparing scientific abstracts generated by chatgpt to real abstracts with detectors and blinded human reviewers. *Npj Digit Med*. 2023;6(1):75. doi:10.1038/s41746-023-00819-6
65. Thorndike EL. A constant error in psychological ratings. *J Appl Psychol*. 1920;4(1):25–29. doi:10.1037/h0071663
66. Morita PP, Abhari S, Kaur J, Lotto M, PADSES M, Oetomo A. Applying chatgpt in public health: a SWOT and PESTLE analysis. *Front Public Health*. 2023;11:1225861. doi:10.3389/fpubh.2023.1225861
67. Cheng K, Li Z, He Y, et al. Potential use of artificial intelligence in infectious disease: take chatgpt as an example. *Ann Biomed Eng*. 2023;51(6):1130–1135. doi:10.1007/s10439-023-03203-3
68. Pandey S, Sharma S. A comparative study of retrieval-based and generative-based chatbots using deep learning and machine learning. *Healthc Anal*. 2023;3:100198. doi:10.1016/j.health.2023.100198
69. Beerbaum D *Generative artificial intelligence (GAI) with chat gpt for accounting -a business case*; 2023.
70. Budhwar P, Chowdhury S, Wood G, et al. Human resource management in the age of generative artificial intelligence: perspectives and research directions on chatgpt. *Hum Resour Manag J*. 2023;33(3):606–659. doi:10.1111/1748-8583.12524
71. McGee R. What are the top 20 questions in sociology? *A ChatGPT Reply*. 2023. doi:10.13140/RG.2.2.36401.04963
72. Kovačević D Use of chatgpt in esp teaching process. In: *2023 22nd International Symposium INFOTEH-JAHORINA (INFOTEH)*.; 2023:1–5. doi:10.1109/INFOTEH57020.2023.10094133.
73. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. 2023:620(7972):172–180. doi:10.1038/s41586-023-06291-2
74. Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial intelligence, bias and clinical safety. *BMJ Qual Saf*. 2019;28(3):231–237. doi:10.1136/bmjqs-2018-008370
75. Kahambing JG. ChatGPT, public health communication and 'intelligent patient companionship'. *J Public Health*. 2023;45(3):e590–e590. doi:10.1093/pubmed/fdad028
76. Tiwari A, Kumar A, Jain S, et al. Implications of chatgpt in public health dentistry: a systematic review. *Cureus*. 2023;15(6):e40367. doi:10.7759/cureus.40367