BMC Oral Health

RESEARCH

Open Access

Performance of artificial intelligence chatbots in responding to the frequently asked questions of patients regarding dental prostheses

Hossein Esmailpour[1], Vanya Rasaie[2], Yasamin Babaee Hemmati[3] and Mehran Falahchai[4*]

Abstract

Background Artificial intelligence (AI) chatbots are increasingly used in healthcare to address patient questions by providing personalized responses. Evaluating their performance is essential to ensure their reliability. This study aimed to assess the performance of three AI chatbots in responding to the frequently asked questions (FAQs) of patients regarding dental prostheses.

Methods Thirty-one frequently asked questions (FAQs) were collected from accredited organizations' websites and the "People Also Ask" feature of Google, focusing on removable and fixed prosthodontics. Two board-certified prosthodontists evaluated response quality using the modified Global Quality Score (GQS) on a 5-point Likert scale. Inter-examiner agreement was assessed using weighted kappa. Readability was measured using the Flesch-Kincaid Grade Level (FKGL) and Flesch Reading Ease (FRE) indices. Statistical analyses were performed using repeated measures ANOVA and Friedman test, with Bonferroni correction for pairwise comparisons ($\alpha = 0.05$).

Results The inter-examiner agreement was good. Among the chatbots, Google Gemini had the highest quality score (4.58 ± 0.50), significantly outperforming Microsoft Copilot (3.87 ± 0.89) ($P = .004$). Readability analysis showed ChatGPT (10.45 ± 1.26) produced significantly more complex responses compared to Gemini (7.82 ± 1.19) and Copilot (8.38 ± 1.59) ($P < .001$). FRE scores indicated that ChatGPT's responses were categorized as fairly difficult (53.05 ± 7.16), while Gemini's responses were in plain English (64.94 ± 7.29), with a significant difference between them ($P < .001$).

Conclusions AI chatbots show great potential in answering patient inquiries about dental prostheses. However, improvements are needed to enhance their effectiveness as patient education tools.

Keywords Artificial intelligence, Prosthodontics, Patient education as topic, Health literacy, Natural Language processing

*Correspondence:
Mehran Falahchai
Mehran.falahchai@gmail.com
[1]School of Dentistry, Guilan University of Medical Sciences, Rasht, Iran
[2]Research Affiliate at Sydney Dental School, Faculty of Medicine and Health, Sydney, Australia
[3]Department of Orthodontics, Dental Sciences Research Center, School of Dentistry, Guilan University of Medical Sciences, Rasht, Iran
[4]Department of Prosthodontics, Dental Sciences Research Center, School of Dentistry, Guilan University of Medical Sciences, Rasht, Iran

Background

Artificial intelligence (AI) chatbots are complex computerized models trained with a large corpus of textual data. They are capable of mimicking human language patterns and making meaningful conversations by benefitting from deep learning algorithms [1]. Various applications have been proposed for chatbots in the healthcare sector due to their ability to provide personalized responses to patient questions [2]. Assisting clinical decision-making, enhancing scientific research processes, and responding to patient questions are among the applications of chatbots [3]. Nonetheless, AI chatbots have several limitations, such as the possibility of giving wrong or misleading answers due to insufficient training data [4]. This is particularly important in the healthcare sector because such responses could be potentially harmful. Therefore, precise and comprehensive assessment of the performance of AI chatbots is imperative to properly benefit from their capabilities in this field [5].

Different companies have designed and introduced several chatbots in recent years. ChatGPT, Google Gemini, and Microsoft Copilot are the three pioneer chatbots in this field [6]. ChatGPT is the AI chatbot of the OpenAI company that has attracted the attention of millions of people worldwide [7]. This popular chatbot uses a transformer model known as the generative pre-trained transformer and has a unique capability to respond to questions in the form of engaging and attractive conversations through supervised learning methods such as reinforcement learning from human feedback [8, 9]. Google Gemini is another chatbot with a unique capability to provide accurate and up-to-date responses by relying on the vast database and powerful search engine of Google [10]. Microsoft Copilot, previously known as Bing, also has numerous applications in treatment documentation processes and providing patients with practical advice by benefiting from GPT-4 and simultaneous access to the Internet [11]. With their evolving capabilities, these chatbots have the potential to enhance patient education by offering accessible and reliable health information.

Patient education is one of the most important components of each medical intervention [12]. Evidence shows that education and knowledge enhancement of patients increase their cooperation, self-care behaviors, and long-term satisfaction with the treatment [13, 14]. However, studies show a lack of patient knowledge regarding dental procedures. A study revealed that edentulous patients had limited knowledge about their prosthetic hygiene [15]. Similarly, inadequate patient knowledge has been reported in other areas of dentistry, such as dental implants and fixed orthodontics [16, 17]. These findings highlight the need for reliable information sources to address patient inquiries effectively.

Given their advanced capabilities, AI chatbots have the potential to help mitigate these shortcomings by providing users with personalized responses and can serve as a viable alternative to traditional methods of obtaining dental health information for patients [18]. Notably, their success in passing professional examinations, such as the European Certification in Implant Dentistry and comprehensive licensing exams in the UK and US, suggests that they might be capable of answering patient inquiries similarly to human dentists [19, 20].

According to the Health Information National Trends Survey, the Internet is the main source of medical information for approximately 80% of the U.S. population [21]. Nonetheless, despite the availability of Internet search engines (like Google), they have information with a highly variable quality level, which makes it difficult to assess the reliability of such sources of information [22]. Ayers et al. [23] found that ChatGPT responds to online patient questions with a higher level of empathy than physicians. Also, another study showed the superior performance of ChatGPT compared to the Google search engine in responding to patient questions regarding symptom-based diagnoses [24]. However, several limitations remain. AI chatbots have also failed certain examinations, indicating inconsistencies in their knowledge base [20, 25]. Concerns persist regarding bias in responses, ethical challenges in their application, and the reliability of the information they provide [5]. Therefore, the assessment of their accuracy and performance in different healthcare fields is imperative.

Several recent studies assessed the performance of AI chatbots as a source of patient information [18, 26]. For instance, a previous study reported that Gemini and Copilot responded to patient questions regarding chest X-rays with better readability but lower accuracy than ChatGPT [27]. Moreover, dental researchers in different fields, such as endodontics, periodontics, orthodontics, and oral and maxillofacial surgery, addressed this topic and unanimously reported that chatbots are efficient in responding to patient questions [28–31].

Prosthodontics is a specialized field of dentistry that includes a wide range of dental procedures that may generate some questions for patients [32]. Freire et al. [33] evaluated the accuracy and reproducibility of ChatGPT in responding to technical and professional prosthodontic questions and reported its poor performance. They highlighted the need for further evaluation of the performance of chatbots in this field, especially as a source of patient medical information.

Several studies have addressed the performance of chatbots in the healthcare field; [34–37] however, to the best of the authors' knowledge, no previous study has evaluated the readability and quality of responses of chatbots as a source of information for patient questions in

the field of prosthodontics. Therefore, the purpose of this study was to assess the performance of AI chatbots in responding to frequently asked questions (FAQs) of patients in the field of prosthodontics. The null hypothesis of the study was that the accuracy and readability of the three chatbots evaluated in this study would not be significantly different.

Materials and methods

The study protocol was approved by the ethics committee of Guilan University of Medical Sciences (IR.GUMS.REC.1403.450). In this analytical cross-sectional study, FAQs of patients in two fields of removable and fixed prosthodontics were collected to assess the performance of AI chatbots. To ensure the representativeness of the selected FAQs, a systematic, multi-source approach was employed, incorporating institutional sources, public interest data, and organically generated patient inquiries.

Institutional sources FAQs were collected from the official websites of accredited prosthodontic organizations, including the American College of Prosthodontics and Washington State Prosthodontics. These institutions provide curated, evidence-based information, ensuring that commonly encountered patient concerns were included in the dataset. The information from these organizations was freely accessible to the public on their official websites, and no special permissions were required for data extraction.

Public interest data (google trends) To capture real-world patient information-seeking behavior, frequently searched phrases related to removable and fixed prosthodontics were extracted from Google Trends (https://trends.google.com). Irrelevant items (e.g., textbook names, cost-related queries) were omitted to focus on clinically relevant inquiries (Table 1).

Google's "people also ask" feature To incorporate organic patient-generated queries, the extracted Google Trends terms were used as inputs in Google Search to retrieve questions listed in the "People Also Ask" section. This approach ensured the inclusion of non-specialist phrasing that reflects how patients naturally ask questions online.

Table 1 Final list of phrases entered in Google search to collect questions from the "people also ask" section after removing unrelated items (2004–2024)

Removable Prosthodontics	Fixed Prosthodontics
Removable Denture	Crown
Denture	Removable Prosthodontics
Removable Partial Denture	Fixed bridge
Partial Denture	Fixed Partial Denture
Fixed Prosthodontics	Fixed Prosthesis

The formula for the comparison of mean values in three groups according to the performance curve and the Flesch Kincaid Grade Level (FKGL) variable were used to calculate the sample size of the study. The minimum sample size was calculated to be 19 questions in each group considering the mean values of 14.3, 12.5, and 12.9 in the three groups, study power of 0.81, type I error of 0.05, and standard deviation of 1.8 [38].

The initial dataset of 65 questions underwent expert validation to enhance clinical relevance and eliminate redundancy. Two board-certified prosthodontists independently reviewed all collected questions and refined the list. Questions that directly addressed common patient concerns regarding post-treatment care, prognosis, terminology, complications, and procedural details were retained, while duplicates, overly technical inquiries, and those lacking sufficient clinical relevance were excluded. Through a consensus-based process, 31 key FAQs were finalized to ensure a balanced representation of patient concerns across different aspects of prosthodontic treatment (Table 2).

Three free commonly used chatbots namely ChatGPT 3.5, Microsoft Copilot, and Google Gemini were evaluated in the present study. To maximize simulation of the clinical setting, the most commonly used Application Programming Interface of each chatbot was used for response collection:

- ChatGPT: Accessed through its official website at https://chat.openai.com.
- Microsoft Copilot: Accessed through its official website at https://copilot.microsoft.com in conversation mode on the "More Balanced" style.
- Google Gemini: Accessed through its official website at https://gemini.google.com.

To prevent memory retention bias, the collected questions were given to the chatbots in a separate new chat page. Also, all questions were asked from all three chatbots with no prior prompt and on the same day, and their first responses were collected. Finally, the examiners were asked to rate the responses. The Global Quality Scale (GQS) is a reliable tool for evaluation of the quality of online information sources for patients, which has been designed according to a 5-point Likert scale [39]. Two Board-certified prosthodontists independently assessed the quality of the responses using the modified-GQS (Table 3) [28]. Accordingly, the responses were evaluated in terms of correctness, accuracy, and completeness to rate their overall quality.

The examiners were blinded to the type of chatbot generating the responses, and the responses were provided

Table 2 Final list of faqs collected on removable and fixed prosthodontics. The questions were provided to the AI chatbots in this exact format and order

	Frequently Asked Questions
Removable Prosthodontics	
1	How do I care for my dentures?
2	Is it normal to have sore spots after wearing dentures?
3	Can I sleep in my dentures?
4	Can I eat and speak normally with dentures?
5	Is it possible to have dentures put in the same day as teeth removal?
6	As a new denture wearer, the bottom denture seems loose. What should I do?
7	Is it possible to perfectly color-match my partial denture to my remaining natural teeth?
8	What do I do if my dentures feel very heavy and it is difficult to close my mouth?
9	My dentures will not stay in no matter what kind of adhesive I use. What should I do?
10	Is it possible to get dentures after not having teeth for a long time?
11	What should I do about bad breath with dentures?
12	Will Dentures change my face?
13	I'm having problem tasting food with my dentures. Is this normal?
14	Is it possible to change or add teeth to my denture?
15	I can't stop gagging on my denture. What's the problem?
Fixed Prosthodontics	
1	What is the difference between crown and Bridge?
2	Why does my tooth look bulky after placing crown
3	Why is my tooth sensitive after crown placement?
4	Why do the crowns fall off?
5	Does my crown stain?
6	What happens if my crown chip off?
7	How long will a dental crown last?
8	What are the disadvantages of crowns?
9	Are crowns stronger than real teeth?
10	Does a crown require a root canal?
11	Is a crown safer than a filling?
12	What is a dental bridge?
13	Is any special care necessary for crown/dental bridges?
14	How long after crown/bridge can I eat?
15	How long should a bridge last in your mouth?
16	Can food get under bridges?

to them without the source name. The weighted kappa was calculated to assess the inter-examiner agreement. According to the Altman's classification [40], kappa values < 0.20, and between 0.21 and 0.40, 0.41–0.60, 0.61–0.80, and 0.81-1.00 indicate poor, fair, moderate, good, and very good agreement, respectively. Finally, disagreements between the assessors were resolved by evidence-based discussion, and one final score was allocated to each response for subsequent statistical analysis. The selection of two evaluators was based on prior studies assessing AI-generated medical and dental content [28, 41, 42].

Readability was assessed by using two reliable tools commonly used for this purpose in the literature, namely the Flesch Reading Ease (FRE) and the FKGL [43] (Table 3). In the FRE, the scores may range from 0 to 100, and lower scores indicate higher difficulty of the text in terms of readability [44]. For example, scores 0–10 indicate highly difficult readability, only comprehendible by the American university graduates; while, scores 90–100 indicate very easy readability comprehendible even by an elementary schooler. The FKGL is another tool for the assessment of the readability of the texts. The FKGL scores may range from 0 to 18, and each score indicates the number of years of education required to understand and comprehend the text [45]. According to the American Medical Association and the National Institute of Health, patient educational content and resources should be readable and comprehensible by individuals with as low as 6 years of education (6th graders). Therefore, scores > 80 in the FRE and < 7 in the FKGL were considered the ideal level of readability for the responses generated by the chatbots [46]. All readability calculations were performed by a well-known online tool available for this purpose at Readable.com (Added Bytes Ltd., Brighton, England).

The Kolmogorov-Smirnov test was applied to test the normality assumption of the data, while the Levene test was used to analyze the homogeneity of the variances. Accordingly, data were analyzed using repeated measures of ANOVA and Friedman test with pairwise comparisons with the Bonferroni test. All statistical analyses were carried out using SPSS version 26 (IBM Corp., NY, USA) at 0.05 level of significance.

Results

As mentioned earlier, for quality assessment, two experienced prosthodontists evaluated the quality of the responses of ChatGPT, Google Gemini, and Microsoft Copilot AI chatbots using a 5-point Likert scale. The weighted kappa values are presented in Table 4. According to the Altman's classification, the inter-examiner agreement was good. Google Gemini acquired the highest (4.58 ± 0.50), and Microsoft Copilot acquired the lowest (3.87 ± 0.89) mean quality score. Also, examiners only gave the lowest score (score 1) to one response, which was the response of Copilot to the question "is any special care necessary for dental crown?".

The Friedman test found a significant difference among the chatbots regarding the quality of their responses ($P < .001$). Subsequent pairwise comparisons of the chatbots regarding the quality of their responses to FAQs revealed the significant superiority of Gemini to Copilot ($P = .004$). Nonetheless, the differences between

Table 3 Summary of the tools used to evaluate the performance of AI chatbots: *: FKGL: Flesch Kincaid grade level, FRE: Flesch reading ease

Accuracy and Completeness	
Modified Global Quality Score (mGQS)	**Score**
Strongly disagree: The answer and the entire content are incorrect or irrelevant.	1
Disagree: The answer is incorrect, but the content includes some correct elements.	2
Neutral: The answer is somewhat correct, but details are primarily incorrect, missing, or irrelevant.	3
Agree: The answer is correct and most of the content is correct, but it lacks information, or contains incorrect information.	4
Strongly agree: The answer is correct, and the content is comprehensive.	5
Readability	
Formula	Index
0.39 ((Total number of words)/(Total number of sentences)) + 11.8((Total number of syllabes)/(Total number of words))-15.59	*FKGL
206.835–1.015 ((Total number of words)/(Total number of sentences))-84.6((total number of syllables)/(total number of words))	*FRE

Table 4 Performance metrics of AI chatbots in responding to prosthodontic frequently asked questions ($n = 31$)

Chatbot	Quality score (Mean ± SD)	weighted kappa	FKGL (Mean ± SD)	FRE (Mean ± SD)
ChatGPT (GPT-3.5)	4.13 ± 0.56	0.78	10.45 ± 1.26	53.05 ± 7.16
Google Gemini (Gemini 1.5)	4.58 ± 0.50	0.76	7.82 ± 1.19	64.94 ± 7.29
Microsoft Copilot (Incorporating GPT-4)	3.87 ± 0.89	0.72	8.38 ± 1.59	59.35 ± 7.42
P value	<.001[a]	-	<.001[b]	<.001[b]
Effect Size	0.21[c]	-	0.52[d]	0.46[d]

Quality Score was evaluated on a 5-point Likert scale, Weighted kappa represents inter-examiner agreement, FKGL: Flesch Kincaid Grade Level, FRE: Flesch Reading Ease, Higher scores of FKGL and Lower Scores of FRE indicate easier readability, SD: Standard Deviation

[a] Statistically significant difference in the mean quality scores of three chatbots using the Friedman test, [b] Statistically significant difference in the mean FKGL and FRE scores of chatbots using repeated measures ANOVA, [c] Eta Squared, [d] Partial Eta Squared

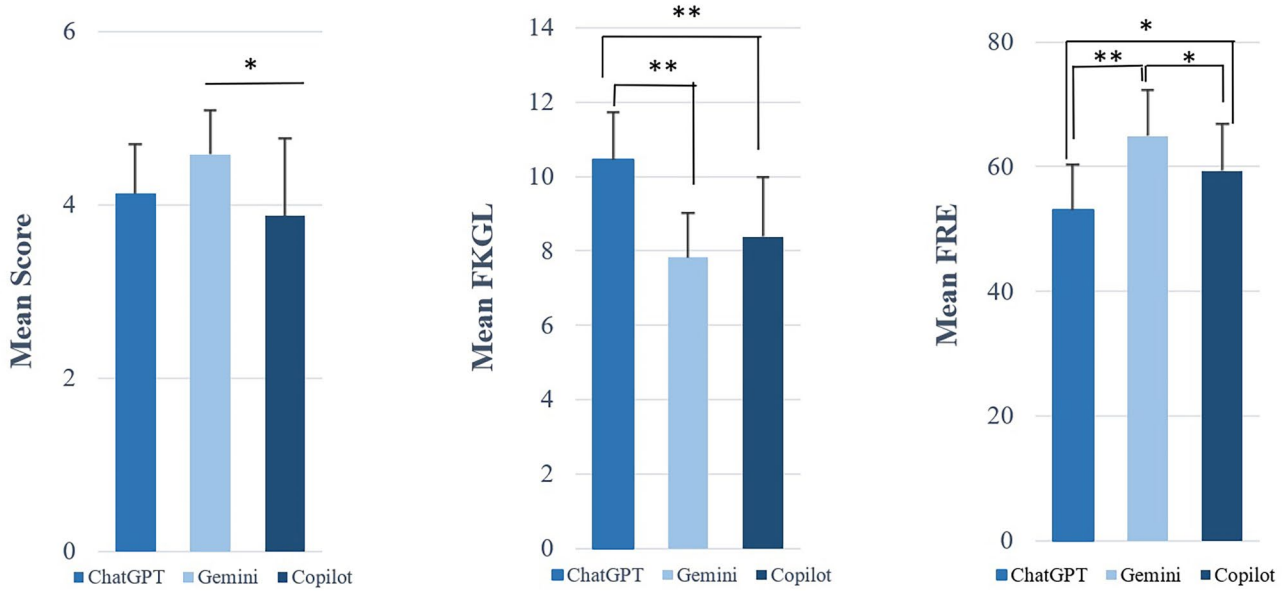

Fig. 1 Comparison of the mean quality, Flesch Kincaid Grade Level (FKGL), and Flesch Reading Ease (FRE) scores of the three AI chatbots. **: $P < 0.001$, *: $P < 0.05$

ChatGPT and Copilot ($P > .000$), and ChatGPT and Gemini ($P = .067$) were not statistically significant (Fig. 1).

Assessment of readability by the FKGL, which provides an objective estimate of the understandability and readability of a text according to educational years, revealed that ChatGPT acquired the highest and Gemini acquired the lowest mean score ($P < .05$). Subsequent pairwise comparisons revealed that ChatGPT had significant differences with both Google Gemini and Microsoft Copilot ($P < .001$ for both), but the difference between Gemini and Copilot was not significant ($P = .478$, Fig. 1).

The FRE used for the assessment of the educational level required for understanding a text revealed that ChatGPT acquired the lowest mean score, and its

responses were significantly more difficult to understand compared with the responses of Gemini ($P<.001$) and Copilot ($P=.002$). Gemini acquired the highest score in this regard, and the difference between Gemini and Copilot was statistically significant ($P=.022$, Fig. 1).

Discussion

The present results revealed a significant difference among the three tested chatbots in terms of the quality and readability of their responses, and therefore, the null hypothesis of the study was rejected. The AI chatbots are unique in providing personalized responses and making conversations and, therefore, can serve as a turning point for patients' access to medical information. However, due to their limitations, it is imperative to assess their performance, especially in the healthcare field. For example, in the present study, all three chatbots acquired a score of 2 (i.e., poor-quality response) in responding to a question. Giving incorrect or irrelevant information, given that it is believable (AI hallucination), is a major limitation of chatbots, which can be potentially harmful. In the present study, reliable tools, including the GQS, FKGL, and FRE, were used to assess the performance of leading AI chatbots in responding to FAQs of patients regarding dental prostheses. The obtained results can aid in the safer and more reliable use of chatbots in responding to patient questions regarding different dental prosthetic treatments.

A comparison of the three chatbots in terms of the quality of the responses revealed that Google Gemini had a significantly superior performance than the other two chatbots. A previous study on chronic kidney diseases also revealed the superiority of Gemini in this respect [47]. Nonetheless, the literature is controversial in this regard, and some others found no significant difference in the accuracy or quality of the responses of different chatbots [48, 49]. Another study showed the superior performance of ChatGPT in this regard [50]. The high variation in tools used for quality assessment of responses may be one reason for this controversy. Developing a reliable tool for the exclusive assessment of the performance of AI chatbots would be useful in solving such ambiguities. Significant differences in training data may be another reason for variations in the reported results. The performance of chatbots is, in fact, a reflection of their training data. Therefore, variations in the training data can lead to differences in the results.

In the current study, the majority of chatbot responses to FAQs of patients acquired excellent scores (4–5). It should be noted that, unlike the present study, ChatGPT had a poor performance in responding to professional and clinical questions regarding dental prostheses in a study by Freire et al. [33] Therefore, it appears that chatbots are more capable of responding to general questions of patients, pointing to their high potential as a patient education tool. Consistent with the present results, other studies in other dental fields, such as orthodontics and periodontics, reported high accuracy and quality of AI chatbot responses to FAQs of patients [34–37]. Nonetheless, the provision of healthcare information is a highly sensitive task, since incorrect and misleading information can be harmful. Therefore, despite their effectiveness, their precise and systematic quality assessment is imperative.

In the assessment of the readability of the responses, Google Gemini was superior to other chatbots, according to both the FKGL and FRE in the present study. In other words, Google Gemini not only showed a superior performance but also gave responses with higher readability. Studies on this topic in different fields have reported conflicting results. Behers et al. [51] reported that ChatGPT acquired the lowest readability score regarding cardiac catheterization, and the responses of other chatbots were significantly more readable. However, another study on appendicitis found no significant difference in readability between ChatGPT and Gemini [48]. This difference may be due to differences in the training data of chatbots in different fields of specialty. Further studies are warranted in this regard to resolve this controversy.

The overall readability level of the responses of the chatbots to FAQs regarding prosthodontics was unfavorable in the current study. Figure 2 shows the frequency distribution of the readability scores of the chatbot responses. Almost all responses had an unfavorable readability level according to the suggested level by the National Institute of Health [46]. Moreover, considering the FRE scores of the chatbots, the readability of ChatGPT and Copilot was at the 10th to 12th grade level, while the readability of Gemini was at the 8th to 9th grade level. Further, Alshehri et al. [52] reported difficult readability scores and poor quality of information available on the Internet regarding dental prostheses. Unlike their findings, the quality of information of chatbots was found to be high in the present study; however, difficult readability can prevent a wide range of patients from benefitting from such information. Therefore, attempts should be made to improve the readability level of the chatbot responses in their next versions.

In this study, two well-established readability formulas were used to provide an objective assessment of the text difficulty of the AI chatbot Responses. Previous studies have also utilized FKGL and FRE in different fields. In appendicitis, Ghanem et al. [53] reported difficult FKGL and FRE scores of AI responses ranging from a high school student to a college graduate level. Similarly, Onder et al. [41] found that AI responses to pregnancy-related hyperthyroidism required college-level education for an adequate understanding. Further,

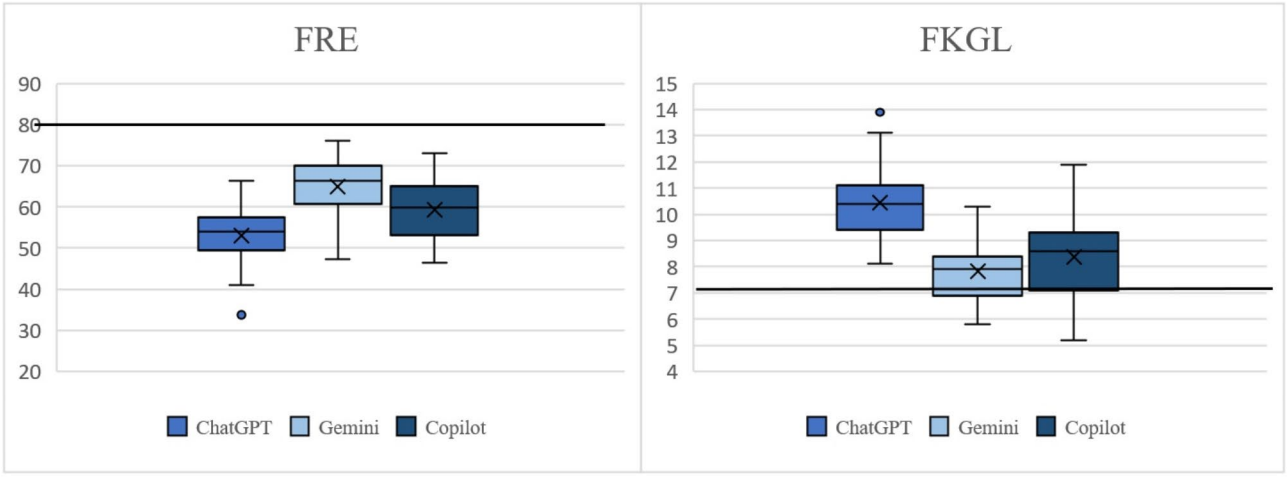

Fig. 2 Distribution of Flesch Kincaid Grade Level (FKGL) and Flesch Reading Ease (FRE) scores of AI chatbot responses to FAQs: The horizontal dashed line indicates the desirable readability level according to the National Library of Health

another study used FKGL and FRE to assess the readability of the responses of chatbots in orthodontics. Unfavorable difficulty was similarly reported in the results of this study [30]. These findings suggest that the readability challenges of AI-generated content are not confined to prosthodontics alone. Measures should be taken to address this gap in the performance of AI chatbots as a patient education tool.

While readability metrics provide valuable insights into the complexity of chatbot-generated responses, they do not directly measure patient comprehension. Studies show that an individual's health literacy could impact their understanding of patient education materials. Gieg et al. [54] showed in their study that patients with low health literacy may struggle to comprehend educational materials written at higher readability levels, limiting their ability to make informed health decisions. Moreover, Low health literacy has been associated with poorer comprehension of health information and subsequent negative health outcomes [55]. Although readability formulas provide an objective, standardized assessment of text difficulty, patient comprehension is multifaceted and extends beyond readability alone. Individuals with lower health literacy may struggle to interpret even moderately readable content. This underscores the need for further research on comprehension-based evaluation. Future studies should explore how AI chatbots can adapt their responses dynamically to suit different literacy levels.

Although this study employed a controlled, unbiased approach by assessing the chatbots' responses with no prior prompt, real-world patient interactions are often more dynamic. The performance of chatbots highly depends on the type of prompt. Patients may engage in multi-turn conversations, where previous exchanges influence subsequent responses [56]. Additionally, chatbots with contextual memory may adapt their answers based on earlier prompts within the same session. Weight et al. [57] reported that when they asked ChatGPT for further explanation, the FKGL score significantly increased, and the readability of the responses significantly improved with no change in accuracy. Nonetheless, techniques to receive better feedback from AI models, known as prompt engineering, need to be taught, although public education in this respect appears to be highly difficult. Therefore, developing chatbots for patient education and health literacy promotion can help eliminate the shortcomings and limitations of public chatbots in this respect. Future research should consider studying chatbot responses in simulated real-world interactions, including multi-turn dialogues and context-aware responses, to further explore their performance in patient education and healthcare applications.

This study is among the first to comprehensively evaluate both the readability and quality of chatbot responses in prosthodontics. By assessing three leading AI chatbots and utilizing well-established metrics, we provide valuable insights into their applicability for patient education. However, we acknowledge some limitations. In the present study, only two examiners evaluated the quality of the responses; benefitting from the opinion of a higher number of experts from different academic institutes and universities can decrease bias and aid in achieving a consensus regarding the performance of AI chatbots.

The modified GQS was used for quality assessment in this study; although it is a well-known, reliable tool for this purpose, it has not been exclusively designed for the assessment of the performance of AI models. Developing a suitable tool for this particular purpose would greatly help in a more precise assessment of the performance of AI chatbots, enabling better comparison of the results.

While Google Trends and Google Search provide valuable insights into public interest, they may not fully

capture clinical inquiries made by patients in dental settings, where spoken communication differs from online search behavior. Additionally, the expert validation process of the FAQs, although crucial for ensuring clinical relevance, inherently involves a degree of subjectivity that could introduce selection bias. Future studies could enhance the representativeness of FAQs by incorporating direct patient surveys, conducting interviews in clinical settings, and expanding the expert panel to include specialists from diverse practice backgrounds.

Conclusion

Comparison of the three pioneer AI chatbots revealed the superiority of Google Gemini in terms of both quality and readability. All three chatbots had acceptable quality despite giving occasional irrelevant answers. However, their responses had difficult readability. Conversational chatbots with their unique capability in mimicking human language can revolutionize public access to healthcare information. Despite some limitations, these chatbots have high potential for replacement or improvement of traditional methods of responding to patient questions. Similar future studies can provide more evidence and pave the way for more reliable application of AI chatbots for patient education.

Abbreviations

AI	Artificial intelligence
FAQs	Frequently asked questions
GQS	Global quality score
FKGL	Flesch kincaid grade level
FRE	Flesch reading ease

Acknowledgements

None.

Author contributions

HE: Conceptualization, Methodology, Writing - Original Draft VR: Methodology, Data Curation YBH: Writing - Review & Editing, Investigation, Visualization MF: Funding acquisition, Project administration, Supervision.

Funding

This work was funded and supported by dental sciences research center of Guilan University of medical sciences (#IR.GUMS.REC.1403.450).

Data availability

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request. Also, the datasets supporting the conclusions of this article are included within the article.

Declarations

Ethics approval and consent to participate

All experimental protocols were approved by Research Ethics Committees of Guilan university of medical sciences. All methods were carried out in accordance to the ethical principles and the national norms and standards for conducting Medical Research in Iran. Ethics code is IR.GUMS.REC.1403.450.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 25 February 2025 / Accepted: 7 April 2025
Published online: 15 April 2025

References

1. Zhang P, Kamel Boulos MN. Generative AI in medicine and healthcare: promises, opportunities and challenges. Future Internet. 2023;15(9):286.
2. Meskó B, Topol EJ. The imperative for regulatory oversight of large Language models (or generative AI) in healthcare. NPJ Digit Med. 2023;6(1):120.
3. Sallam M, Basel. Switzerland) Mar. 2023;19(6). https://doi.org/10.3390/healthcare11060887
4. Reddy S. Evaluating large Language models for use in healthcare: A framework for translational value assessment. Inf Med Unlocked. 2023;41:101304. https://doi.org/10.1016/j.imu.2023.101304
5. Wang Y, Zhao Y, Petzold L. Are large Language models ready for healthcare? A comparative study on clinical Language Understanding. PMLR 2023;804–23.
6. Alhur A, Redefining Healthcare With Artificial Intelligence (AI). The contributions of ChatGPT, Gemini, and Co-pilot. Cureus Apr. 2024;16(4):e57795. https://doi.org/10.7759/cureus.57795
7. Ray PP. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet Things Cyber-Physical Syst. 2023;3:121–54.
8. Roumeliotis KI, Tselikas ND. Chatgpt and open-ai models: A preliminary review. Future Internet. 2023;15(6):192.
9. Goodman RS, Patrinely JR, Stone CA, et al. Accuracy and reliability of chatbot responses to physician questions. JAMA Netw Open. 2023;6(10):e2336483–2336483.
10. Rane N, Choudhary S, Rane J. Gemini versus ChatGPT: applications, performance, architecture, capabilities, and implementation. Performance, Architecture, Capabilities, and Implementation 2024;2024.
11. Tepe M, Emekli E. Decoding medical jargon: the use of AI Language models (ChatGPT-4, BARD, Microsoft copilot) in radiology reports. Patient Educ Couns Sep. 2024;126:108307. https://doi.org/10.1016/j.pec.2024.108307
12. Albano MG, d'Ivernois JF, de Andrade V, Levy G. Patient education in dental medicine: A review of the literature. Eur J Dent Education: Official J Association Dent Educ Europe May. 2019;23(2):110–8. https://doi.org/10.1111/eje.12409
13. Çevik L, Rizalar S. The effect on anxiety and satisfaction of Video-Assisted education given before an ERCP procedure. gastroenterology nursing: the official journal of the society of gastroenterology nurses and associates. Jan-Feb. 2024;01(1):19–26. https://doi.org/10.1097/sga.0000000000000781
14. Thorat V, Rao P, Joshi N, Talreja P, Shetty AR. Role of artificial intelligence (AI) in patient education and communication in dentistry. Cureus May. 2024;16(5):e59799. https://doi.org/10.7759/cureus.59799
15. de Castellucci Barbosa L, Ferreira MR, de Carvalho Calabrich CF, Viana AC, de Lemos MC, Lauria RA. Edentulous patients' knowledge of dental hygiene and care of prostheses. Gerodontology Jun. 2008;25(2):99–106. https://doi.org/10.1111/j.1741-2358.2007.00190.x
16. Alhaija ESA, Al-Saif EM, Taani DQ. Periodontal health knowledge and awareness among subjects with fixed orthodontic appliance. Dental press journal of orthodontics. 2018;23(5):40e. 1-40.e9.
17. Dolińska E, Węglarz A, Jaroma W, Kornowska G, Zapaśnik Z, Włodarczyk P, Wawryniuk J, Pietruska M. Periodontal patients' perceptions and knowledge of dental implants—A questionnaire study. J Clin Med. 2024;13(16):4859. https://doi.org/10.3390/jcm13164859
18. Choi J, Kim JW, Lee YS, et al. Availability of ChatGPT to provide medical information for patients with kidney cancer. Sci Rep Jan. 2024;17(1):1542. https://doi.org/10.1038/s41598-024-51531-8
19. Revilla-León M, Barmak BA, Sailer I, Kois JC, Att W. Performance of an artificial Intelligence-Based chatbot (ChatGPT) answering the European certification in implant dentistry exam. Int J Prosthodont Apr. 2024;22(2):221–4. https://doi.org/10.11607/ijp.8852
20. Jung YS, Chae YK, Kim MS, Lee H-S, Choi SC, Nam OH. Evaluating the accuracy of artificial Intelligence-Based chatbots on pediatric dentistry questions in the Korean National dental board exam. J Korean Acad Pediatr Dentistry. 2024;51(3):299–309.
21. Calixte R, Rivera A, Oridota O, Beauchamp W, Camacho-Rivera M. Social and demographic patterns of health-related internet use among adults in the United States: a secondary data analysis of the health information national trends survey. Int J Environ Res Public Health. 2020;19(18). https://doi.org/10.3390/ijerph17186856

22. McClung HJ, Murray RD, Heitlinger LA. The internet as a source for current patient information. Pediatr Jun. 1998;101(6):E2. https://doi.org/10.1542/peds.101.6.e2
23. 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):589–96.
24. Shen SA, Perez-Heydrich CA, Xie DX, Nellis JC. ChatGPT vs. web search for patient questions: what does ChatGPT do better? European archives of Oto-Rhino-Laryngology. 2024;281(6):3219–25.
25. Danesh A, Pazouki H, Danesh K, Danesh F, Danesh A. The performance of artificial intelligence Language models in board-style dental knowledge assessment: A preliminary study on ChatGPT. J Am Dent Assoc. 1939;154(11):970–4. https://doi.org/10.1016/j.adaj.2023.07.016. Nov 2023.
26. Walker HL, Ghani S, Kuemmerli C, et al. Reliability of medical information provided by ChatGPT: assessment against clinical guidelines and patient information quality instrument. J Med Internet Res Jun. 2023;30:25:e47479. https://doi.org/10.2196/47479
27. Tepe M, Emekli E. Assessing the responses of large Language models (ChatGPT-4, Gemini, and Microsoft Copilot) to frequently asked questions in breast imaging: A study on readability and accuracy. Cureus May. 2024;16(5):e59960. https://doi.org/10.7759/cureus.59960
28. Mohammad-Rahimi H, Ourang SA, Pourhoseingholi MA, Dianat O, Dummer PMH, Nosrat A. Validity and reliability of artificial intelligence chatbots as public sources of information on endodontics. Int Endod J. 2024;57(3):305–14.
29. Alan R, Alan BM. Utilizing ChatGPT-4 for providing information on periodontal disease to patients: A DISCERN quality analysis. Cureus Sep. 2023;15(9):e46213. https://doi.org/10.7759/cureus.46213
30. Kılınç DD, Mansız D. Examination of the reliability and readability of chatbot generative pretrained Transformer's (ChatGPT) responses to questions about orthodontics and the evolution of these responses in an updated version. Am J Orthod Dentofac Orthop. 2024;165(5):546–55.
31. Balel Y. Can ChatGPT be used in oral and maxillofacial surgery? J Stomatology Oral Maxillofacial Surg. 2023;124(5):101471.
32. Sondell K, Söderfeldt B, Palmqvist S. Dentist-patient communication and patient satisfaction in prosthetic dentistry. Int J Prosthodont Jan-Feb. 2002;15(1):28–37.
33. Freire Y, Laorden AS, Pérez JO, Sánchez MG, García VD-F, Suárez A. ChatGPT performance in prosthodontics: assessment of accuracy and repeatability in answer generation. J Prosthet Dent. 2024;131(4):659. e1-659. e6.
34. Babayiğit O, Tastan Eroglu Z, Ozkan Sen D, Ucan Yarkac F. Potential use of ChatGPT for patient information in periodontology: A descriptive pilot study. Cureus Nov. 2023;15(11):e48518. https://doi.org/10.7759/cureus.48518
35. Aguiar de Sousa R, Costa SM, Almeida Figueiredo PH, Camargos CR, Ribeiro BC, Alves ESMRM. Is ChatGPT a reliable source of scientific information regarding third-molar surgery? Journal of the American Dental Association (1939). Mar 2024;155(3):227–232.e6. https://doi.org/10.1016/j.adaj.2023.11.004
36. Cai Y, Zhao R, Zhao H, Li Y, Gou L. Exploring the use of ChatGPT/GPT-4 for patient follow-up after oral surgeries. Int J Oral Maxillofacial Surg Oct. 2024;53(10):867–72. https://doi.org/10.1016/j.ijom.2024.04.002
37. Daraqel B, Wafaie K, Mohammed H, et al. The performance of artificial intelligence models in generating responses to general orthodontic questions: ChatGPT vs Google bard. Am J Orthod Dentofac Orthop. 2024;165(6):652–62.
38. Şahin MF, Ateş H, Keleş A, et al. Responses of five different artificial intelligence chatbots to the top searched queries about erectile dysfunction: a comparative analysis. J Med Syst. 2024;48(1):38.
39. Bernard A, Langille M, Hughes S, Rose C, Leddin D, Veldhuyzen van Zanten S. A systematic review of patient inflammatory bowel disease information resources on the world wide web. Am J Gastroenterol Sep. 2007;102(9):2070–7. https://doi.org/10.1111/j.1572-0241.2007.01325.x
40. Altman DG. Practical statistics for medical research. Chapman and Hall/CRC. 1990.
41. Onder C, Koc G, Gokbulut P, Taskaldiran I, Kuskonmaz S. Evaluation of the reliability and readability of ChatGPT-4 responses regarding hypothyroidism during pregnancy. Sci Rep. 2024;14(1):243.
42. Sezgin E, Chekeni F, Lee J, Keim S. Clinical accuracy of large Language models and Google search responses to postpartum depression questions: Cross-Sectional study. J Med Internet Res Sep. 2023;11:25:e49240. https://doi.org/10.2196/49240
43. Daraz L, Morrow AS, Ponce OJ, et al. Readability of online health information: A Meta-Narrative systematic review. Am J Med Quality: Official J Am Coll Med Qual Sep/Oct. 2018;33(5):487–92. https://doi.org/10.1177/1062860617751639
44. Flesch R. A new readability yardstick. J Appl Psychol Jun. 1948;32(3):221–33. https://doi.org/10.1037/h0057532
45. Kincaid JP, Fishburne RP Jr, Rogers RL, Chissom BS. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Defense Technical Information Center, 1975. https://apps.dtic.mil/sti/citations/tr/ADA006655
46. Weiss BD. Health literacy. Am Med Assoc. 2003;253:358.
47. Naz R, Akacı O, Erdoğan H, Açıkgöz A. Can large Language models provide accurate and quality information to parents regarding chronic kidney diseases? J Evaluation Clin Pract Dec. 2024;30(8):1556–64. https://doi.org/10.1111/jep.14084
48. Joseph T, Sanghavi N, Kanyal S, Majumder K, Seidu-Aroza H, Godavarthi A. Comparative analysis of ChatGPT and Google gemini in the creation of patient education materials for acute appendicitis, cholecystitis, and hydrocele. Indian J Surg. 2024;87:1–6. https://doi.org/10.1007/s12262-024-04112-y
49. Lee TJ, Campbell DJ, Patel S, et al. Unlocking health literacy: the ultimate guide to hypertension education from ChatGPT versus Google gemini. Cureus May. 2024;16(5):e59898. https://doi.org/10.7759/cureus.59898
50. Geantă M, Bădescu D, Chirca N, et al. The emerging role of large Language models in improving prostate cancer literacy. Bioengineering (Basel, Switzerland). Jun. 2024;27(7). https://doi.org/10.3390/bioengineering11070654
51. Behers BJ, Vargas IA, Behers BM, et al. Assessing the readability of patient education materials on cardiac catheterization from artificial intelligence chatbots: an observational Cross-Sectional study. Cureus Jul. 2024;16(7):e63865. https://doi.org/10.7759/cureus.63865
52. Alshehri A, Alghofaili N, Alkadi RAL. Quality and readability assessment of Internet-Based information on common prosthodontic treatments. Int J Prosthodont January/February. 2022;35(1):62–7. https://doi.org/10.11607/ijp.7063
53. Ghanem YK, Rouhi AD, Al-Houssan A, et al. Dr. Google to dr. ChatGPT: assessing the content and quality of artificial intelligence-generated medical information on appendicitis. Surg Endosc. 2024;38(5):2887–93.
54. Gieg SD, Stannard JP, Cook JL. Evaluating the role and impact of health literacy for patients undergoing knee surgery. J Knee Surg Dec. 2023;36(14):1399–404. https://doi.org/10.1055/a-2106-3638
55. Shahid R, Shoker M, Chu LM, Frehlick R, Ward H, Pahwa P. Impact of low health literacy on patients' health outcomes: a multicenter cohort study. BMC Health Serv Res Sep. 2022;12(1):1148. https://doi.org/10.1186/s12913-022-08527-9
56. Authors, Clark M, Bailey S. CADTH horizon scans. Chatbots in health care: connecting patients to information: emerging health technologies. Canadian Agency for Drugs and Technologies in Health 2024.
57. Wright BM, Bodnar MS, Moore AD, et al. Is ChatGPT a trusted source of information for total hip and knee arthroplasty patients? Bone & joint open. Feb. 2024;15(2):139–46. https://doi.org/10.1302/2633-1462.52.Bjo-2023-0113.R1

Publisher's note