

Comparative Evaluation of Reliability, Quality, Readability, and Originality of AI Chatbot Responses to Patient Questions in Implant Dentistry



Lana Alatrash¹  and Gülfem Ergun^{1,*} 

¹Department of Prosthodontics, Faculty of Dentistry, Gazi University, Ankara, Turkey

Abstract:

Introduction/Objective: This study aimed to evaluate and compare the reliability, informational quality, readability, and originality of responses generated by three widely used artificial intelligence chatbot platforms (Google Gemini Flash, ChatGPT Plus (OpenAI), and Claude Sonnet (Anthropic)) to patient-oriented questions in implant dentistry.

Methods: Twenty frequently asked questions in implant dentistry were developed and categorized as theoretical (n = 10) or clinical (n = 10). Each question was submitted to the three chatbots on the same day, and only the first response was analyzed. Responses were evaluated for reliability, informational quality, readability, and originality using validated assessment tools. Statistical comparisons were performed using appropriate parametric and nonparametric tests, with significance set at 0.05.

Results: Google Gemini Flash demonstrated significantly higher reliability scores than ChatGPT Plus and Claude Sonnet ($p < 0.001$). Informational quality also differed significantly among models ($p = 0.045$), with Gemini providing higher-quality responses than Claude. Readability analyses showed that chatbot responses generally required a high school or college-level reading ability. Similarity index values indicated high originality, with no significant differences among platforms ($p = 0.423$). Differences were more pronounced for theoretical questions, whereas clinical questions showed limited variation.

Discussion: Although AI chatbots can provide useful information for implant dentistry patients, variability in reliability and readability highlights the need for cautious use without professional supervision.

Conclusion: AI-powered chatbots can support patient education in implant dentistry; however, response quality and accessibility vary across models. Expert oversight remains essential, and future research should focus on improving readability, transparency, and standardized evaluation of AI-generated dental information.

Keywords: Artificial intelligence, Chatbots, Dental implants, Patient education, Reliability, Readability.

© 2026 The Author(s). Published by Bentham Open.

This is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International Public License (CC-BY 4.0), a copy of which is available at: <https://creativecommons.org/licenses/by/4.0/legalcode>. This license permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

*Address correspondence to this author at the Department of Prosthodontics, Faculty of Dentistry, Gazi University, Ankara, Turkey; Tel: +90 532 551 61 04; E-mail: gulfem@gazi.edu.tr

Cite as: Alatrash L, Ergun G. Comparative Evaluation of Reliability, Quality, Readability, and Originality of AI Chatbot Responses to Patient Questions in Implant Dentistry. Open Dent J, 2026; 20: e187421063068. <http://dx.doi.org/10.2174/01187421063068261001092949>



CrossMark

Received: February 09, 2026

Revised: June 16, 2026

Accepted: September 14, 2026

Published: October 05, 2026



Send Orders for Reprints to
reprints@benthamscience.net

1. INTRODUCTION

Dental implants are a safe and long-lasting treatment option for replacing missing teeth, offering guaranteed functional and aesthetic results. With the increasing prevalence of dental implants in modern dentistry, more and more patients seek information about the implant procedure, its risks, costs, and aftercare before consulting their dentists [1-4].

Lately, there has been a marked increase in dental patients turning to digital sources to better understand their treatment options. This trend includes consulting AI chatbots and searching various social media outlets such as YouTube, Facebook, and Twitter alongside web-based forums for medical advice.

Dental implant treatment has become a premier modality for oral rehabilitation due to its predictable long-

term outcomes in preserving alveolar bone structure and restoring masticatory function [2]. Depending on the specific clinical requirements and surgical timing, various configurations of dental implants are utilized, including endosteal, subperiosteal, and transosteal designs, with endosteal root-form implants being the most frequently deployed in modern clinical practice [4]. Regarding material composition, commercially pure Titanium (cpTi) and Titanium alloys (such as Ti-6Al-4V) remain the primary biomaterials utilized in implant dentistry owing to their exceptional biocompatibility, high mechanical strength, and proven osseointegration rate [1]. Recently, zirconia ceramics have also emerged as a viable metal-free alternative, particularly for patients exhibiting high aesthetic demands or specific metal hypersensitivities [4].

Influenced by developments in the fields of Large Language Modeling (LLM) and Natural Language Processing (NLP), AI chatbots have achieved unprecedented levels of conversational fluency [5]. By leveraging deep learning and artificial neural networks, these tools process vast quantities of heterogeneous language data. Recent developmental cycles have enhanced the models' ability to produce high-fidelity, realistic responses, with latest-generation versions consistently outperforming prior iterations in complex domain-specific tasks [6]. Given their versatility, these AI tools are now integral to a wide array of sectors, including medical diagnostics, educational platforms, and automated technical troubleshooting [7].

The rapid development of artificial intelligence, particularly chatbots based on massive linguistic models, has fundamentally transformed how health information is accessed and used [8, 9]. AI platforms such as ChatGPT, Google Gemini, and Claude are widely used to provide instant explanations on medical and dental topics. ChatGPT, developed by OpenAI, is designed to produce coherent, context-aware responses and is commonly utilized for conversational and educational purposes. Google Gemini, developed by Alphabet Inc., is optimized for interactive dialogue and is recognized for generating structured and informative answers, with an emphasis on accessing recent information [10]. ChatGPT, which was first introduced in November 2022, rapidly gained popularity, amassing 100 million users within three months [11]. Claude, developed by Anthropic, is designed with a focus on safety, clarity, and balanced content generation, and is known for producing nuanced and user-oriented responses [12]. Given the increasing public reliance on these platforms for medical and dental information, it is important to systematically evaluate their performance in delivering accurate, reliable, and patient-appropriate content. These tools offer numerous advantages, including responsiveness, accessibility, and interactive communication. However, concerns remain regarding the reliability, quality, readability, and credibility of AI-generated health information, especially when used by patients without professional medical supervision [5, 8, 9].

Previous studies have assessed the performance of AI-

powered chatbots in several dental specialties, including pediatric dentistry, orthodontics, endodontics, and maxillofacial rehabilitation [13]. Although these studies indicate that AI tools are generally capable of generating helpful responses, they also highlight limitations related to a lack of references, inconsistent accuracy, and complex text, which can hinder patient comprehension. Research specifically evaluating AI-generated information about dental implant treatment remains limited, despite the clinical complexity and patient interest in implant-related therapies [8, 13, 14].

Given the potential consequences of misinformation in implant dentistry, it is essential to comprehensively evaluate the performance of AI-powered chatbots in answering patient inquiries related to implant dentistry. Assessing factors such as information reliability, content quality, readability, and credibility can help determine the suitability of these tools as patient education aids [5, 15, 16].

The aim of this study is to evaluate and compare the reliability, quality, readability, and credibility of responses generated by three commonly used AI-powered chatbots regarding implant dentistry treatment. The findings of this study may contribute to understanding the potential role of AI tools in educating implant dentistry patients, while emphasizing the importance of professional supervision. The null hypothesis was that there would be no difference between the responses generated by different AI-powered chatbots.

2. MATERIALS AND METHODS

This study followed a structured methodological framework. Ethical approval was not required as no human participants, patient data, or animal-derived material were used.

Information regarding dental implant treatment was obtained using three AI-based chatbot models: Google Gemini Flash (<https://gemini.google.com>), ChatGPT Plus (OpenAI; <https://chatgpt.com>), and Claude Sonnet (Anthropic; <https://claude.ai>). The primary objective was to evaluate and compare the reliability, overall quality, readability, and specificity of the responses these models provided to patient-centered questions about dental implant treatment.

A total of 20 patient-oriented questions concerning dental implant therapy were developed by the author based on current implant dentistry guidelines and frequently asked patient concerns. To enhance clinical validity and relevance, the draft questions were reviewed, revised, and formally approved by a faculty member of a prosthodontic graduate program with extensive expertise in implant dentistry. Although no a priori power analysis was performed, this methodology and the targeted sample size are consistent with recent dental and prosthodontic literature benchmarking AI chatbots, in which cohorts of 15 to 20 specific questions are effectively used [13, 17].

The finalized questions were structured to represent a broad range of patient inquiries and were categorized into two domains:

(1) Theoretical questions (fundamental concepts, indications, contraindications, and implant materials) 10 questions;

(2) Clinical questions (treatment procedures, risks, complications, postoperative care, and maintenance) 10 questions.

A. Theoretical (Knowledge-Based) Questions:

(1) What is a dental implant and how does it replace a missing tooth?

(2) How successful are dental implants in the long term?

(3) How long do dental implants usually last?

(4) What materials are dental implants made of?

(5) How long does dental implant treatment take?

(6) What factors affect the success of a dental implant?

(7) What is bone graft (bone powder) and who is it applied to?

(8) Can implants be placed immediately after tooth extraction?

(9) Is there an age limit for getting dental implants?

(10) Are dental implants better than bridges or dentures?

B. Clinical (Patient-Oriented) Questions

(11) Will I feel pain during or after dental implant surgery?

(12) How long does it take to recover after getting a dental implant?

(13) Can my implant be used immediately after it is placed?

(14) What happens if I do not have enough bone for an implant?

(15) Does smoking reduce the success of dental implants?

(16) Can people with diabetes safely receive dental implants?

(17) What are the most common complications of dental implants?

(18) How do I clean and take care of my dental implant?

(19) Can gum inflammation occur around dental implants?

(20) Can dental implants fail, and what are the warning signs?

All questions were submitted to each AI chatbot on the same day (January 20, 2026) using their web-based interfaces. To prevent previous responses from influencing subsequent answers, a new conversation session was created for each question. No guiding prompts or follow-up instructions were provided, and only the first response generated by each chatbot was recorded. The collected responses were saved verbatim in a Microsoft Word document without any modifications for subsequent analysis. A total of 60 responses were obtained from the three AI chatbots, corresponding to 20 questions per model. The collected answers were evaluated in terms of informational quality, reliability, originality, and readability

using multiple validated assessment tools, including the adapted DISCERN instrument, the Global Quality Scale (GQS), the Flesch Reading Ease Score (FRES), the Flesch-Kincaid Reading Grade Level (FKRGL), and iThenticate plagiarism detection software. The responses produced by three AI-powered chatbots were anonymized prior to evaluation and categorized as Model A, Model B, and Model C. Specifically, responses from Gemini were labeled as Model A, responses from Claude as Model B, and responses from ChatGPT as Model C. All responses were then presented to two prosthodontists for evaluation: one a junior prosthodontist and the other a faculty member with over 20 years of clinical and academic experience. The evaluators were unaware of the AI models' identities and independently assessed the accuracy, completeness, and quality of the chatbot responses.

To ensure consistency in scoring, a calibration process was implemented prior to the evaluation. During this process, the assessors reviewed all evaluation criteria and established a shared understanding of the scoring framework. Ten pilot questions that were not included in the main study were rated and subsequently re-evaluated after a 14-day interval, and the results were compared. Inter-rater and intra-rater reliability were examined using the Intraclass Correlation Coefficient (ICC), with values exceeding the commonly accepted threshold for good reliability ($ICC > 0.75$).

The reliability of chatbot-generated responses was assessed using a five-item version of the DISCERN instrument adapted for AI-generated health information. The modified DISCERN is an adapted version of the original DISCERN tool to assess the reliability of information content, tailored to specific research needs [18, 19], as shown in Fig. (1). In the adapted DISCERN test, features were rated on a scale from 1 to 5, with 5 being the highest possible score. The responses were independently assessed by 2 experts in prosthodontics, and a consensus was reached.

Overall informational quality was evaluated using the Global Quality Scale (GQS). This scale assesses the clarity, flow, depth of information, and usefulness of content for patient education, with scores ranging from 1 (very poor quality) to 5 (excellent quality) as shown in Fig. (2). Higher scores indicated more comprehensive, coherent, and clinically valuable responses [20]. The responses were assessed independently by 2 experts in prosthodontics, and a consensus was reached.

Readability of the AI-generated responses was evaluated using the Flesch Reading Ease Score (FRES) and the Flesch-Kincaid Reading Grade Level (FKRGL). Both indices estimate text readability based on sentence length and the average number of syllables per word. The FRES score was calculated using the following formula:

$$FRES = 206.835 - (1.015 \times \text{average words per sentence}) - (84.6 \times \text{average syllables per word})$$
 [21], while the FKRGL score was computed as:

$$FKRGL = (0.39 \times \text{average words per sentence}) + (11.8 \times \text{average syllables per word}) - 15.59$$
 [22].

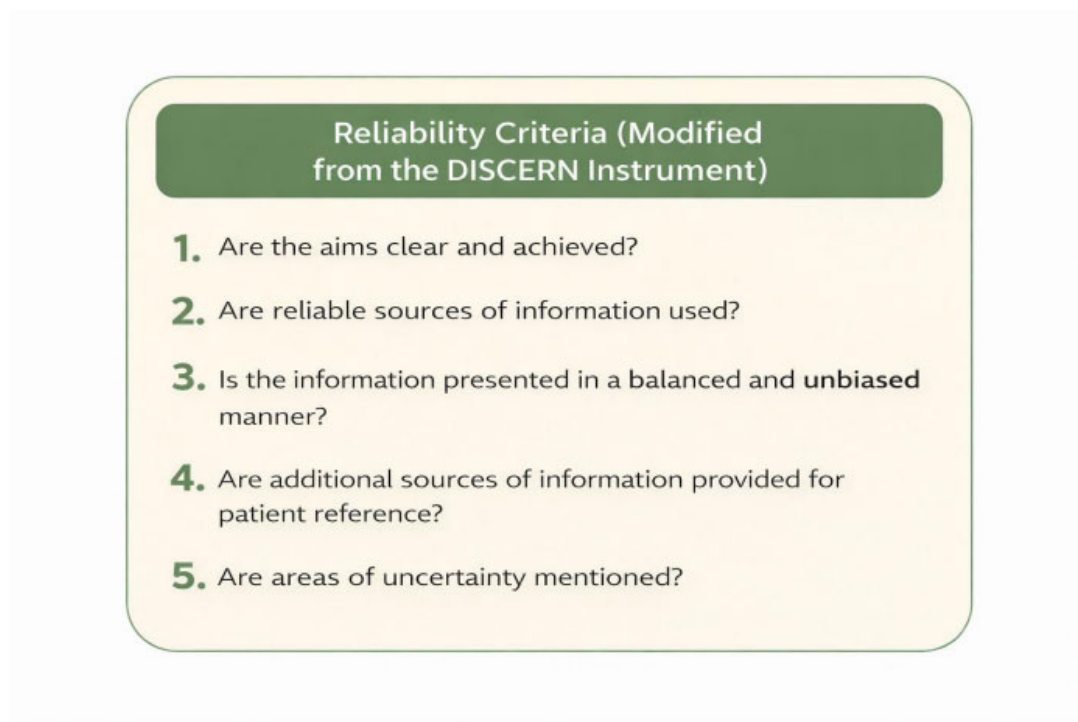


Fig. (1). Reliability criteria adapted from the DISCERN instrument.



Fig. (2). Global quality scale (GQS) used for informational quality assessment.

Higher FRES values indicate easier-to-read and more patient-friendly text, whereas lower scores reflect greater linguistic complexity. The FKRGL index estimates the educational grade level required to comprehend the text, with higher values corresponding to more advanced reading proficiency. These metrics were used to assess the accessibility and understandability of the chatbot-generated responses for patients [19, 21, 22].

The originality of chatbot-generated responses was evaluated using iThenticate plagiarism detection software (<http://www.ithenticate.com>). For each chatbot, all responses were compiled into a single document and analyzed separately. Similarity percentages were recorded and categorized as follows: 0–10% indicating high originality, 11–20% indicating acceptable similarity, 21–40% indicating high similarity, and greater than 40% indicating very high similarity.

2.1. Statistical Analysis

Five outcome measures were included in the statistical analysis: reliability (DISCERN), overall information quality (Global Quality Scale [GQS]), readability (Flesch Reading Ease Score [FRES] and Flesch–Kincaid Reading Grade Level [FKRGL]), and originality (similarity index). Prior to statistical comparisons, the distribution of all variables was assessed to determine the appropriate analytical approach.

Normality was evaluated using the Shapiro–Wilk test and further verified by examining skewness and kurtosis values, as well as Q–Q plots. FRES scores were normally distributed (Shapiro–Wilk $W = 0.971$, $p = 0.164$; skewness = -0.548 ; kurtosis = 0.179), as were FKRGL scores (Shapiro–Wilk $W = 0.983$, $p = 0.550$; skewness = 0.197 ; kurtosis = 0.256). Additionally, both variables satisfied the homogeneity of variance assumption (FRES: Levene's $F = 1.748$, $p = 0.183$; FKRGL: Levene's $F = 0.339$, $p = 0.714$). In contrast, among the ordinal score-type variables, DISCERN, GQS, and similarity index scores did not demonstrate normal distribution characteristics.

For normally distributed variables, comparisons among the three AI-powered chatbots were performed using one-way Analysis of Variance (ANOVA). Results were interpreted based on mean values, and partial eta-squared (η^2) was reported as the measure of effect size. When statistically significant differences were detected, Tukey's post hoc test was applied for pairwise comparisons.

For variables that did not follow a normal distribution (DISCERN, GQS, and similarity), the Kruskal–Wallis test was used to compare distributions among chatbots. Median values were used for interpretation, and epsilon-squared (ϵ^2) was reported as the effect size measure. For the five individual DISCERN items (D1–D5), the family-wise error rate was controlled using the Holm–Bonferroni sequential correction method. When significant differences were identified, the Dwass–Steel–Critchlow–Fligner (DSCF) post hoc test was conducted for multiple comparisons.

All statistical tests were two-tailed, and the level of statistical significance was set at $\alpha = 0.05$. Statistical analyses were performed using IBM SPSS Statistics version

24.0 (IBM Corp., Armonk, NY, USA), Jamovi version 2.4 (The Jamovi Project, Sydney, Australia), Microsoft Excel 2016 (Microsoft Corporation, Redmond, WA, USA), and RStudio version 2023.09.0 (Posit Software, Boston, MA, USA).

3. RESULTS

A total of 60 responses generated by three AI-powered chatbots (Gemini 3, ChatGPT 5.2, and Claude 4.5 Sonnet) were evaluated, corresponding to 20 patient-oriented questions per model. The responses were analyzed in terms of reliability (DISCERN), overall information quality (GQS), readability (FRES and FKRGL), and originality (similarity index).

A statistically significant difference was observed among the AI chatbots for all individual DISCERN items as well as for the total DISCERN reliability score ($p < 0.05$ for all comparisons). Gemini 3 consistently achieved higher median scores across all five DISCERN questions compared with ChatGPT 5.2 and Claude 4.5 Sonnet.

For DISCERN Question 1, a significant difference was found among the chatbots ($\chi^2 = 10.23$; $\epsilon^2 = 0.17$; $p = 0.006$), with Gemini 3 obtaining a higher median score than both ChatGPT 5.2 and Claude 4.5 Sonnet. Similar statistically significant differences were observed for DISCERN Questions 2 and 3 ($p < 0.05$), where Gemini 3 demonstrated higher median scores than Claude 4.5 Sonnet, while ChatGPT 5.2 did not differ significantly from either chatbot.

For DISCERN Questions 4 and 5, the differences among chatbots were highly significant ($p < 0.001$). Gemini 3 achieved significantly higher median scores than both ChatGPT 5.2 and Claude 4.5 Sonnet for these items.

Regarding the overall DISCERN reliability score, a statistically significant difference was found among the AI chatbots ($\chi^2 = 14.45$; $\epsilon^2 = 0.25$; $p < 0.001$). Gemini 3 demonstrated a higher total DISCERN score compared with ChatGPT 5.2 and Claude 4.5 Sonnet, whereas no significant difference was observed between ChatGPT 5.2 and Claude 4.5 Sonnet. As shown in Table 1 and (Fig. 3).

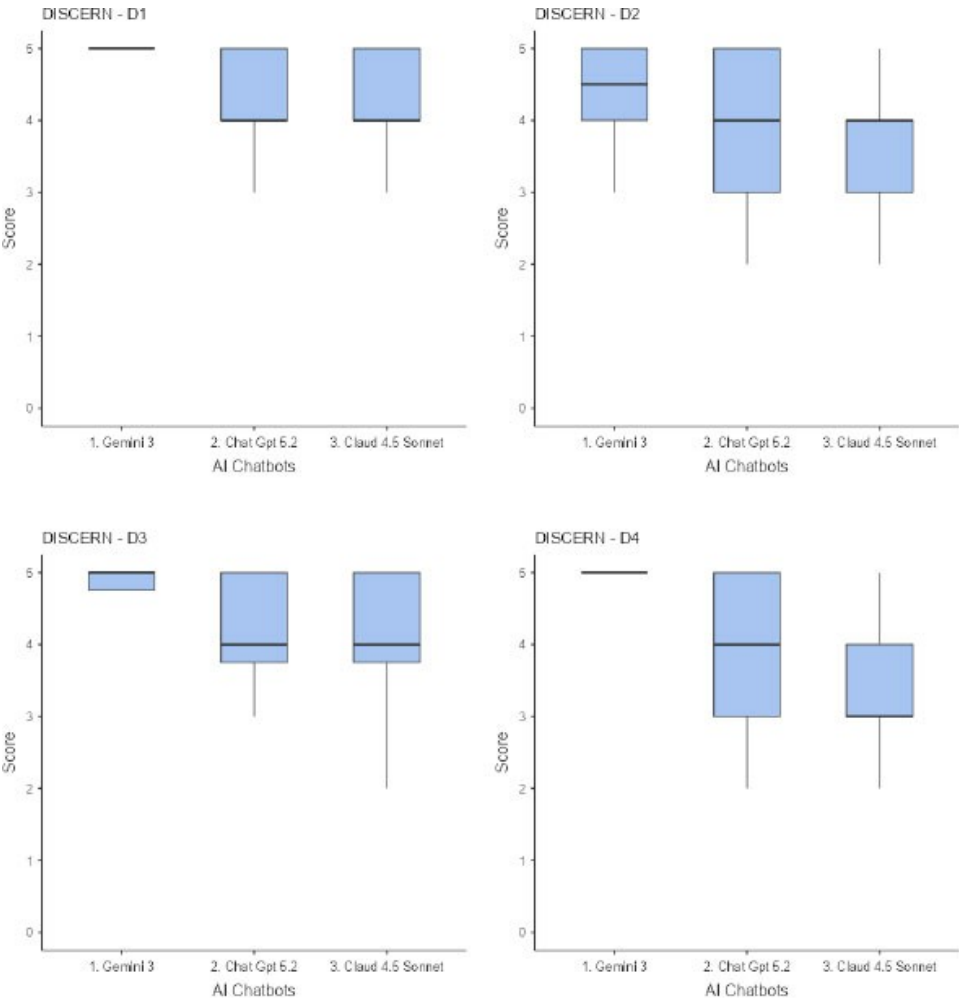
The Global Quality Scale analysis revealed a statistically significant difference among the AI chatbots ($\chi^2 = 6.20$; $\epsilon^2 = 0.11$; $p = 0.045$). Gemini 3 achieved a higher GQS score distribution compared with Claude 4.5 Sonnet. ChatGPT 5.2 did not demonstrate a statistically significant difference when compared with either Gemini 3 or Claude 4.5 Sonnet. The mean and median GQS values of information obtained from different AI chatbots are given in Table 2 and (Fig. 4).

Readability analysis using the Flesch Reading Ease Score (FRES) demonstrated a statistically significant difference among the AI chatbots ($F = 3.94$; $\eta^2 = 0.12$; $p = 0.025$). Gemini 3 achieved higher mean FRES scores than Claude 4.5 Sonnet, indicating relatively easier readability. No statistically significant difference was observed between Gemini 3 and ChatGPT 5.2 or between ChatGPT 5.2 and Claude 4.5 Sonnet, as shown in Table 3 and (Fig. 5).

Table 1. Comparison of reliability measures obtained from different artificial intelligence tools.

DISCERN	Gemini 3			ChatGPT 5.2			Claude 4.5 Sonnet				ϵ^2	p*
	n	Mean ($\pm$ SD)	Median (IQR; R)	n	Mean ($\pm$ SD)	Median (IQR; R)	n	Mean ($\pm$ SD)	Median (IQR; R)			
D1	20	4.75 ($\pm$ 0.55)	^a 5.00 (0; 3-5)	20	4.10 ($\pm$ 0.72)	^b 4.00 (1; 3-5)	20	4.25 ($\pm$ 0.72)	^b 4.00 (1; 3-5)	10.23	0.17	0.018
D2	20	4.40 ($\pm$ 0.68)	^a 4.50 (1; 3-5)	20	3.75 ($\pm$ 1.16)	^{ab} 4.00 (2; 2-5)	20	3.60 ($\pm$ 0.82)	^b 4.00 (1; 2-5)	8.10	0.14	0.022
D3	20	4.70 ($\pm$ 0.57)	^a 5.00 (0.25; 3-5)	20	4.15 ($\pm$ 0.81)	^{ab} 4.00 (1.25; 3-5)	20	3.95 ($\pm$ 0.94)	^b 4.00 (1.25; 2-5)	9.04	0.15	0.022
D4	20	4.80 ($\pm$ 0.41)	^a 5.00 (0; 4-5)	20	3.85 ($\pm$ 0.99)	^b 4.00 (2; 2-5)	20	3.45 ($\pm$ 0.94)	^b 3.00 (1; 2-5)	20.30	0.34	<0.001
D5	20	4.80 ($\pm$ 0.41)	^a 5.00 (0; 4-5)	20	3.90 ($\pm$ 0.97)	^b 4.00 (2; 2-5)	20	3.85 ($\pm$ 0.93)	^b 4.00 (1.25; 2-5)	14.83	0.25	0.002
DISCERN	20	23.45 ($\pm$ 1.70)	^a 23.50 (2; 19-25)	20	19.75 ($\pm$ 4.34)	^b 19.50 (7.25; 13-25)	20	19.10 ($\pm$ 3.81)	^b 20.00 (4.5; 11-25)	14.45	0.25	<0.001

Note:SD: Standard deviation, IQR: Interquartile range (Q_3-Q_1), R: Range.
 χ^2 : Kruskal-Wallis test statistic, ϵ^2 : epsilon-squared effect size, p: P value from the Kruskal-Wallis test.
^{ab}: Different superscript letters indicate statistically significant differences between chatbots (DSCF pairwise, $p < 0.05$).
D1: Clear and achieved aims; D2: Explicit/reliable information sources; D3: Balanced and unbiased presentation; D4: Provision of additional patient reference sources; D5: Mentioning areas of uncertainty.
*For the D1-D5 comparisons, p-values were adjusted for multiple testing using the Holm-Bonferroni sequential correction method.



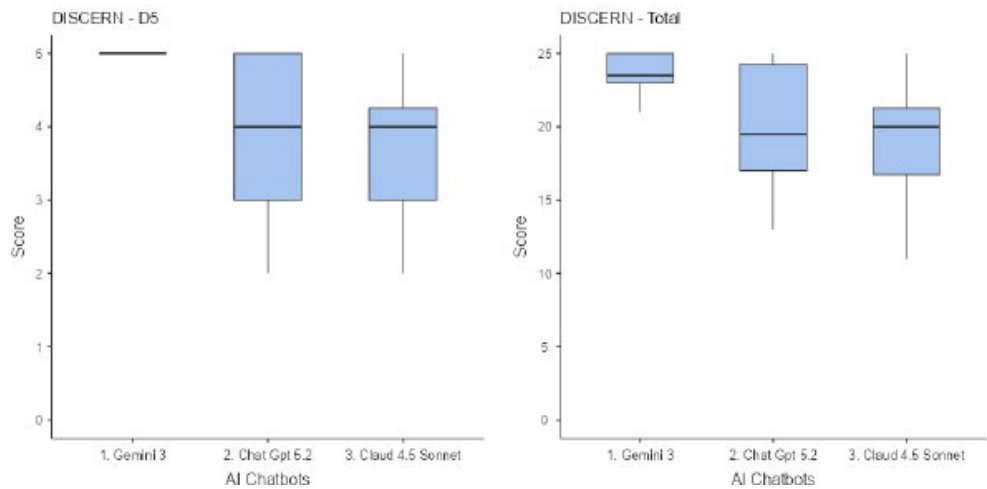


Fig. (3). DISCERN reliability scores across AI chatbots. Boxplots showing the distribution of DISCERN item scores (D1–D5) and total DISCERN scores for Gemini 3 Flash, ChatGPT 5.2, and Claude 4.5 Sonnet.

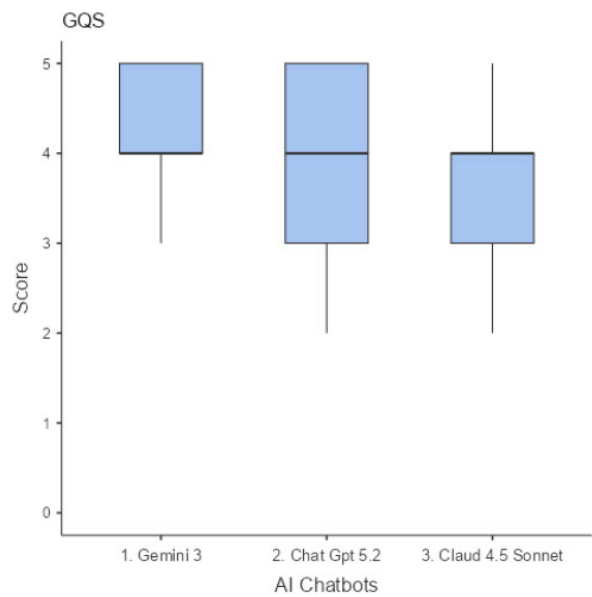


Fig. (4). Global quality scale scores across AI chatbots.

Table 2. Comparison of quality scores obtained from different artificial intelligence tools.

	Gemini 3			ChatGPT 5.2			Claud 4.5 Sonnet				ϵ^2	p
	n	Mean ($\pm$ SD)	Median (IQR; R)	n	Mean ($\pm$ SD)	Median (IQR; R)	n	Mean ($\pm$ SD)	Median (IQR; R)			
GQS	20	4.35 ($\pm$ 0.59)	^a 4.00 (1; 3-5)	20	3.85 ($\pm$ 0.99)	^{ab} 4.00 (2; 2-5)	20	3.65 ($\pm$ 0.93)	^b 4.00 (1; 2-5)	6.20	0.11	0.045

SD: Standard deviation, IQR: Interquartile range (Q_3 - Q_1), R: Range.

χ^2 : Kruskal-Wallis test statistic, ϵ^2 : epsilon-squared effect size, p: P value from the Kruskal-Wallis test.

^{ab}: Different superscript letters indicate statistically significant differences between chatbots (DSCF pairwise, $p < 0.05$).

GQS: Global Quality Score.

Table 3. Comparison of reading features obtained from different artificial intelligence tools.

	Gemini 3			ChatGPT 5.2			Claud 4.5 Sonnet			F	η_p^2	p
	n	Mean ($\pm$ SD)	Median	n	Mean ($\pm$ SD)	Median	n	Mean ($\pm$ SD)	Median			
FRES	20	^a 54.77 ($\pm$ 7.54)	55.45	20	^{ab} 50.41 ($\pm$ 10.69)	48.85	20	^b 46.36 ($\pm$ 9.92)	49.00	3.94	0.12	0.025
FKRGL	20	^{ab} 8.77 ($\pm$ 1.47)	8.70	20	^a 8.56 ($\pm$ 1.66)	8.45	20	^b 9.82 ($\pm$ 1.60)	9.85	3.67	0.11	0.032

SD: Standard deviation, F: ANOVA F statistic, η_p^2 : partial eta-squared effect size, p: ANOVA P value.
^{ab}: Different superscript letters indicate statistically significant differences between chatbots (Tukey pairwise, $p < 0.05$).

Table 4. Comparison of similarity index obtained from different artificial intelligence tools.

	Gemini 3			ChatGPT 5.2			Claud 4.5 Sonnet			χ^2	ϵ^2	p
	n	Mean ($\pm$ SD)	Median (IQR; R)	n	Mean ($\pm$ SD)	Median (IQR; R)	n	Mean ($\pm$ SD)	Median (IQR; R)			
Similarity	20	2.40 ($\pm$ 3.70)	0.00 (3.25; 0-12)	20	4.55 ($\pm$ 7.01)	0.00 (6.5; 0-24)	20	4.45 ($\pm$ 5.25)	4.00 (5.25; 0-18)	1.72	0.03	0.423

SD: Standard deviation, IQR: Interquartile range (Q_3-Q_1), R: Range.
 χ^2 : Kruskal-Wallis test statistic, ϵ^2 : epsilon-squared effect size, p: P value from the Kruskal-Wallis test.
Different superscript letters indicate statistically significant differences between chatbots (DSCF pairwise, $p < 0.05$).

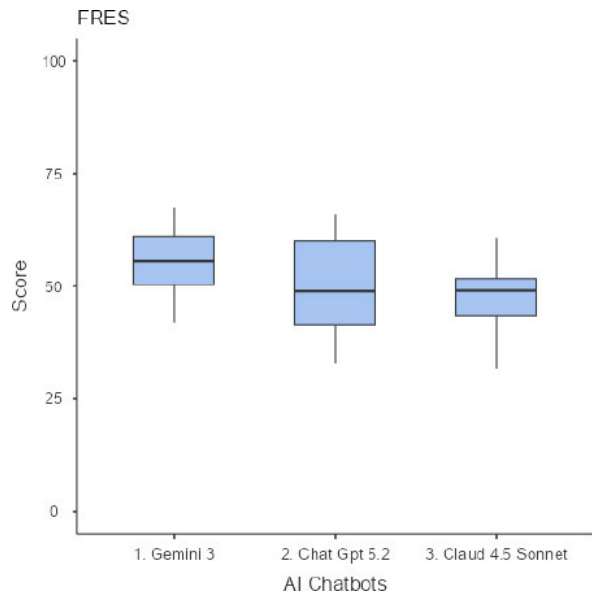


Fig. (5). FRES readability scores of chatbot responses.

For the Flesch-Kincaid Reading Grade Level (FKRGL), a statistically significant difference was also observed among the AI chatbots ($F = 3.67$; $\eta_p^2 = 0.11$; $p = 0.032$). Claude 4.5 Sonnet demonstrated higher mean FKRGL scores compared with ChatGPT 5.2, indicating a higher required reading grade level. Gemini 3 did not differ significantly from the other chatbots. As shown in Table 3 and (Fig. 6).

Overall, the readability scores indicated that chatbot-

generated responses ranged from “difficult” to “fairly difficult” and required an early high school-level reading ability.

The similarity index analysis revealed that the median similarity values for all AI chatbots fell within the “high originality” range. No statistically significant difference was found among the chatbots with respect to similarity scores ($\chi^2 = 1.72$; $\epsilon^2 = 0.03$; $p > 0.05$). The similarity rates (%) of answers provided by AI chatbots are given in Table 4 and (Fig. 7).

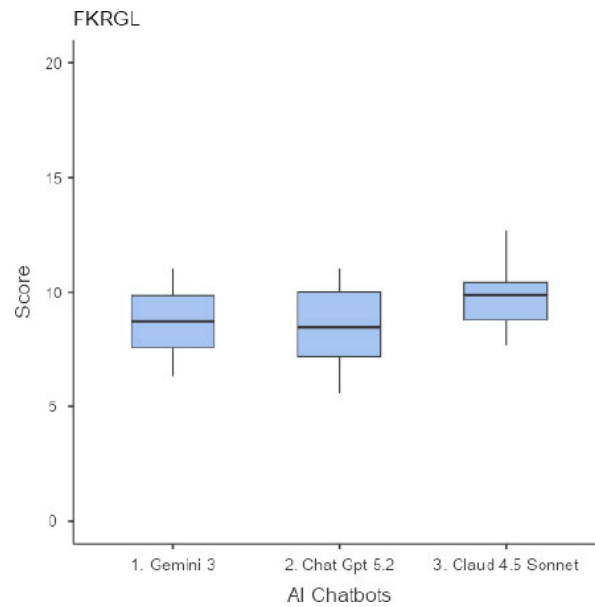


Fig. (6). FKRGL readability levels of chatbot responses.

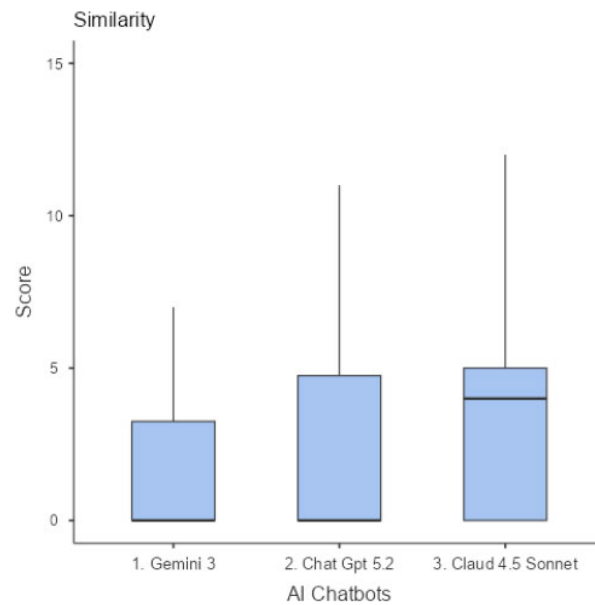


Fig. (7). Similarity scores of chatbot-generated responses.

When theoretical (Questions 1-10) and clinical questions (Questions 11-20) were analyzed separately, statistically significant differences were observed among the AI chatbots for theoretical questions in terms of DISCERN reliability ($\chi^2 = 13.27$; $\varepsilon^2 = 0.46$; $p < 0.001$) and GQS quality scores ($\chi^2 = 9.13$; $\varepsilon^2 = 0.31$; $p = 0.010$). Gemini 3 demonstrated significantly higher reliability and quality scores for theoretical questions compared with ChatGPT 5.2 and Claude 4.5 Sonnet, as shown in Table 5.

For the theoretical questions, mean FRES readability scores ranged from 45.86 ("difficult") to 56.88 ("fairly difficult") across chatbots. No statistically significant difference was found among the AI chatbots in FRES readability scores ($F = 2.69$; $\eta_p^2 = 0.17$; $p > 0.05$). The mean FKRGL scores, another readability metric, ranged from 8.09 to 9.78, indicating readability at the early high school level. No statistically significant difference was found among the AI chatbots in FKRGL scores ($F = 2.84$; $\eta_p^2 = 0.17$; $p > 0.05$).

Table 5. Distribution and comparison of measurements by artificial intelligence tools for question types.

	DISCERN	Gemini 3			ChatGPT 5.2			Claude 4.5 Sonnet			χ^2/F	ϵ^2/η_p^2	p
		n	Mean (±SD)	Median (IQR; R)	n	Mean (±SD)	Median (IQR; R)	n	Mean (±SD)	Median (IQR; R)			
Theoretical Questions	DISCERN	10	23.00 (±1.41)	^a 23.00 (1.5; 21-25)	10	18.90 (±3.93)	^b 18.50 (4; 13-25)	10	17.10 (±3.63)	^b 17.50 (4.5; 11-22)	13.27	0.46	<0.001
	GQS	10	4.20 (±0.42)	^a 4.00 (0; 4-5)	10	3.60 (±0.97)	^{ab} 3.50 (1; 2-5)	10	3.10 (±0.74)	^b 3.00 (0.75; 2-4)	9.13	0.31	0.010
	FRES	10	56.88 (±7.18)	57.70	10	52.78 (±12.01)	56.65	10	45.86 (±12.27)	50.90	2.69	0.17	0.086*
	FKRGL	10	8.27 (±1.40)	8.40	10	8.09 (±1.75)	7.85	10	9.78 (±2.02)	8.95	2.84	0.17	0.076*
	Similarity	10	1.30 (±3.47)	0.00 (0; 0-11)	10	7.00 (±8.72)	4.00 (9.50; 0-24)	10	4.50 (±5.50)	4.50 (5.75; 0-18)	4.35	0.15	0.114
Clinical Questions	DISCERN	10	23.90 (±1.91)	25.00 (1.75; 19-25)	10	20.60 (±4.77)	22.00 (8; 13-25)	10	21.10 (±2.92)	20.50 (3; 16-25)	4.36	0.15	0.113
	GQS	10	4.50 (±0.71)	5.00 (1; 3-5)	10	4.10 (±0.99)	4.50 (2; 3-5)	10	4.20 (±0.79)	4.00 (1; 3-5)	1.07	0.04	0.586
	FRES	10	52.65 (±7.64)	53.25	10	48.05 (±9.19)	45.85	10	46.85 (±7.51)	46.20	1.41	0.09	0.261*
	FKRGL	10	9.27 (±1.42)	9.30	10	9.03 (±1.50)	9.45	10	9.86 (±1.14)	10.20	0.98	0.07	0.388*
	Similarity	10	3.50 (±3.78)	2.50 (4.25; 0-12)	10	2.10 (±3.78)	0.00 (3; 0-11)	10	4.40 (±5.27)	3.50 (5; 0-15)	2.02	0.07	0.365

SD: Standard deviation, IQR: Interquartile range (Q_3-Q_1), R: Range.

χ^2 : Kruskal-Wallis test statistic, F: ANOVA F statistic, ϵ^2 : epsilon-squared effect size, η_p^2 : partial eta-squared effect size.

p: P value from the Kruskal-Wallis test, p*: ANOVA P value.

^{ab}: Different superscript letters indicate statistically significant differences between chatbots (DSCF/Tukey pairwise, $p < 0.05$)

GQS: Global Quality Score.

Median similarity index values across chatbots were distributed within a “high originality” range (0.00–4.50); however, no statistically significant difference was found between chatbots ($\chi^2 = 4.35$; $\epsilon^2 = 0.15$; $p > 0.05$).

For clinical questions, no statistically significant differences were found among the AI chatbots for DISCERN, GQS, readability, or similarity parameters ($p > 0.05$ for all comparisons).

4. DISCUSSION

The increasing use of AI-based chatbots has transformed how patients access health-related information, including in implant dentistry. While these tools provide quick and accessible answers to patient-centered questions, concerns remain regarding the reliability, quality, readability, and clinical relevance of the information they generate [23]. Therefore, this study aimed to compare three commonly used AI-powered chatbots in terms of these parameters using valid assessment tools. The null hypothesis that there would be no difference between responses generated by different AI chatbots was rejected. However, the findings revealed that although the responses were generally original, significant differences in reliability, quality, and readability were observed depending on the chatbot model and question type. These results indicate that AI chatbots can support patient education in implant dentistry but should not be used interchangeably or without professional supervision.

One of the most notable findings of this study is that Gemini 3 consistently achieved higher DISCERN reliability

scores compared to ChatGPT 5.2 and Claude 4.5 Sonnet. Gemini 3 obtained a significantly higher total DISCERN score (median = 23.50) than both ChatGPT 5.2 (median = 19.50) and Claude 4.5 Sonnet (median = 20.00) ($\chi^2 = 14.45$; $p < 0.001$). Additionally, Gemini 3 scored significantly higher across all individuals' DISCERN items ($p < 0.05$). This suggests that Gemini 3 provides more balanced, transparent, and patient-centered information, particularly in terms of presenting treatment benefits, risks, and uncertainties. The superior performance observed on DISCERN items related to clarity of objectives and acknowledgment of treatment limitations may indicate a more cautious and safety-oriented response structure. Similar findings have been reported in recent prosthodontic research, in which different AI chatbots demonstrated varying levels of accuracy in answering patient questions about prosthodontic treatment [14]. However, this contrasts with other studies evaluating multiple large language models in prosthodontics, where ChatGPT demonstrated higher validity in specific domains or no significant superiority was observed among chatbots [24]. These differences highlight that AI chatbot performance can vary depending on the clinical subject, model version, and evaluation criteria used. Overall, these findings emphasize that not all AI models perform equally when evaluated using validated tools, and their outputs are not interchangeable in the context of patient education.

In contrast, ChatGPT 5.2 and Claude 4.5 Sonnet demonstrated lower DISCERN scores, which may reflect a

tendency to provide more generalized explanations with less emphasis on uncertainty, alternative treatment options, or the need for professional clinical judgment. While these models still produced largely accurate and coherent responses, the reduced reliability scores raise concerns regarding their unsupervised use by patients seeking guidance on complex treatments such as dental implants. Given the potential consequences of misinformation or oversimplification in implant dentistry, these differences in reliability are clinically relevant.

Previous investigations have demonstrated that the performance of AI chatbots in prosthodontics remains variable across models and clinical domains. Freire *et al.* [25] evaluated ChatGPT-4 in answering questions related to removable dental prostheses and tooth-supported fixed dental prostheses, reporting limited accuracy and only moderate-to-substantial repeatability. Similarly, Gheisarifar *et al.* [26] found that newer chatbot systems, such as Bing, Gemini, and ChatGPT-4, demonstrated acceptable consistency in responding to prosthodontics-related patient queries, whereas earlier versions, such as ChatGPT-3.5, performed less reliably. These findings suggest that, despite their growing popularity, AI chatbots may still generate incomplete or occasionally inaccurate information, and clinicians should be aware of the potential risks when patients rely on these tools without professional supervision. Concerns about the lack of citations, limited source transparency, and suboptimal readability further reinforce the need for cautious interpretation of chatbot-generated health information. Nevertheless, other studies have also highlighted the promising role of AI chatbots in patient education. For example, Sultanoğlu *et al.* [27] reported highly accurate chatbot responses regarding edentulism treatment, and Esmailpour *et al.* [28] emphasized their potential usefulness for answering common questions about dental prostheses, while noting that further improvements are required before widespread clinical adoption.

Overall, the Global Quality Scale findings indicated that chatbot responses generally ranged from good to excellent, suggesting that current AI models are capable of generating structured and clinically relevant explanations for implant-related patient questions. In the present study, a statistically significant difference was observed among the chatbots in terms of overall informational quality ($\chi^2 = 6.20$; $\epsilon^2 = 0.11$; $p = 0.045$). Gemini 3 achieved higher GQS scores (mean = 4.35 ± 0.59 ; median = 4.00) compared with Claude 4.5 Sonnet (mean = 3.65 ± 0.93 ; median = 4.00), while ChatGPT 5.2 (mean = 3.85 ± 0.99 ; median = 4.00) did not differ significantly from either model. These findings suggest that Gemini 3 generally offered clearer, more structured, and more comprehensive explanations, particularly for theoretical questions. Similar trends have been reported in previous prosthodontic research. For example, Esmailpour *et al.* [28] noted that although most chatbots achieved relatively high GQS scores, certain platforms such as Copilot tended to provide comparatively lower-quality responses regarding dental prostheses. Differences

in chatbot performance may also be influenced by whether free or advanced paid versions are evaluated [20]. In this study, ChatGPT 5.2 was assessed using a paid model, whereas Gemini 3 Flash and Claude 4.5 Sonnet were evaluated in their freely available versions, reflecting real-world patient access to these tools. Nevertheless, the literature remains inconsistent, as some investigations have reported minimal or no significant differences in response quality across chatbot systems [15]. Moreover, Freire *et al.* [25] demonstrated that even advanced versions of ChatGPT may exhibit limited accuracy when addressing professional-level prosthodontic questions, emphasizing that important shortcomings persist despite ongoing technological improvements. Variations in evaluation methods and scoring instruments across studies may partly explain these conflicting results, highlighting the need for standardized assessment frameworks to better determine the clinical usefulness of AI-generated dental information. Finally, the absence of significant differences among chatbots for clinical questions in the present study may indicate that responses to practical, procedure-related inquiries are more standardized and less variable across AI models.

Readability refers to how easily written health information can be understood by the general population [29]. Ensuring that patient-oriented materials are presented in clear and accessible language is essential for improving comprehension and supporting informed decision-making. However, previous research has shown that many online patient education resources are written at a level that exceeds recommended readability standards, potentially limiting their usefulness for individuals with lower health literacy. This issue represents an important concern in public health, as overly complex information may reduce patients' ability to fully understand treatment options and related risks [15, 30].

Readability is a critical factor in determining whether AI-generated health information can be effectively understood by the general patient population. In the present study, statistically significant differences were observed among the chatbots in terms of FRES readability scores ($F = 3.94$; $\eta^2 = 0.12$; $p = 0.025$). Gemini 3 achieved the highest mean FRES score (54.77), indicating relatively easier readability compared with Claude 4.5 Sonnet (46.36), whereas ChatGPT 5.2 showed an intermediate value (50.41). Nevertheless, the overall FRES range across all models corresponded to "fairly difficult" to "difficult" readability, suggesting that implant-related chatbot responses may still be challenging for many patients to interpret without assistance.

Similarly, the Flesch-Kincaid Reading Grade Level results revealed significant variation among the models ($F = 3.67$; $\eta^2 = 0.11$; $p = 0.032$). Claude 4.5 Sonnet produced responses that require a significantly higher educational level (mean FKRL = 9.82) than ChatGPT 5.2 (mean = 8.56), while Gemini 3 showed comparable values (mean = 8.77). These findings indicate that most chatbot-generated implant dentistry responses require an early

high school reading level, which exceeds the commonly recommended sixth-grade level for patient education materials. Therefore, despite acceptable informational quality, linguistic complexity may limit accessibility for individuals with lower health literacy. This result was consistent with previous studies reporting the difficulty of ChatGPT [13, 20].

Several recent studies across different dental and medical specialties have highlighted both the potential benefits and important limitations of AI chatbot-generated patient information. Giannakopoulos *et al.* [31] evaluated multiple chatbot systems, including ChatGPT-3.5, ChatGPT-4, Gemini, and Copilot, in the context of patient inquiries about clear aligner therapy. Their findings indicated that chatbot responses were generally accurate and of fair-to-good quality; however, a major limitation was the lack of cited references or verifiable sources, which may reduce clinical trustworthiness. Importantly, they also emphasized that the readability of chatbot-generated content was consistently low, potentially restricting accessibility for the broader public. In addition, Duran *et al.* [32] assessed ChatGPT-generated information on cleft lip and palate and reported that although ChatGPT-4 produced reliable and good-quality responses, the readability level remained “difficult” according to FRES scores, reinforcing concerns that linguistic complexity continues to be a barrier to effective patient understanding. Collectively, these findings align with the present study, suggesting that while AI chatbots may serve as supportive educational tools, improvements in readability, transparency, and source attribution are essential before widespread unsupervised patient use can be recommended.

The similarity index analysis showed that responses from all three chatbots fell within the “high originality” range, and no statistically significant differences were found among the models. This suggests that the AI chatbots were capable of generating original responses rather than reproducing existing textual content verbatim. While high originality is a desirable feature, it does not necessarily guarantee accuracy or clinical appropriateness. Therefore, originality should be interpreted as a complementary metric rather than a standalone indicator of content quality.

When theoretical and clinical questions were analyzed separately, Gemini 3 demonstrated a clear advantage in reliability and quality scores for theoretical questions, whereas no significant differences were found among chatbots for clinical questions. This finding suggests that AI models may be more effective at delivering structured foundational knowledge than at addressing nuanced, patient-specific clinical scenarios. Consequently, while AI-powered chatbots may serve as useful adjuncts for general patient education in implant dentistry, their role in clinical decision-making should remain limited.

Furthermore, the lack of significant differences in clinical questions may be due to the nature of these questions, which primarily cover general, highly standardized post-operative risks and care guidelines.

Since this information is widely uniform across online sources, all chatbots generated similarly high-quality answers, which does not necessarily mean they possess equal clinical reasoning capabilities.

Overall, this study demonstrates that AI-powered chatbots can support patient education in implant dentistry; however, reliability, quality, and readability vary among models, highlighting the need for expert supervision. Limitations of this study include the evaluation of a limited number of chatbot platforms, the use of both paid and free versions, and the analysis of only initial responses to a standardized set of questions. Because AI chatbot systems are continuously updated, the responses reflect only the versions available at the time of data collection. This dynamic evolution underscores the importance of continuously benchmarking large language models against the most recent literature up to 2026, which demonstrates how rapid software and dataset updates consistently alter the landscape of readability, accuracy, and reliability in online public information sources in implant dentistry [33]. Future research should incorporate a wider range of AI models, multilingual assessments, and standardized evaluation frameworks, with a particular focus on improving readability, transparency, and reference documentation to ensure safe clinical use.

CONCLUSION

Based on the findings of this study, the following important points can be concluded:

- AI-powered chatbots are generally capable of generating high-quality and reliable responses to patient questions related to dental implants; however, their performance is not uniform across models.
- The Gemini 3 model features a more patient-centered and transparent information architecture, achieving significantly higher reliability (DISCERN) and quality (QOS) scores compared to other chatbots.
- Readability analyses revealed that the chatbot-generated responses often require a secondary school-level reading comprehension level, exceeding the recommended standards for patient education materials, which may limit accessibility for individuals with limited health knowledge.
- While similarity scores demonstrated high reliability, the significant variation in reliability, quality, and readability underscores the need to avoid using AI-powered chatbot outputs interchangeably or as standalone clinical guidance resources.
- Therefore, while these tools can contribute to patient education, professional supervision is essential to ensure accurate understanding and sound decision-making in implant dentistry.
- Future research should focus on improving the readability, transparency of references, and clinical reliability of AI-generated dental health information through standardized assessment frameworks.

AUTHORS' CONTRIBUTIONS

The authors confirm their contributions to the paper as follows: L.A.: Investigation, data curation, formal analysis, visualization, writing original draft; G.E.: Conceptualization, methodology, supervision, validation, writing review and editing.

LIST OF ABBREVIATIONS

GQS	=	Global Quality Scale
LLM	=	Large Language Modeling
FRES	=	Flesch Reading Ease Score
NLP	=	Natural Language Processing
ICC	=	Intraclass Correlation Coefficient

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

HUMAN AND ANIMAL RIGHTS

Not applicable.

CONSENT FOR PUBLICATION

Not applicable. This study did not include identifiable personal data or human participants requiring consent for publication.

AVAILABILITY OF DATA AND MATERIALS

The datasets generated and/or analyzed during the current study are available from the corresponding author upon reasonable request.

FUNDING

None.

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

- [1] Shenoy VK, Rodrigues SJ, Mukherjee S, Ramanathan A. Dental implant failure and retrieval techniques; a scoping review. *Front Oral Health* 2025; 6: 1667808. <http://dx.doi.org/10.3389/froh.2025.1667808> PMID: 41384141
- [2] Krupińska AM, Bogucki Z. Evaluation of patients' awareness and knowledge regarding dental implants among patients of the Department of Prosthetic Dentistry at Wrocław Medical University in Poland. *Adv Clin Exp Med* 2023; 32(11): 1319-25. <http://dx.doi.org/10.17219/acem/174032> PMID: 37999636
- [3] Rashidi Maybodi F, Shayeghi M. Patients' knowledge and attitude regarding Dental Implants. *J Community Health Res* 2022; 11(2): 99-105. <http://dx.doi.org/10.18502/jchr.v11i2.9998>
- [4] Morton D, Wismeijer D, Chen S, *et al.* Group 5 ITI Consensus Report: Implant placement and loading protocols. *Clin Oral Implants Res* 2023; 34(S26): 349-56. <http://dx.doi.org/10.1111/clr.14137> PMID: 37750529
- [5] Tuzlali M, Baki N, Aral K, Aral CA, Bahçe E. Evaluating the performance of AI chatbots in responding to dental implant FAQs: A comparative study. *BMC Oral Health* 2025; 25(1): 1548. <http://dx.doi.org/10.1186/s12903-025-06863-w> PMID: 41063105
- [6] Deiana G, Dettori M, Arghittu A, Azara A, Gabutti G, Castiglia P. Artificial intelligence and public health: Evaluating ChatGPT responses to vaccination myths and misconceptions. *Vaccines* 2023; 11(7): 1217. <http://dx.doi.org/10.3390/vaccines11071217> PMID: 37515033
- [7] Wah JNK. Revolutionizing e-health: The transformative role of AI-powered hybrid chatbots in healthcare solutions. *Front Public Health* 2025; 13: 1530799. <http://dx.doi.org/10.3389/fpubh.2025.1530799> PMID: 40017541
- [8] Metin U, Goymen M. Information from digital and human sources: A comparison of chatbot and clinician responses to orthodontic questions. *Am J Orthod Dentofacial Orthop* 2025; 168(3): 348-57. <http://dx.doi.org/10.1016/j.ajodo.2025.04.008> PMID: 40327024
- [9] Patel K, Radcliffe R. Evaluating the readability and quality of bladder cancer information from AI chatbots: A comparative study between ChatGPT, Google Gemini, Grok, Claude and DeepSeek. *J Clin Med* 2025; 14(21): 7804. <http://dx.doi.org/10.3390/jcm14217804> PMID: 41227199
- [10] Rahsepar AA, Tavakoli N, Kim GHJ, *et al.* How AI responds to common lung cancer questions: ChatGPT versus Google Bard. *Radiology* 2023; 307(5): e230922. <http://dx.doi.org/10.1148/radiol.230922> PMID: 37310252
- [11] Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese Medical Licensing Examination: Comparison study. *JMIR Med Educ* 2023; 9: e48002. <http://dx.doi.org/10.2196/48002> PMID: 37384388
- [12] Aydin S, Karabacak M, Vlachos V, Margetis K. Large language models in patient education: A scoping review of applications in medicine. *Front Med* 2024; 11: 1477898. <http://dx.doi.org/10.3389/fmed.2024.1477898> PMID: 39534227
- [13] Özyemişçi N, Bal BT, Güngör MB, Öztürk EK, Canvar A, Nemli SK. Evaluation of information provided by artificial intelligence chatbots on extraoral maxillofacial prostheses. *J Prosthet Dent* 2025; 134(6): 2623.e1-8. <http://dx.doi.org/10.1016/j.prosdent.2025.08.028> PMID: 40925817
- [14] Özcivelek T, Özcan B. Comparative evaluation of responses from DeepSeek-R1, ChatGPT-o1, ChatGPT-4, and dental GPT chatbots to patient inquiries about dental and maxillofacial prostheses. *BMC Oral Health* 2025; 25(1): 871. <http://dx.doi.org/10.1186/s12903-025-06267-w> PMID: 40450291
- [15] Ozduran E, Akkoc I, Büyükçoban S, Erkin Y, Hancı V. Readability, reliability and quality of responses generated by ChatGPT, gemini, and perplexity for the most frequently asked questions about pain. *Medicine* 2025; 104(11): e41780. <http://dx.doi.org/10.1097/MD.00000000000041780> PMID: 40101096
- [16] Ozduran E, Hancı V, Erkin Y, Özbek İC, Abdulkirimov V. Assessing the readability, quality and reliability of responses produced by ChatGPT, Gemini, and Perplexity regarding most frequently asked keywords about low back pain. *PeerJ* 2025; 13: e18847. <http://dx.doi.org/10.7717/peerj.18847> PMID: 39866564
- [17] Aydin Varol E, Ozturk Z, Bal C, Karamuftuoglu N. Comparison of two different artificial intelligence chatbots that provide information to patients and parents about primary tooth pulpotomy treatments. *BMC Oral Health* 2025; 26(1): 685. <http://dx.doi.org/10.1186/s12903-025-07449-2> PMID: 40399852
- [18] Charnock D, Shepperd S, Needham G, Gann R. DISCERN: An instrument for judging the quality of written consumer health information on treatment choices. *J Epidemiol Community Health* 1999; 53(2): 105-11. <http://dx.doi.org/10.1136/jech.53.2.105> PMID: 10396471
- [19] Shan Y, Xing Z, Dong Z, Ji M, Wang D, Cao X. Translating and adapting the DISCERN instrument into a simplified Chinese version and validating its reliability: Development and usability study. *J Med Internet Res* 2023; 25: e40733. <http://dx.doi.org/10.2196/40733> PMID: 36729573

- [20] Öztürk Z, Bal C, Çelikkaya BN. Evaluation of information provided by ChatGPT versions on traumatic dental injuries for dental students and professionals. *Dent Traumatol* 2025; 41(4): 427-36. <http://dx.doi.org/10.1111/edt.13042> PMID: 39853713
- [21] Bothun LS, Feeder SE, Poland GA. Readability of participant informed consent forms and informational documents. *Mayo Clin Proc* 2021; 96(8): 2095-101. <http://dx.doi.org/10.1016/j.mayocp.2021.05.025> PMID: 34226027
- [22] Zai AH, Faro JM, Allison J. Unveiling readability challenges: An extensive analysis of consent document accessibility in clinical trials. *J Clin Transl Sci* 2024; 8(1): e125. <http://dx.doi.org/10.1017/cts.2024.595> PMID: 39345692
- [23] Wong C, Harrison C, Britt H, Henderson J. Patient use of the internet for health information. *Aust Fam Physician* 2014; 43(12): 875-7. PMID: 25705739
- [24] 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. <http://dx.doi.org/10.1111/iej.14014> PMID: 38117284
- [25] Freire Y, Santamaría Laorden A, Orejas Pérez J, Gómez Sánchez M, Díaz-Flores García V, Suárez A. ChatGPT performance in prosthodontics: Assessment of accuracy and repeatability in answer generation. *J Prosthet Dent* 2024; 131(4): 659.e1-6. <http://dx.doi.org/10.1016/j.prosdent.2024.01.018> PMID: 38310063
- [26] Gheisarifar M, Shembesh M, Koseoglu M, *et al.* Evaluating the validity and consistency of artificial intelligence chatbots in responding to patients' frequently asked questions in prosthodontics. *J Prosthet Dent* 2025; 134(1): 199-206. <http://dx.doi.org/10.1016/j.prosdent.2025.03.009> PMID: 40199631
- [27] Güzelce Sultanoğlu E. Can natural language processing (NLP) provide consultancy to patients about edentulism teeth treatment? *Cureus* 2024; 16(10): e70945. <http://dx.doi.org/10.7759/cureus.70945> PMID: 39502997
- [28] Esmailpour H, Rasaie V, Babae Hemmati Y, Falahchai M. Performance of artificial intelligence chatbots in responding to the frequently asked questions of patients regarding dental prostheses. *BMC Oral Health* 2025; 25(1): 574. <http://dx.doi.org/10.1186/s12903-025-05965-9> PMID: 40234820
- [29] Avra TD, Le M, Hernandez S, Thure K, Ulloa JG. Readability assessment of online peripheral artery disease education materials. *J Vasc Surg* 2022; 76(6): 1728-32. <http://dx.doi.org/10.1016/j.jvs.2022.07.022> PMID: 35931399
- [30] Erkin Y, Hanci V, Ozduran E. Evaluating the readability, quality and reliability of online patient education materials on transcutaneous electrical nerve stimulation (TENS). *Medicine* 2023; 102(16): e33529. <http://dx.doi.org/10.1097/MD.00000000000033529> PMID: 37083809
- [31] Giannakopoulos K, Kavadella A, Aaqel Salim A, Stamatopoulos V, Kaklamanos EG. Evaluation of the performance of generative AI large language models ChatGPT, Google Bard, and Microsoft Bing chat in supporting evidence-based dentistry: Comparative mixed methods study. *J Med Internet Res* 2023; 25: e51580. <http://dx.doi.org/10.2196/51580> PMID: 38009003
- [32] Duran GS, Yurdakurban E, Topsakal KG. The quality of CLP-related information for patients provided by ChatGPT. *Cleft Palate Craniofac J* 2025; 62(4): 588-95. <http://dx.doi.org/10.1177/10556656231222387> PMID: 38128909
- [33] Yagci F, Eraslan R, Albayrak H, İpekten F. Accuracy and reliability of artificial intelligence chatbots as public information sources in implant dentistry. *Int J Oral Maxillofac Implants* 2025; 41(2): 1-23. <http://dx.doi.org/10.11607/jomi.11280> PMID: 40560758

DISCLAIMER: Please note that this article is currently in the 'Early view' stage and is not the final 'Version of Record'. While it has been accepted, copy-edited, and formatted, however, it is still undergoing proofreading and corrections by the authors. Therefore, the text may still change before the final publication. Although 'Early view' may not have all bibliographic details available, the DOI and the year of online publication can still be used to cite them. The article title, DOI, publication year, and author(s) should all be included in the citation format. Once the 'Final Version of Record' becomes available the 'Early view' will be replaced by that.