BMC Medical Informatics and Decision Making

RESEARCH

Open Access

Can artificial intelligence models serve as patient information consultants in orthodontics?

Derya Dursun[1] and Rumeysa Bilici Geçer[2*]

Abstract

Background To evaluate the accuracy, reliability, quality, and readability of responses generated by ChatGPT-3.5, ChatGPT-4, Gemini, and Copilot in relation to orthodontic clear aligners.

Methods Frequently asked questions by patients/laypersons about clear aligners on websites were identified using the Google search tool and these questions were posed to ChatGPT-3.5, ChatGPT-4, Gemini, and Copilot AI models. Responses were assessed using a five-point Likert scale for accuracy, the modified DISCERN scale for reliability, the Global Quality Scale (GQS) for quality, and the Flesch Reading Ease Score (FRES) for readability.

Results ChatGPT-4 responses had the highest mean Likert score (4.5 ± 0.61), followed by Copilot (4.35 ± 0.81), ChatGPT-3.5 (4.15 ± 0.75) and Gemini (4.1 ± 0.72). The difference between the Likert scores of the chatbot models was not statistically significant ($p > 0.05$). Copilot had a significantly higher modified DISCERN and GQS score compared to both Gemini, ChatGPT-4 and ChatGPT-3.5 ($p < 0.05$). Gemini's modified DISCERN and GQS score was statistically higher than ChatGPT-3.5 ($p < 0.05$). Gemini also had a significantly higher FRES compared to both ChatGPT-4, Copilot and ChatGPT-3.5 ($p < 0.05$). The mean FRES was 38.39 ± 11.56 for ChatGPT-3.5, 43.88 ± 10.13 for ChatGPT-4 and 41.72 ± 10.74 for Copilot, indicating that the responses were difficult to read according to the reading level. The mean FRES for Gemini is 54.12 ± 10.27, indicating that Gemini's responses are more readable than other chatbots.

Conclusions All chatbot models provided generally accurate, moderate reliable and moderate to good quality answers to questions about the clear aligners. Furthermore, the readability of the responses was difficult. ChatGPT, Gemini and Copilot have significant potential as patient information tools in orthodontics, however, to be fully effective they need to be supplemented with more evidence-based information and improved readability.

Keywords Clear aligner, Chatbots, Artificial Intelligence, ChatGPT, Gemini, Copilot

*Correspondence:
Rumeysa Bilici Geçer
drrumeysabilici@gmail.com
[1]Department of Orthodontics, Hamidiye Faculty of Dentistry, University of Health Sciences, Istanbul, Turkey
[2]Department of Orthodontics, Faculty of Dentistry, Istanbul Aydin University, Istanbul, Turkey

Introduction

The term "artificial intelligence" (AI), defined as computerized synthetic human cognitive function, first appeared in 1956 and is widely used today [1]. AI is the term used to describe machines or software systems designed to perform tasks that typically require human intelligence, including making decisions, solving problems, and learning from experience [2]. Machine Learning (ML) and Large Language Models (LLM), which are part of AI, are also used in medicine and dentistry to help professionals provide better oral health services [3, 4]. LLM is often used as a large-scale language model trained with deep learning techniques, and they form the basis for various Natural Language Processing (NLP) tasks, a subset of AI that enables machines to understand, interpret and produce human-like texts. Chatbots are language models that automatically understand and respond to human user queries using NLP and ML algorithms [5]. Today there are different language models with different characteristics.

Firstly, a new artificial intelligence LLM called ChatGPT was created by OpenAI Inc. (San Francisco, CA, USA). The ChatGPT version that is freely accessible is based on the GPT-3.5 language model. In contrast, the newer GPT-4 version is available exclusively under the ChatGPT Plus paid subscription. In the initial three-month period following its inauguration, the platform attracted a remarkable 100 million new users [4]. Following the popularity of ChatGPT, in February 2023 Microsoft (Microsoft Corporation, Redmond, WA, USA) introduced the Bing Chat AI chatbot using the GPT-4 language model, which is currently being relaunched as Copilot [4]. Copilot uses GPT-4 technology as well as Code Interpreter and DALL-E 3 for encoding and rendering respectively [6]. Copilot also works effectively with Microsoft 365 applications. It has been stated that Copilot addresses some of the important issues commonly encountered in ChatGPT, such as keeping up to date with current events via internet access, providing footnotes with links to sources for the information received, and having live internet access [6]. In March 2023, Google (Google Ireland Limited, Dublin, Ireland) introduced the Google Bard language model initially powered by LaMDA and later by PaLM 2 LLM, which is currently being relaunched as Gemini [4]. These models and their respective updates have several distinguishing features. ChatGPT's responses are based on pre-existing training data, while Gemini uses real-time access to the Internet to incorporate up-to-date information when generating responses [7]. Although Copilot has some advantages, such as live Internet access, it has a limit of 100 requests per day compared to ChatGPT's 70 requests per hour [4]. Another difference is that ChatGPT-3.5, Gemini and Copilot are publicly available, while ChatGPT-4 requires a paid subscription and is more difficult to access.

The use of AI chatbots is becoming increasingly popular among healthcare professionals and patients as a convenient source of medical and dental information due to technological advancements [8]. AI chatbots offer patients 24/7 accessibility, fast responses to questions, and minimize the need for appointments, enabling them to access information at any time [2]. In addition to these features, the ability of the system to provide incorrect answers, generate irrelevant content, and present false information and disinformation as fact raises serious concerns in critical areas such as health [3]. It is essential to evaluate health-related content created by AI objectively [9]. The reliability and quality of the information source in chatbots is crucial as it can affect patients' cooperation and compliance in treatment, as well as doctor-patient communication and trust.

Clear aligner treatment is one of the most popular orthodontic developments due to the increasing demand for aesthetic perception in orthodontic treatment [8]. Clear aligners have advantages, such as meeting aesthetic expectations and ease of use. However, there are also issues to consider, such as the duration of use, patient compliance, personal motivation, and cleaning [10]. For successful treatment, patients must receive accurate information about clear aligners. According to a survey, 57% of patients prefer to consult the internet first for health-related information [11]. In particular, the objectivity of promotional and advertising content information of social media sources is controversial [12]. Therefore, the accuracy, reliability, and content knowledge of chatbots trained on large text datasets from the internet (including Wikipedia, digitized books, articles, and web pages) are important in terms of clear aligner treatments and information [4].

As the field of health decision-making continues to evolve, patients are increasingly utilising AI tools. Despite the potential benefits of AI chatbots, there are several limitations that require consideration. These include the risk of misinformation, the lack of specialized medical knowledge, and the potential for producing unrealistic outputs, such as hallucinations [13]. As chatbot responses can cause serious problems in areas where critical information is needed (medicine, dentistry, etc.), it is important that the responses are verified by physicians.

Recent literature has investigated the capacity of AI models, such as ChatGPT, to provide answers to general questions across different fields, including medicine and dentistry [1, 5]. Daraqel et al. [7] highlighted that there is still a notable gap in the comprehensive evaluation of the performance of different AI models in orthodontics. To our knowledge, this study is the first to examine

Table 1 The queries that were asked to ChatGPT-3.5, ChatGPT-4, Gemini, and Copilot

1. What are clear aligners?
2. How do clear aligners straighten the teeth?
3. What's the difference between clear aligners and braces?
4. What malocclusions can be treated with clear aligners?
5. What are the benefits of clear aligners?
6. Can clear aligners be more effective than braces in solving orthodontic problems?
7. Are clear aligners more expensive than braces?
8. Is clear aligner treatment covered by health insurance?
9. What is the recommended duration for wearing clear aligners during the day?
10. How often do I need to change my clear aligners?
11. How long does the clear aligner treatment last?
12. How frequently do I have to visit my orthodontist during clear aligner treatment?
13. Do clear aligners affect eating?
14. Is it possible to drink tea/coffee or smoke with clear aligners?
15. How to clean clear aligners?
16. Is there any effect of clear aligners on speech?
17. Are clear aligners painful?
18. What are the clear aligners made of?
19. What is interproximal reduction in the treatment of clear aligners?
20. What are clear aligner attachments?

the responses of the most used AI models, including, Google's Gemini, OpenAI's ChatGPT-3.5, and ChatGPT-4, and Microsoft's Copilot, to questions about clear aligner treatments that are frequently asked by patients on websites. Reliability, quality, readability, and accuracy were evaluated to provide a comprehensive assessment of the responses. The main contribution of our study is an in-depth content analysis of the information provided by different AI chatbots to frequently asked questions by patients, especially about clear aligners, which are a popular and common orthodontic treatment today.

Materials and methods

Ethical approval was not required as no human/animal participants were involved in the study. A search was conducted using the Google Search Tool to identify websites that responded to the search term "frequently asked questions about clear aligners" [3]. In the literature, it is reported that 90% of search-engine users only viewed the first three pages of search results [14]. Therefore, the first 30 sites were analysed and a pool of 180 questions was created by excluding 'irrelevant', 'duplicate', non-English and 'sponsored advertising' sites. The 20 most frequently asked questions were included by the authors, with the exclusion of repetitive (64 questions), similar (50 questions), irrelevant (36 questions) and brand-related (10 questions) questions (Table 1)(Fig. 1).

The chatbots tested were: (i) ChatGPT model GPT-3.5, which is currently available for free. (ii) ChatGPT model GPT-4, which is available as part of a subscription in ChatGPT Plus. (iii) Gemini. (iv) Copilot.

It has been reported that ChatGPT may potentially provide different and faster responses when asked the same question again or at different time points [15]. Accordingly, all questions were posed only once. Furthermore, Gemini generates 3 versions, or drafts, of each response [7]. The initial draft was selected for evaluation. Copilot has three different speaking style modes, and the 'more balanced' standard mode was chosen. To prevent

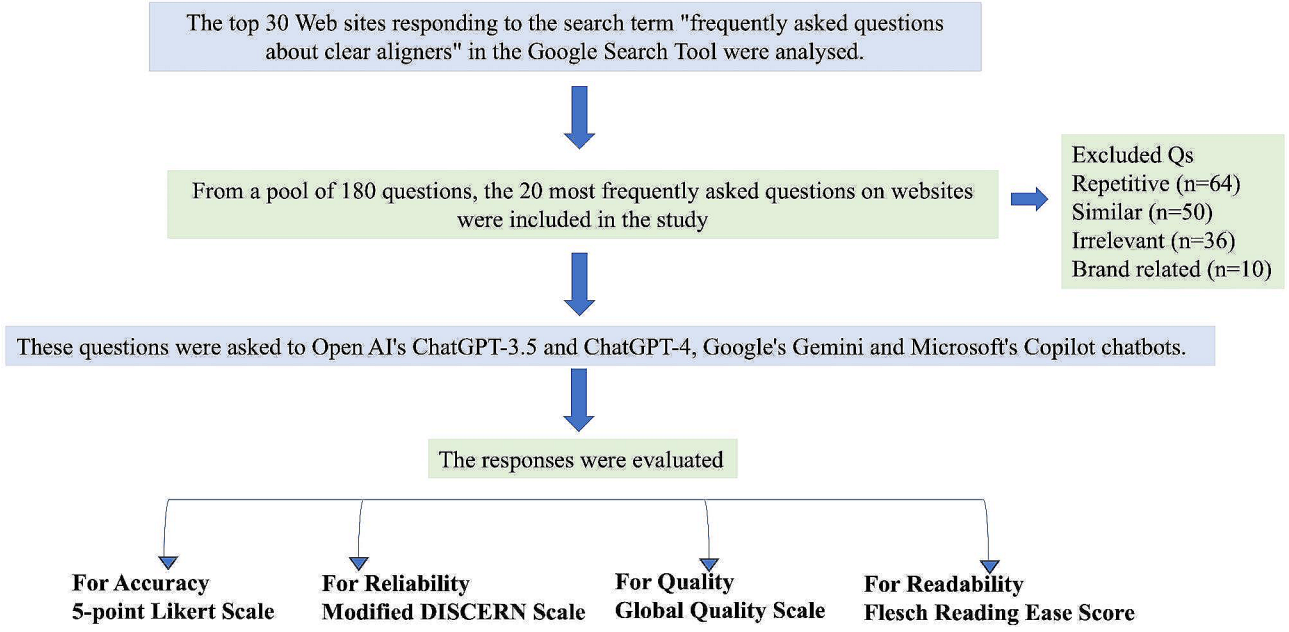

Fig. 1 Flowchart of the study

the correlation of answers, a new user record was created for each model and a separate chat window was opened for each question. All questions were posed on the same day using the same laptop (Windows 11 Home, Intel(R) Core(TM) i7-12700 H 2.70 GHz, 16 GB RAM, Nvidia GeForce RTX 4060 8 GB Graphics Card) and the same fixed fiber internet network (100 Mbps). Consequently, responses from ChatGPT-3.5, ChatGPT-4, Gemini and Copilot were entered into four separate forms (A, B, C and D), with the removal of all words related to each AI model, to ensure the blindness of the evaluators. The responses of AI chatbots to the questions are presented in the Additional file 1. All responses were subjected to independent assessment by experienced study authors (D.D and R.B.G), with reference to the current literature and clinical practice.

For accuracy, a five-point Likert scale was employed [16, 17]. According to this scale; Score 1; the chatbot's answers are completely incorrect, Score 2; the chatbot's answers contain more incorrect items than correct items, Score 3; the chatbot's answers contain an equal balance of correct and incorrect items, Score 4; the chatbot's answers contain more correct items than incorrect items and Score 5; the chatbot's answers are completely correct.

The DISCERN scale is a three-part scale that has been used in previous studies to evaluate the reliability and quality of online health information [3, 18] In this study, only the initial eight-question section of the DISCERN scale was used to assess the reliability of AI chatbot responses [19]. The 8 questions of the modified DISCERN scale evaluate the following; (i) clarity of objectives, (ii) achievement of objectives, (iii) relevant, (iv) what sources of information were used, (v) when the information used or reported was produced, (vi) balance and unbiased, (vii) details of additional support and sources of information, (viii) reference to areas of uncertainty. For each question in the modified DISCERN scale, the total score was calculated by scoring the no answer as 1, the partial answer as 2–3–4, and the yes answer as 5. The total score was then categorized as poor (8–15 point), moderate (16–31 point), or good (32–40 point) [20].

Global Quality Scale (GQS) analyze the quality of written sources in the field of medicine [2, 21]. In the GQS, the lowest score is 1 and the highest score is 5. According to this scale; Score 1; poor quality, most information missing, not useful for patients, Score 2; generally poor quality, many important topics missing and very limited use for patients, Score 3; moderate quality, some important information adequately discussed but others insufficiently discussed,

Score 4; good quality, most of the relevant information listed is useful for patients, and Score 5; excellent quality and excellent flow, very useful for patients. According to GQS scoring; 1–2 points representing low quality, 3 points moderate quality, and 4–5 points high quality [20].

For readability, the Flesch Reading Ease Score (FRES) was employed [3, 20]. Readability of the response texts was measured using the Microsoft Word for Mac Flesch-Reading Ease Score calculator (version 16.75 [23,070,901]; Microsoft, Redmond, Wash). FRES is as follows: 206.835–1.015 x (total words/total sentences) − 84.6 3 (total syllables/total words). In this scoring system, a score between 0 and 100 is obtained according to a calculation and a reading level is determined between easy and difficult (Table 2).

Table 2 The interpretation of the core and reading levels

Flesch Reading Ease Score	Reading Level
0–29	Very difficult
30–49	Difficult
50–59	Fairly difficult
60–69	Standard and/or plain
70–79	Fairly easy
80–89	Easy
90–100	Very easy

Statistical analysis

Statistical analysis was performed using NCSS (Number Cruncher Statistical System) 2007 statistical software (Utah, USA). Quantitative data were described using the mean, standard deviation (SD), median and interquartile range (IQR). For comparisons between the groups, Shapiro-Wilk test was used to verify the normality of distribution. One-way analysis of variance (ANOVA) and post-hoc Tukey's multiple comparison test were used for intergroup comparisons of normally distributed variables. Otherwise, Kruskal-Wallis test and post-hoc Dunn's multiple comparison test were used. Interobserver agreement was evaluated with the intraclass correlation coefficient (ICC). The results were evaluated at a significance level of $p<0.05$.

Results

ICC values ranged between from 0.829 to 0.979 for interrater reliability (Table 3).

Likert, modified DISCERN, GQS and FRES scores of the chatbots' answers to questions about clear aligners are shown in Table 4. Normally distributed data are presented as mean±SD, and non-normally distributed data are presented as median (IQR).

ChatGPT-4 responses had the highest mean Likert score (4.5±0.61), followed by Copilot (4.35±0.81), ChatGPT-3.5 (4.15±0.75) and Gemini (4.1±0.72) (Table 4). The difference between the Likert scores of the chatbot models was not statistically significant ($p>0.05$) (Table 5).

Table 3 The intraclass correlation coefficient of evaluators data

		Intraclass Correlation Coefficient %95 CI
Likert Score	ChatGPT-3.5	0.919 (0.796–0.968)
	ChatGPT-4	0.829 (0.767–0.932)
	Gemini	0.946 (0.864–0.979)
	Copilot	0.911 (0.774–0.965)
Modified DISCERN Score	ChatGPT-3.5	0.952 (0.879–0.981)
	ChatGPT-4	0.979 (0.947–0.992)
	Gemini	0.967 (0.917–0.987)
	Copilot	0.945 (0.860–0.978)
GQS Score	ChatGPT-3.5	0.930 (0.824–0.972)
	ChatGPT-4	0.829 (0.768–0.932)
	Gemini	0.915 (0.784–0.966)
	Copilot	0.882 (0.799–0.922)

Copilot responses had the highest mean modified DISCERN score (25.4±3.39), followed by Gemini (21.1±2.63), ChatGPT-4 (19.7±2.15) and ChatGPT-3.5 (18.65±2.25) (Table 4). Copilot had a significantly higher modified DISCERN score compared to both Gemini, ChatGPT-4 and ChatGPT-3.5 ($p<0.05$). Gemini's modified DISCERN score was statistically higher than ChatGPT-3.5 ($p<0.05$) (Table 5). According to the modified DISCERN classification, 95% of Copilot, Gemini, and ChatGPT-4 responses and 85% of ChatGPT-3.5 responses were moderately reliable (Table 6).

Copilot responses had the highest mean GQS score (4.4±0.68), followed by Gemini (3.95±0.69), ChatGPT-4 (3.8±0.62) and ChatGPT-3.5 (3.5±0.61) (Table 4). Copilot had a significantly higher GQS score compared to Gemini, ChatGPT-4 and ChatGPT-3.5 ($p<0.05$). Gemini's GQS score was statistically higher than ChatGPT-3.5 ($p<0.05$) (Table 5). According to the GQS classification, 90% of Copilot's responses were high quality, 80% of Gemini's responses were high quality, 65% of ChatGPT-4's responses were high quality, 55% of ChatGPT-3.5's responses were moderate quality (Table 6).

The mean FRES was 38.39±11.56 for ChatGPT-3.5, 43.88±10.13 for ChatGPT-4 and 41.72±10.74 for Copilot, indicating that the responses were difficult to read according to the reading level (Table 2). The mean FRES for Gemini was 54.12±10.27, indicating that the responses were fairly difficult to read according to the reading level (Table 2). Gemini also demonstrated a significantly higher FRES compared to both ChatGPT-4, Copilot and ChatGPT-3.5 ($p<0.05$). There was no statistically significant difference between ChatGPT-4, Copilot and ChatGPT-3.5 ($p>0.05$) (Table 5).

Table 5 Post-hoc pairwise comparison of scores in AI chatbots

	Modified DISCERN Score *p*	Flesch Reading Ease Score *p*	GQS Score *p*
ChatGPT-3.5 vs. ChatGPT-4	0.597	0.371	0.165
ChatGPT-3.5 vs. Gemini	**0.023**†	**0.001**†	**0.036**†
ChatGPT-3.5 vs. Copilot	**0.0001**†	0.758	**0.0001**†
ChatGPT-4 vs. Gemini	0.348	**0.017**†	0.483
ChatGPT-4 vs. Copilot	**0.0001**†	0.919	**0.007**†
Gemini vs. Copilot	**0.0001**†	**0.002**†	**0.042**†

†$p<0.05$, Tukey's post hoc test was used for Modified DISCERN and Flesch Reading Ease Score, and Dunn's post hoc test was used for GQS.

Table 6 Score distribution of chatbot responses according to the modified DISCERN scale and GQS classification

		ChatGPT-3.5	ChatGPT-4	Gemini	Copilot
		n (%)	n (%)	n (%)	n (%)
Modified DISCERN Score	Poor (8–15 point)	3 (15%)	1 (5%)	0 (0%)	0 (0%)
	Moderate (16–31 point)	17 (85%)	19 (95%)	19 (95%)	19 (95%)
	Good (32–40 point)	0 (0%)	0 (0%)	1 (0%)	1 (5%)
GQS Score	Low quality (Score 1 or 2)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
	Moderate quality (Score 3)	11 (55%)	7 (35%)	4 (20%)	2 (10%)
	High quality (Score 4 or 5)	9 (45%)	13 (65%)	16 (80%)	18 (90%)

Categorical variables (number of questions) are shown as n (%) in the table.

Table 4 A comparison of the Likert, modified DISCERN, GQS and FRES scores of four different chatbots

		ChatGPT-3.5	ChatGPT-4	Gemini	Copilot	*p*
Likert Score	mean±SD	4.15±0.75	4.5±0.61	4.1±0.72	4.35±0.81	0.259‡
	median (IQR)	4 (4-5)	5 (4-5)	4 (4-5)	5 (4-5)	
Modified DISCERN Score	mean±SD	18.65±2.25	19.7±2.15	21.1±2.63	25.4±3.39	**0.0001**†
GQS Score	mean±SD	3.5±0.61	3.8±0.62	3.95±0.69	4.4±0.68	**0.001**‡
	median (IQR)	3 (3-4)	4 (3-4)	4 (3,25 – 4)	4,5 (4-5)	
Flesch Reading Ease Score	mean±SD	38.39±11.56	43.88±10.13	54.12±10.27	41.72±10.74	**0.0001**†

†ANOVA, ‡Kruskal–Wallis analysis

Discussion

In the 21st century, a significant proportion of patients and parents use the internet and social media platforms to find out more about health-related issues. Clear aligners are one of the most discussed orthodontic developments due to the increasing demand for aesthetic orthodontic treatment [8]. The objectivity of claims made by commercial companies and advertising sources about clear aligners, especially on social media and the internet, is questionable [12]. Therefore, the quality of health information about clear aligners, which is of great interest to patients, is important. AI chatbots are a potential source for patients to access information about clear aligners. The objective of this study is to evaluate the responses to frequently asked questions about the clear aligner using a comprehensive overview and various AI models.

AI chatbots interact with users using NLP and machine learning techniques, with the overarching goal of facilitating access to healthcare services, providing quick access to medical information, and reducing the burden on healthcare systems [13]. At this point, it is of great importance that factors such as accuracy, reliability, quality, and readability of chatbot responses in the medical field are audited by clinicians. A study in the literature evaluating the quality of information provided by ChatGPT-4 about periodontal disease using the DISCERN tool found that, despite some limitations, the AI consistently provided accurate guidance for most responses [22]. Bhattacharyya et al. [23] emphasize the need for caution when searching for medical information on ChatGPT, as many of the references provided in the medical content were found to be inaccurate. In addition, Eggman et al. [9] evaluated the impact of language models such as ChatGPT on dentistry and stated that further research and development is needed to fully realize the potential benefits of language models in areas such as clinical decision support, patient education and dental education.

In the literature, different indices have been used to evaluate the accuracy of AI chatbot responses, in this study a five-point Likert scale was used. [7, 8, 16]. In the study, all AI chatbot models generally produced reasonably accurate to the most frequently asked questions of patients about clear aligner. Both evaluators scored ChatGPT-4's responses as the most accurate, followed by the answers of Copilot, Gemini, and ChatGPT-3.5, respectively. In a study evaluating the responses of four different chatbots to clinical questions about general orthodontics, Makrygiannakis et al. [4] found that the best answers were given by Bing, ChatGPT-4, Google Bard, and ChatGPT-3.5, respectively. Similarly, Daraqel et al. [7] evaluated the accuracy of ChatGPT and Google Bard's answers to general orthodontic questions and reported that the average response accuracy level was high in both AI models. Tanaka et al. [24] also used a 5-point Likert Scale to assess accuracy and found that ChatGPT-4 provided useful information on clear aligners, temporary anchorage devices and digital imaging. On the other hand, Arqud et al. [8] have indicated that the overall accuracy of ChatGPT-3.5 responses to queries regarding clear aligners is inadequate. This difference is thought to be related to the content of the questions. While Arqud et al. [8] asked detailed and technical questions about clear aligners, this study included questions frequently asked by patients about clear aligners. Indeed, Balel [25] reported that ChatGPT provided good to excellent reliable responses to patient (layman) queries in accordance with our study, while responses to physician queries were of a moderate quality. AI chatbot models may have significant potential as a patient information tool, but their use in technical questions and training may not be completely safe, as Balel and Arqud et al. [8, 25].

Today, the reliability of health information is of great importance for patients utilizing the Internet as a source of information. It is possible for AI chatbots such as ChatGPT to present information that is false or misleading, and patients may be inclined to believe this information [3]. Topics such as sources of information, citations and references are important in evaluating reliability of health information and were evaluated in our study using the modified DISCERN scale. In the study, all AI chatbot models generally produced moderately reliable answers to the questions patients most frequently asked about the clear aligner. Both evaluators rated Copilot's responses as the most reliable, followed by Gemini, ChatGPT-4, and ChatGPT-3.5. This result is in accordance with Copilot's and Gemini's high score in the DISCERN tool in the questions related to 'specifying the information sources used in the creation of the content'. Copilot used a total of 87 references in all questions. Gemini used 4 references in total, only in 4 questions. The fact that Gemini and Copilot provide references or citations can be attributed to their real-time access to up-to-date information [4]. A significant limitation of chatbots such as ChatGPT-3.5 and ChatGPT-4 in responding to health-related questions is their inability to provide references or citations for the information they generate [26]. Similarly, the present study found that Chat-GPT 3.5 and ChatGPT-4 did not provide references or citations to questions about clear aligners. In the study, the 19th question in Additional file 1 asked: "What is interproximal reduction in clear aligner treatment?" According to the modified DISCERN scoring, overall, all AI chatbot responses were generally clear, relevant, and unbiased. However, Copilot provided referenced information about the clinical practice of interproximal reduction. Similarly, Gemini provided referenced information about the methods of interproximal reduction and referred to its website. As

a result, Copilot and Gemini had a higher modified DISCERN score for this question.

Kılınç et al. [3] evaluated the reliability of ChatGPT answers to orthodontic questions and reported that the answers were not scientifically based and did not provide peer-reviewed references. The availability of accurate, evidence-based information from reliable sources empowers patients to make informed decisions about their health [26]. Copilot cited 75 websites and 9 academic articles as references, while Gemini's references were all websites. Seth et al. reported that Bing's references were below average, offering few academic articles [27]. Although Copilot provided the most citations and references from academic articles in this study, more use of evidence-based scientific data is needed for the reliability of language models. Chatbots are language models trained on a variety of internet data that produce text based on statistical models covering a wide range of topics [27]. Accordingly, the content of the websites also becomes of significant importance. In the literature, the quality of web-based content about clear aligners was evaluated with the DISCERN scale, and found the quality of website content poor [28, 29]. The improved content resources of websites prepared by experts, supported by evidence-based scientific data, can enhance the learning and responses of language models.

Balel [25] asked ChatGPT the questions patients ask about maxillofacial surgery and the answers were found to be good to excellent and reliable according to the GQS scale. Another study, Acar [2] reported that the responses provided by ChatGPT, Bard, and Bing regarding oral surgery exhibited high GQS scores. Similarly in the present study, Copilot, Gemini, and ChatGPT-4 gave good quality answers to frequently questions about clear aligners, while ChatGPT-3.5 provided medium/good quality answers. In the study, the 20th question in Additional file 1 asked: "What are clear aligner attachments?" According to GQS, ChatGPT-3.5's answer was given a score of 3. The answer was of medium quality, there was some information, but not enough. There was no information about power transmission, what tooth movements it is needed for, that it should only be used for more complex movements that are difficult to achieve with clear aligners. Gemini also received the same score. ChatGPT-4 gave a good quality response with more information listed for patients on topics such as improved grip, directional forces, precise movements and was given a score of 4. Copilot gave very useful and excellent quality responses for patients, providing additional information on the different types of attachments, passive and active surfaces of attachments, dimensions of attachments and was given a score of 5.

Health literacy is defined as the ability to read and understand health-related information, enabling individuals to make informed decisions and manage treatment processes [3]. In the literature, readability is identified as an essential element of health literacy, ensuring that documents are understood [30]. Readability is evaluated using different indices, including sentence length and the use of difficult words [31]. FRES between 0 and 100, with a higher score indicating better readability; 60–70 is standard English, and 65 is an acceptable target [32]. Although Gemini had the highest FRES score in the study, it was found that all AI chatbots were below 60 points and difficult to read. Similarly, in the study evaluating the readability of responses about rhinoplasty, Google Bard showed the highest FRES, followed by ChatGpt and Bing [27]. Onder et al. reported that the readability of responses generated by ChatGPT-4 regarding hypothyroidism in pregnancy was difficult. The difficulty in readability of chatbots will limit their easy use by the public. Expert produced websites can improve the performance of language models trained on a wide range of internet data by providing high quality, clear and readable health information.

There were some limitations of our study. First, responses were analyzed only in English, so results cannot be generalized to all languages. Chatbots, which are based on deep learning, collect new data that is uploaded to the internet every second and create new answers. As a result, changing answers makes control difficult. Should the same question be posed again or at different points in time, it is possible that an AI chatbots may provide a different answer. It should be noted that all questions were posed only once, and the initial responses were evaluated. In addition, there are currently studies involving prompting of language models [13]. Studies and solutions to increase the effectiveness of AI language models in providing accurate and unbiased information are discussed. Based on the findings of the study, it is evident that further studies are necessary to refine the responses of AI language models and enhance their utility in the field of orthodontics.

Conclusions

All chatbot models provided generally accurate, moderate reliable and moderate to good quality answers to questions about the clear aligners. Copilot's responses were the most reliable and quality, followed by Gemini, ChatGPT-4, and Chat GPT-3.5. Copilot and Gemini provided references or citations in responses about clear aligners. A major limitation of ChatGPT-3.5 and ChatGPT-4 is that they did not provide references or citations for the answers they generated. Furthermore, the FRES indicated that the reading level of the AI chatbots responses was difficult. ChatGPT, Gemini and Copilot have significant potential as patient information tools in orthodontics, however, to be fully effective they need to

be supplemented with more evidence-based information and improved readability.

Abbreviations

AI	Artificial Intelligence
ML	Machine Learning
LLM	Large Language Models
NLP	Natural Language Processing
GQS	Global Quality Score
FRES	Flesch Reading Ease Score

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-024-02619-8.

Supplementary Material 1

Supplementary Material 2

Acknowledgements
Not applicable.

Author contributions
All authors contributed to the understanding and design of the study. R.B.G formulated the idea and designed the study. R.B.G and D.D collected data. R.B.G and D.D conducted the analysis. The first draft of the article was written by R.B.G and D.D. R.B.G analyzed the manuscript in a critical way. All authors have read and approved the final article.

Funding
The authors did not receive support from any organization for the submitted work.

Data availability
No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate
Present study is a public application, and there is no human/animal participant, ethics committee approval was not required.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 12 June 2024 / Accepted: 23 July 2024
Published online: 29 July 2024

References

1. Thurzo A, Urbanova W, Novak B, Czako L, Siebert T, Stano P et al. Where is the Artificial Intelligence Applied in Dentistry? Systematic Review and Literature Analysis. Healthc (Basel). 2022;10(7).
2. Acar AH. Can natural language processing serve as a consultant in oral surgery? J Stomatol Oral Maxillofac Surg. 2023;125(3):101724.
3. Kilinc DD, Mansiz 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.
4. Makrygiannakis MA, Giannakopoulos K, Kaklamanos EG. Evidence-based potential of generative artificial intelligence large language models in orthodontics: a comparative study of ChatGPT, Google Bard, and Microsoft Bing. Eur J Orthod. 2024.
5. Chakraborty C, Pal S, Bhattacharya M, Dash S, Lee SS. Overview of Chatbots with special emphasis on artificial intelligence-enabled ChatGPT in medical science. Front Artif Intell. 2023;6:1237704.
6. Cascella M, Semeraro F, Montomoli J, Bellini V, Piazza O, Bignami E. The breakthrough of large Language models Release for Medical Applications: 1-Year Timeline and perspectives. J Med Syst. 2024;48(1):22.
7. Daraqel B, Wafaie K, Mohammed H, Cao L, Mheissen S, Liu Y, Zheng L. The performance of artificial intelligence models in generating responses to general orthodontic questions: ChatGPT vs Google Bard. Am J Orthod Dentofac Orthop. 2024.
8. Abu Arqub S, Al-Moghrabi D, Allareddy V, Upadhyay M, Vaid N, Yadav S. Content analysis of AI-generated (ChatGPT) responses concerning orthodontic clear aligners. Angle Orthod. 2024.
9. Eggmann F, Weiger R, Zitzmann NU, Blatz MB. Implications of large language models such as ChatGPT for dental medicine. J Esthetic Restor Dentistry. 2023;35(7):1098–102.
10. Wajekar N, Pathak S, Mani S. Rise & review of invisalign clear aligner system. IP Indian J Orthod Dentofac Res. 2022;8(1):7–11.
11. PM360. Howconsumersfindanduseonlinehealthrelatedcontent in 2017. [https://www.pm360online.com/how-consumers-find-and-use-online-health-related-content-in-2017/.
12. Alkadhimi A, Al-Moghrabi D, Fleming PS. The nature and accuracy of Instagram posts concerning marketed orthodontic products. Angle Orthod. 2022;92(2):247–54.
13. Shi W, Zhuang Y, Zhu Y, Iwinski H, Wattenbarger M, Wang MD, editors. Retrieval-augmented large language models for adolescent idiopathic scoliosis patients in shared decision-making. Proceedings of the 14th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics; 2023.
14. Ustdal G, Guney AU. YouTube as a source of information about orthodontic clear aligners. Angle Orthod. 2020;90(3):419–24.
15. Vaishya R, Misra A, Vaish A. ChatGPT: Is this version good for healthcare and research? Diabetes & Metabolic Syndrome: Clinical Research & Reviews. 2023;17(4):102744.
16. Hatia A, Doldo T, Parrini S, Chisci E, Cipriani L, Montagna L, et al. Accuracy and completeness of ChatGPT-Generated information on interceptive orthodontics: a Multicenter Collaborative Study. J Clin Med. 2024;13(3):735.
17. Johnson D, Goodman R, Patrinely J, Stone C, Zimmerman E, Donald R et al. Assessing the accuracy and reliability of AI-generated medical responses: an evaluation of the Chat-GPT model. Res Square. 2023.
18. Meade MJ, Dreyer CW. Web-based information on orthodontic clear aligners: a qualitative and readability assessment. Aust Dent J. 2020;65(3):225–32.
19. DISCERN online quailty criteria for consumer health information. http://www.discern.org.uk. Access 14 Jun 2024.
20. 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.
21. Cakir H, Caglar U, Sekkeli S, Zerdali E, Sarilar O, Yildiz O, Ozgor F. Evaluating ChatGPT ability to answer urinary tract infection-related questions. Infect Dis Now. 2024;54(4):104884.
22. Alan R, Alan BM. Utilizing ChatGPT-4 for providing information on periodontal disease to patients: a DISCERN quality analysis. Cureus. 2023;15(9).
23. Bhattacharyya M, Miller VM, Bhattacharyya D, Miller LE. High rates of fabricated and inaccurate references in ChatGPT-generated medical content. Cureus. 2023;15(5).
24. Tanaka OM, Gasparello GG, Hartmann GC, Casagrande FA, Pithon MM. Assessing the reliability of ChatGPT: a content analysis of self-generated and self-answered questions on clear aligners, TADs and digital imaging. Dent Press J Orthod. 2023;28:e2323183.
25. Balel Y. Can ChatGPT be used in oral and maxillofacial surgery? J Stomatology oral Maxillofacial Surg. 2023;124(5):101471.
26. 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.
27. Seth I, Lim B, Xie Y, Cevik J, Rozen WM, Ross RJ, Lee M, editors. Comparing the efficacy of large language models ChatGPT, BARD, and Bing AI in providing information on rhinoplasty: an observational study. Aesthetic Surgery Journal Open Forum; 2023: Oxford University Press US.
28. Meade M, Dreyer C. Web-based information on orthodontic clear aligners: a qualitative and readability assessment. Aust Dent J. 2020;65(3):225–32.

29. Alpaydın MT, Buyuk SK, Bavbek NC. Information on the internet about clear aligner treatment—an assessment of content, quality, and readability. J Orofac Orthop. 2022;83(Suppl 1):1.
30. Meade MJ, Dreyer CW. Orthodontic treatment consent forms: a readability analysis. J Orthodont. 2022;49(1):32–8.
31. McInnes N, Haglund BJ. Readability of online health information: implications for health literacy. Inf Health Soc Care. 2011;36(4):173–89.
32. Abou-Abdallah M, Dar T, Mahmudzade Y, Michaels J, Talwar R, Tornari C. The quality and readability of patient information provided by ChatGPT: can AI reliably explain common ENT operations? European Archives of Oto-Rhino-Laryngology. 2024:1–7.

Publisher's Note