

Comparison of Chatbot Responses to Frequently Asked Patient Questions in Periodontology

Mert KELEŞ^{1*}  Elif ÖNCÜ² 

¹ Specialist Dentist, Department of Periodontology, Dental Clinic Academy Oral and Dental Health Clinic, Ankara, Türkiye, mertkeles05@gmail.com

² Assoc. Prof., Department of Periodontology, Panorama Oral and Dental Health Clinic, Ankara, Türkiye, oncu.elif@hotmail.com

Article Info

Article History

Received: 28.01.2026

Accepted: 02.07.2026

Published: 31.08.2026

Keywords:

Artificial Intelligence,
Health Communication,
Patient Education,
Periodontology,
Readability.

ABSTRACT

Aim: This study compared the accuracy, quality, understandability, actionability, and readability of responses generated by ChatGPT-4o, Gemini-2.5 Flash, and DeepSeek-V3 to common patient questions related to periodontal diseases.

Material and Methods: Twenty open-ended patient-based questions were submitted to each chatbot. Two periodontologists independently evaluated the responses for scientific accuracy, overall quality (GQS), understandability, and actionability (PEMAT-P), while readability was calculated using the Ateşman Readability Index. Inter-rater reliability was good to excellent for the expert-rated domains.

Results: No significant differences were found among chatbots in accuracy or overall quality ($p > 0.05$). Significant differences were observed in understandability, actionability, and readability ($p < 0.001$). ChatGPT-4o and DeepSeek-V3 showed comparable understandability and outperformed Gemini-2.5 Flash, whereas ChatGPT-4o and Gemini-2.5 Flash showed comparable actionability and outperformed DeepSeek-V3. For readability, ChatGPT-4o achieved the highest scores, followed by DeepSeek-V3 and Gemini-2.5 Flash.

Conclusion: Although all chatbots produced accurate responses, their performance varied across communication domains. ChatGPT-4o showed the most favorable overall profile, particularly in readability and patient-oriented clarity; however, no single model should be considered uniformly superior across all criteria.

Periodontolojide Sık Sorulan Hasta Sorularına Chatbot Yanıtlarının Karşılaştırılması

Makale Bilgisi

Makale Geçmişi

Geliş Tarihi: 28.01.2026

Kabul Tarihi: 02.07.2026

Yayın Tarihi: 31.08.2026

Anahtar Kelimeler:

Hasta Eğitimi,
Okunabilirlik,
Periodontoloji,
Sağlık İletişimi,
Yapay Zeka.

ÖZET

Amaç: Bu çalışmada, ChatGPT-4o, Gemini-2.5 Flash ve DeepSeek-V3 tarafından periodontal hastalıklarla ilgili yaygın hasta sorularına verilen yanıtların doğruluğu, kalitesi, anlaşılabilirliği, eyleme dönüştürülebilirliği ve okunabilirliği karşılaştırılmıştır.

Gereç ve Yöntemler: Hasta ifadelerine dayanan yirmi açık uçlu soru her bir sohbet robotuna yöneltilmiştir. İki periodontolog, yanıtları bilimsel doğruluk, genel kalite (GQS), anlaşılabilirlik ve eyleme dönüştürülebilirlik (PEMAT-P) açısından bağımsız olarak değerlendirmiş; okunabilirlik ise Ateşman Okunabilirlik İndeksi kullanılarak hesaplanmıştır. Uzman değerlendirmesine dayalı alanlarda değerlendiriciler arası güvenilirlik iyi ile mükemmel düzeydeydi.

Bulgular: Sohbet robotları arasında doğruluk veya genel kalite açısından anlamlı fark bulunmamıştır ($p > 0.05$). Anlaşılabilirlik, eyleme dönüştürülebilirlik ve okunabilirlik açısından ise anlamlı farklılıklar gözlenmiştir ($p < 0.001$). ChatGPT-4o ve DeepSeek-V3 anlaşılabilirlik açısından benzer skorlar göstermiş ve Gemini-2.5 Flash'tan daha yüksek performans sergilemiştir. Eyleme dönüştürülebilirlik açısından ChatGPT-4o ve Gemini-2.5 Flash benzer skorlar göstermiş ve DeepSeek-V3'ten daha yüksek performans göstermiştir. Okunabilirlik açısından ChatGPT-4o en yüksek skorları elde etmiş, bunu DeepSeek-V3 ve Gemini-2.5 Flash izlemiştir.

Sonuç: Tüm sohbet robotları doğru yanıtlar üretmesine rağmen, performansları iletişim alanlarına göre farklılık göstermiştir. ChatGPT-4o özellikle okunabilirlik ve hasta odaklı açıklık açısından en olumlu genel profili göstermiştir; ancak hiçbir model tüm kriterlerde mutlak olarak üstün kabul edilmemelidir.

To cite this article: Keleş M, & Öncü E. Comparison of Chatbot Responses to Frequently Asked Patient Questions in Periodontology. NEU Dent J. 2026;8:271-82. <https://doi.org/10.51122/neudentj.2026.201>

***Corresponding Author:** Mert KELEŞ, mertkeles05@gmail.com



This article is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0)

INTRODUCTION

Periodontal diseases are characterized by destruction of the gums and supporting tissues and are one of the most common causes of tooth loss in adults.¹ It is known that early diagnosis and proper patient education regarding these diseases are critically important not only for oral health but also for general health.²

The success of periodontal diagnosis and treatment depends not only on professional clinical interventions but also on the patient's accurate understanding of the disease process, adherence to recommended care protocols, and long-term motivation.³ Therefore, access to reliable information directly affects both individual treatment prognosis and oral and dental health indicators at the community level.⁴

With the rise of digital health literacy today, patients are seeking quick answers to questions they have about topics such as periodontal diseases and their treatments, and they often turn to online resources.⁵ However, the scientific accuracy, reliability, and understandability of the information found on these platforms is often debatable.^{6, 7} Incorrect or incomplete information can lead to increased anxiety about treatment, delayed treatment, and inadequate oral hygiene practices.⁸

Recently, artificial intelligence (AI)-powered chatbots have emerged as innovative tools that facilitate healthcare communication. Large language models (LLMs) such as ChatGPT, Gemini, and DeepSeek can generate contextually appropriate, personalized responses to user questions thanks to their advanced natural language processing capabilities.^{9,10} Although there are studies examining chatbot performance in dentistry on various topics, there is a limited number of multidimensional studies in the literature that comprehensively evaluate periodontology-specific patient questions in terms of scientific

accuracy, quality (Global Quality Scale (GQS)), understandability (Patient Education Materials Assessment Tool for Printable Materials (PEMAT-P)), actionability (PEMAT-P), and readability (Ateşman Index).^{11,12} Despite increasing evidence in other areas of dentistry, to our knowledge, responses provided by LLM-supported chatbots to patient questions specific to periodontology have not been comprehensively compared in Turkish to date.

Therefore, the aim of our study is to evaluate the responses provided by three different artificial intelligence-supported chatbots (ChatGPT-4o, Gemini-2.5 Flash, and DeepSeek-V3) to questions frequently asked by patients in the field of periodontology and assess them comparatively in terms of scientific accuracy, quality (GQS), understandability (PEMAT-P), actionability (PEMAT-P), and readability (Ateşman Index). Additionally, the study explored the expert-assessed understandability and actionability of chatbot-generated responses as indicators of their potential usefulness for patient-oriented communication.

The hypotheses tested in this context are as follows:

H₀ (Null hypothesis): There is no statistically significant difference between chatbots in terms of the specified criteria.

H₁ (Alternative hypothesis): There is a statistically significant difference between chatbot performances in terms of accuracy, quality, understandability, actionability, and readability.

MATERIALS and METHODS

This study was approved by the University Scientific Research Ethics Committee (Approval No: 2025/209) and was conducted in accordance with the principles of the Declaration of Helsinki. Informed consent was not required as the study did not involve

patient intervention or identifiable personal data. Twenty original, open-ended questions reflecting concerns frequently expressed by patients in the field of periodontology were developed for use in the study. The initial question pool was generated based on the authors' clinical experience with common patient inquiries and patient discussions on digital platforms such as Reddit. The questions were grouped into four clinically relevant topic

areas: periodontal diagnosis and health, periodontal treatment, halitosis, and dental implants. The final set of 20 questions was reviewed by two periodontologists for clinical relevance, clarity, and suitability for patient-oriented chatbot evaluation. No formal psychometric validation was performed, as the questions were intended to reflect common real-world patient expressions rather than constitute a validated questionnaire (Table 1).

Table 1: Patient questions addressed to chatbots categorized by periodontal topics.

Periodontal Diagnosis and Health	
1	My gums bleed easily when brushing. Is this a sign of a serious disease?
2	My teeth feel loose. Is that normal?
3	If gum disease runs in my family, am I likely to have it too?
4	Does gum disease cause pain, or can it progress without noticeable symptoms?
5	Can gum disease progress to tooth loss if left untreated?
Periodontal Treatment	
6	What is the purpose of dental scaling?
7	People say scaling scratches the teeth-is that true or just a misconception?
8	I had sensitivity for a few days after scaling. Is that normal?
9	Does root planing (curettage) cause pain, and is anesthesia sufficient for comfort?
10	After gum treatment, what daily habits should I pay special attention to so the disease doesn't return?
Halitosis (Bad Breath)	
11	My mouth smells constantly. Could my gums be the cause?
12	What can I do to get rid of bad breath?
13	How can I tell if my bad breath comes from my teeth or my stomach?
14	Do habits like smoking and drinking coffee worsen bad breath, and would quitting help?
15	I have bad breath-which type of doctor should I see?
Dental Implants	
16	Do dental implants last a lifetime, or do they need to be replaced after a certain period?
17	I'm considering dental implants but I'm scared-will I feel pain during or after the procedure?
18	How long does the full healing process take after implant treatment?
19	How can I care for my implants to extend their lifespan?
20	How can I tell if there's gum inflammation around my implant?

In this study, three different large language model (LLM)-supported chatbots, which are among the most advanced and capable artificial intelligence systems in their fields as of the date of publication, were selected. These models are ChatGPT-4o, developed by OpenAI; Gemini-2.5 Flash, offered by Google; and DeepSeek-V3, an open-source model. The study used the most recent versions of these chatbots that were most easily and freely accessible to patients. This approach aims to accurately replicate the scenario a patient would encounter when searching for health information in the real world and to increase the generalizability of the results to patients' daily practices.

The 20 original questions prepared for the study were sent to an independent third party. All chatbot interactions were performed on 22 November 2025, after ethics committee approval and before manuscript submission. To reflect the ordinary user experience, the freely accessible versions of ChatGPT-4o, Gemini-2.5 Flash, and DeepSeek-V3 available at the time of data collection were used. Each question was entered in Turkish in the same format as presented in Table 1, without any additional prompt, role assignment, clarification, or follow-up instruction. For each chatbot and each question, a separate fresh chat session was initiated to minimize the influence of previous conversational context. No additional account-level custom instructions, saved role settings, or medical role prompts were intentionally used during the chatbot interactions. The first complete response generated by each chatbot was recorded without editing. After collecting the initial, unedited raw responses from each chatbot, this person coded them as "Chatbot A," "Chatbot B," and "Chatbot C," concealing which chatbot each response belonged to. The prepared questions and chatbot responses are included in Supplementary Material 1.

Evaluation of Responses

The 60 responses obtained (20 questions x 3 chatbots) were independently evaluated by two expert periodontists with at least 3 years of clinical experience in periodontology. The assessment was performed under blinded conditions, with the evaluators unaware of the identity of the chatbot responsible for each response. Each expert scored the responses based on four expert-rated criteria: accuracy, GQS, PEMAT-P understandability, and PEMAT-P actionability. Readability was assessed separately using the Ateşman Readability Index, which is a formula-based objective metric for Turkish texts and does not require subjective expert scoring.

Accuracy: The suitability of responses to scientific facts, their timeliness, and whether they contain misleading information were evaluated using a 5-point Likert scale (1: Completely wrong, 5: Completely correct and complete).¹³

Quality (GQS): The completeness, usefulness, and scope of responses were rated on a 5-point Likert scale (1: Very poor, 5: Very good, complete, and useful).¹⁴

Understandability (PEMAT-P): The 17-item "Understandability" subscale of the PEMAT-P scale used in the evaluation of patient education materials was used. The suitability of the responses to these items was scored as "Yes=1" and "No=0" and the results were converted to a percentage score.¹⁵

Actionability (PEMAT-P): Using the 7-item "Actionability" subscale of the PEMAT-P scale, responses were evaluated based on how much they encouraged the patient to take concrete steps. Scoring was done as "Yes=1", "No=0" and the results were converted to a percentage score.¹⁵

Readability: To measure the linguistic difficulty of the responses, the Ateşman Readability Index, which is standardized for

Turkish texts, was used. Using this formula, a score between 0 and 100 was calculated for each response (a higher score indicates easier readability), and these scores were classified between “Very Easy” and “Very Difficult”.¹⁶ Because this index is calculated using a standardized formula, readability was treated as an objective text-based measure rather than an expert-rated variable.

Statistical Analysis

All statistical procedures were conducted with SPSS software (IBM SPSS Statistics for Windows, v26.0; IBM Corp., Armonk, NY, USA). Inter-rater agreement between the two periodontologists was quantified using the intraclass correlation coefficient (ICC) based on a two-way random-effects model with absolute agreement for the expert-rated domains, including accuracy, GQS, PEMAT-P understandability, and PEMAT-P actionability. Readability was calculated separately using the Ateşman Readability Index and was therefore treated as a formula-based objective text metric rather than an expert-rated variable. Accordingly, readability was not included in the inter-rater reliability analysis. For expert-rated variables, the median scores of the two evaluators were calculated and used for subsequent analyses.

Distributional assumptions were assessed with the Shapiro–Wilk test. As the variables did not demonstrate normal distribution, nonparametric methods were selected. Group comparisons among the three chatbot models

(ChatGPT-4o, Gemini-2.5 Flash, and DeepSeek-V3) were performed using the Kruskal–Wallis test. When statistically significant differences were detected, pairwise comparisons were conducted using Dunn’s procedure with Bonferroni adjustment.

Data are reported as median and interquartile range (IQR). The threshold for statistical significance was defined as $p < 0.05$. Effect sizes (η^2) were interpreted according to conventional benchmarks: 0.01 (small), 0.06 (moderate), and ≥ 0.14 (large). Prior to analysis, the dataset was screened for assumption violations, missing values, and potential outliers; no extreme observations were detected.

RESULTS

A total of 20 patient-based questions regarding periodontal diseases were evaluated by two periodontologists across three chatbot platforms (ChatGPT-4o, Gemini-2.5 Flash, and DeepSeek-V3).

The inter-rater reliability analysis demonstrated good to excellent agreement between evaluators for the expert-rated domains (ICC range: 0.79–0.91) (Table 2). ICC values were reported to document the consistency of the blinded expert-based evaluation process across the predefined scoring domains. Readability was not included in the inter-rater reliability analysis because the Ateşman Readability Index was calculated using a standardized formula rather than expert scoring.

Table 2: Inter-rater reliability between periodontologists.

Measure	ICC (95% CI)	Level of Agreement
Accuracy	0.91 (0.81–0.96)	Excellent
Global Quality Scale (GQS)	0.88 (0.76–0.94)	Good
PEMAT-P Understandability (%)	0.83 (0.68–0.91)	Good
PEMAT-P Actionability (%)	0.79 (0.62–0.89)	Moderate–Good

ICC: Intraclass Correlation Coefficient; GQS: Global Quality Scale; PEMAT-P: Patient Education Materials Assessment Tool for Printable Materials. Readability was calculated using the Ateşman Readability Index and was not included in the inter-rater reliability analysis because it was a formula-based objective text metric rather than an expert-rated variable.

Specifically, the reliability was excellent for accuracy ($ICC = 0.91$) and good for GQS and PEMAT-P understandability.

The actionability subscale showed moderate-to-good agreement ($ICC = 0.79$), indicating consistent but slightly variable interpretation between evaluators. These findings supported the reliability of the evaluator ratings used in the subsequent comparative analyses.

Descriptive statistics are presented as median [interquartile range, IQR] (Table 3).

No statistically significant differences were found among chatbots for accuracy ($p = 0.752$, $\eta^2 = 0.01$) or GQS ($p = 0.516$, $\eta^2 = 0.02$).

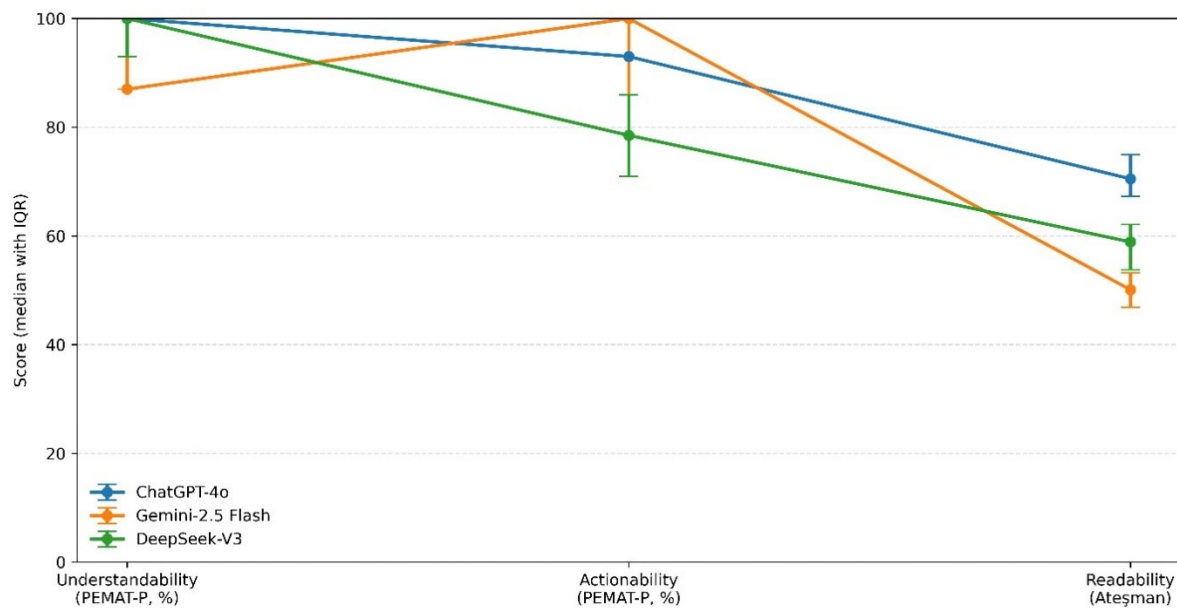
However, significant differences were observed in PEMAT-P understandability, actionability, and readability scores ($p < 0.001$) (Figure 1). The ChatGPT-4o and DeepSeek-V3 groups achieved the highest median scores for understandability, while Gemini-2.5 Flash showed lower understandability scores. Regarding readability, ChatGPT-4o achieved the highest median score (70.5 [67.3–75.0]), followed by DeepSeek-V3 (59.1 [53.4–62.3]) and Gemini-2.5 Flash (50.5 [46.3–53.5]). For actionability, both Gemini-2.5 Flash (100.0 [86.0–100.0]) and ChatGPT-4o (94.8 [86.0–100.0]) showed higher median scores than DeepSeek-V3 (82.3 [71.0–86.0]).

Table 3: Evaluation outcomes of chatbot groups: median [IQR] and statistical comparison.

Measure	ChatGPT-4o Median [IQR]	Gemini-2.5 Flash Median [IQR]	DeepSeek-V3 Median [IQR]	<i>p</i> -value	Effect Size (η^2)
Accuracy (Likert)	5.0 [5.0–5.0]	5.0 [5.0–5.0]	5.0 [4.9–5.0]	0.752	0.01
Global Quality Scale (GQS)	5.0 [4.9–5.0]	5.0 [4.0–5.0]	5.0 [4.0–5.0]	0.516	0.02
PEMAT-P Understandability (%)	100.0 [96.5– 100.0]	87.0 [87.0– 93.0]	100.0 [93.0– 100.0]	<0.001*	0.43
PEMAT-P Actionability (%)	94.8 [86.0– 100.0]	100.0 [86.0– 100.0]	82.3 [71.0– 86.0]	<0.001*	0.36
Readability (Ateşman Index)	70.5 [67.3–75.0]	50.5 [46.3– 53.5]	59.1 [53.4– 62.3]	<0.001*	0.67

Values are presented as median [interquartile range, IQR]. Accuracy and GQS were scored on a 5-point Likert scale (1 = lowest, 5 = highest). PEMAT-P scores are presented as percentages. Readability was assessed using the Ateşman Index, where higher scores indicate easier readability. Kruskal–Wallis H test results are reported; * $p < 0.05$ indicates statistical significance. $\eta^2 = 0.01$ small, 0.06 medium, ≥ 0.14 large effect size.

Figure 1: Comparison of three chatbot models (ChatGPT-4o, Gemini-2.5 Flash, and DeepSeek-V3) for PEMAT-P understandability, PEMAT-P actionability, and readability (Ateşman Index).



Post-hoc pairwise comparisons using Dunn's test with Bonferroni adjustment are presented in Table 4. For PEMAT-P understandability, ChatGPT-4o and DeepSeek-V3 showed significantly higher scores than Gemini-2.5 Flash, whereas no significant difference was observed between ChatGPT-4o and DeepSeek-V3. For PEMAT-P actionability, ChatGPT-4o and Gemini-2.5 Flash scored significantly higher than DeepSeek-V3, while the difference between ChatGPT-4o and Gemini-2.5 Flash was not significant. For readability, ChatGPT-4o scored significantly higher than both Gemini-2.5 Flash and DeepSeek-V3, and DeepSeek-V3 also scored significantly higher than Gemini-2.5 Flash.

Effect size analysis revealed large effects for readability ($\eta^2 = 0.67$), understandability ($\eta^2 = 0.43$), and actionability ($\eta^2 = 0.36$), indicating meaningful performance differences among chatbots in these communication-related domains. Overall, while all three chatbots provided highly accurate and qualitatively consistent responses, ChatGPT-4o demonstrated the most favorable overall profile, particularly in readability and patient-oriented clarity. However, the post-hoc findings also showed that DeepSeek-V3 had comparable understandability to ChatGPT-4o, whereas Gemini-2.5 Flash had comparable actionability to ChatGPT-4o.

Table 4: Dunn–Bonferroni post-hoc pairwise comparisons for variables with significant Kruskal–Wallis results.

Variable	Pairwise comparison	Adjusted p-value
PEMAT-P Understandability	ChatGPT-4o vs Gemini-2.5 Flash	<0.001
PEMAT-P Understandability	ChatGPT-4o vs DeepSeek-V3	1.000
PEMAT-P Understandability	Gemini-2.5 Flash vs DeepSeek-V3	<0.001
PEMAT-P Actionability	ChatGPT-4o vs Gemini-2.5 Flash	1.000
PEMAT-P Actionability	ChatGPT-4o vs DeepSeek-V3	<0.001
PEMAT-P Actionability	Gemini-2.5 Flash vs DeepSeek-V3	<0.001
Readability / Ateşman Index	ChatGPT-4o vs Gemini-2.5 Flash	<0.001
Readability / Ateşman Index	ChatGPT-4o vs DeepSeek-V3	<0.001
Readability / Ateşman Index	Gemini-2.5 Flash vs DeepSeek-V3	0.031

Adjusted p-values were obtained using Dunn's post-hoc test with Bonferroni correction. Only variables with statistically significant Kruskal–Wallis results are presented.

DISCUSSION

The aim of this study was to compare artificial intelligence responses to frequently asked patient questions in periodontology with criteria of accuracy, overall quality, understandability, actionability, and readability. All three models scored high in terms of accuracy and overall quality; however, significant differences were observed, particularly in the dimensions of understandability, actionability, and readability. Based on these findings, the H_0 hypothesis could not be rejected in terms of accuracy and quality, while the H_1 hypothesis was supported in the other criteria. The results show that it is not sufficient for AI-supported chatbots to simply provide accurate information in healthcare communication; it is equally important that the information is conveyed in a way that is understandable, actionable, and safe for patients to use.

According to the results obtained, all three chatbots performed at a high level of accuracy and overall quality, with median scores close to the maximum possible value. This finding is consistent with a previous study reporting that ChatGPT's responses to patient questions related to periodontology were generally accurate.⁶ An evaluation conducted for oral, dental, and maxillofacial surgery also

showed that ChatGPT answered patient questions with a high degree of accuracy. Furthermore, comparative studies evaluating questions related to traumatic dental injuries and prosthetics have reported that ChatGPT-4o exhibits higher response quality and consistency compared to other models.^{11, 12} Although no statistically significant difference was found in our study, the tendency for ChatGPT-4o's accuracy and quality scores to be higher than other models is consistent with these findings.

While it is important for chatbots to be able to answer patient questions with high accuracy, providing only correct information is not sufficient for effective healthcare communication. Furthermore, it is crucial that these responses are understandable, readable, and capable of prompting the patient to take action.^{17, 18} From this perspective, the findings of the present study revealed domain-specific differences among chatbot models rather than uniform superiority of a single model across all evaluated criteria. Post-hoc analyses showed that ChatGPT-4o and DeepSeek-V3 had comparable understandability scores and both performed better than Gemini-2.5 Flash in this domain. For actionability, ChatGPT-4o and Gemini-2.5 Flash showed comparable performance and both scored higher than

DeepSeek-V3. In contrast, readability differed across all three models, with ChatGPT-4o achieving the highest scores, followed by DeepSeek-V3 and Gemini-2.5 Flash. ChatGPT-4o's ability to generate simple and structured responses in Turkish may explain its favorable readability profile, whereas Gemini-2.5 Flash's tendency to provide practical suggestions may have contributed to its high actionability scores. This apparent discrepancy may be explained by the fact that readability and actionability assess different dimensions of patient education material: a response may contain clear behavioral recommendations and concrete steps, thereby achieving high actionability, while still being linguistically more complex and therefore less readable according to the Ateşman Index. Similar to our study, a study on patient education states that ChatGPT and Gemini chatbots are more understandable and actionable compared to other chatbots.¹⁹ Furthermore, a recent study on the readability levels of questions asked in restorative dentistry similarly indicates that ChatGPT-4o outperformed other chatbots.²⁰ These results largely parallel the findings observed in our study.

The findings of this study support the literature by demonstrating that AI-supported chatbots can serve as supportive educational and communication tools for patients, going beyond merely being information-sharing systems.²¹ ChatGPT-4o, in particular, showed the most favorable overall profile, especially in readability and patient-oriented clarity; however, the findings also indicate that chatbot performance should be interpreted according to the specific communication domain being evaluated.²² Such models may help provide clear and accessible preliminary information on topics of interest to patients, including oral hygiene, post-treatment care, or systemic risk factors. Nevertheless, because actual patient comprehension, behavioral change, and clinical outcomes were not directly evaluated in this

study, these findings should be interpreted as expert-based evidence of response quality rather than direct evidence of patient benefit. Therefore, AI-supported chatbot responses may be considered adjunctive educational resources that can support professional communication, although their effects on patient behavior and oral health outcomes should be confirmed in future patient-based studies.^{23, 24}

This study's strengths include that the questions were based on real patient experiences and clinical expressions, not just theory. This allowed for patient-centered assessments and ensured that the results were more closely aligned with clinical practice. To our knowledge, this study is one of the first in the literature to compare responses to patient questions about periodontal diseases across different artificial intelligence models in a multidimensional manner, using criteria such as accuracy, overall quality, understandability, actionability, and readability. Another important aspect is that the questions and answers were evaluated in Turkish. This approach ensured the preservation of the local linguistic context and contributed to the objective analysis of Turkish texts by enabling the use of the Ateşman readability index. Furthermore, examining the most recent versions of chatbots enhances the study's timeliness, while the preference for freely accessible versions adds additional value in terms of reflecting real user experience and societal accessibility.

However, some limitations of this study should also be considered. First, the number of questions evaluated is limited, and only text-based responses were analyzed; the impact of visual or interactive content was not investigated. In addition, no patient participants were included in the present study; therefore, actual patient comprehension, satisfaction, behavioral response, or clinical decision-making was not directly evaluated. Accordingly, the findings should be interpreted

as expert-based assessments of response quality, understandability, and actionability, together with formula-based readability analysis, rather than direct evidence of how patients would understand or use chatbot-generated information. Furthermore, the fact that only free versions were used does not rule out the possibility of different results being obtained in paid or professional versions. Since chatbot responses may vary depending on version updates, the findings should be considered specific to the study period. In addition, only a single response was evaluated for each question from each chatbot; therefore, potential response variability across repeated interactions could not be assessed. Although this approach reflects a typical real-world user experience based on the first generated response, future studies should examine repeated prompts to evaluate the consistency of chatbot responses. Finally, although hallucinated or potentially misleading responses may occur in AI-generated health information, the present study did not include a predefined framework for systematically coding hallucinations according to chatbot model, question type, frequency, or clinical severity. Therefore, the frequency and distribution of hallucinated responses across chatbot models could not be reported as a separate outcome. In the context of periodontal health, hallucinated or misleading responses may result in inappropriate self-care practices, delayed professional consultation, misunderstanding of disease severity, or unrealistic expectations regarding treatment outcomes. Therefore, chatbot-generated responses should be regarded as supportive educational content rather than a substitute for individualized diagnosis, clinical decision-making, or professional dental advice. Future studies should include a predefined hallucination assessment protocol to evaluate the frequency, severity, and clinical relevance of potentially misleading or fabricated information generated by chatbots.

CONCLUSIONS

No significant differences were found in terms of accuracy and overall quality among the three different AI-supported chatbots; however, distinct differences were observed in the criteria of understandability, actionability, and readability. ChatGPT-4o demonstrated the most favorable overall profile, particularly in readability and expert-assessed patient-oriented clarity. However, DeepSeek-V3 showed comparable understandability to ChatGPT-4o, whereas Gemini-2.5 Flash showed comparable actionability to ChatGPT-4o. Therefore, chatbot performance should not be interpreted as uniformly superior for a single model across all domains, but rather as domain-specific according to the evaluated communication criterion. The findings suggest that AI-supported systems may serve as supportive tools for patient communication and education; however, because no patient participants were included, actual patient comprehension, usability, behavioral response, and clinical outcomes could not be directly evaluated. Future studies involving patient participants, interactive designs, and multilingual populations may provide further insights into the practical usability of AI-supported chatbots in dentistry.

Ethical Approval

This study was approved by the Lokman Hekim University Scientific Research Ethics Committee (Approval No: 2025/209) and was conducted in accordance with the principles of the Declaration of Helsinki.

Financial Support

The authors declare that this study received no financial support.

Conflict of Interest

The authors deny any conflicts of interest related to this study.

Author Contributions

Design: MK, EO, Data collection or access: MK, EO, Analysis and comments: MK, EO, Literature search: MK, EO, Writing: MK, EO.

REFERENCES

1. Duran-Pinedo AE, Solbiati J, Teles F, Frias-Lopez J. Subgingival host-microbiome metatranscriptomic changes following scaling and root planing in grade II/III periodontitis. *J Clin Periodontol*. 2023;50:316-30.
2. Røslund A, Bertelsen RJ, Bunaes DF, Drengenes C, et al. Periodontitis is associated with airflow obstruction in the Malmö Offspring Dental Study. *J Clin Periodontol*. 2024;51:86-96.
3. Castañeda-Avila MA, Pérez CM, Vivaldi-Oliver JA, Díaz-Toro EC, et al. Comparison of oral human papilloma virus detection methods among Hispanic adults. *Clin Exp Dent Res*. 2022;8:169-75.
4. Wehmeyer MM, Corwin CL, Guthmiller JM, Lee JY. The impact of oral health literacy on periodontal health status. *J Public Health Dent*. 2014;74:80-7.
5. Xu Y, Wang M, Bao L, Cheng Z, et al. A cross-sectional study based on the Comprehensive Model of Information seeking: which factors influence health information-seeking behavior in patients with periodontitis. *BMC Oral Health*. 2024;24:1307.
6. Babayiğit O, Tastan Eroglu Z, Ozkan Sen D, Ucan Yarkac F. Potential use of ChatGPT for patient information in periodontology: a descriptive pilot study. *Cureus*. 2023;15:e48518.
7. Eggmann F, Weiger R, Zitzmann NU, Blatz MB. Implications of large language models such as ChatGPT for dental medicine. *J Esthet Restor Dent*. 2023;35:1098-1102.
8. Badran A, Keraa K, Farghaly MM. The impact of oral health literacy on dental anxiety and utilization of oral health services among dental patients: a cross-sectional study. *BMC Oral Health*. 2023;23:146.
9. Wu X, Cai G, Guo B, Ma L, Shao S, et al. A multi-dimensional performance evaluation of large language models in dental implantology: comparison of ChatGPT, DeepSeek, Grok, Gemini and Qwen across diverse clinical scenarios. *BMC Oral Health*. 2025;25:1272.
10. Tuzlalı M, Baki N, Aral K, Aral CA, et al. Evaluating the performance of AI chatbots in responding to dental implant FAQs: a comparative study. *BMC Oral Health*. 2025;25:1548.
11. Guven Y, Ozdemir OT, Kavan MY. Performance of artificial intelligence chatbots in responding to patient queries related to traumatic dental injuries: a comparative study. *Dent Traumatol*. 2025;41:338-47.
12. Gheisarifar M, Shembesh M, Koseoglu M, Fang Q, et al. Evaluating the validity and consistency of artificial intelligence chatbots in responding to patients' frequently asked questions in prosthodontics. *J Prosthet Dent*. 2025;134:199-206.
13. 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;25:871.
14. Bernard A, Langille M, Hughes S, Rose C, et al. A systematic review of patient inflammatory bowel disease information resources on the World Wide Web. *Am J Gastroenterol*. 2007;102:2070-2077.
15. Shoemaker SJ, Wolf MS, Brach C. Development of the Patient Education Materials Assessment Tool (PEMAT): a new measure of understandability and actionability for print and audiovisual patient information. *Patient Educ Couns*. 2014;96:395-403.
16. Ateşman E. Türkçede okunabilirliğin ölçülmesi. *Dil Dergisi*. 1997;58:71-4.
17. Gökcek Taraç M, Nale T. Artificial intelligence in pediatric dental trauma: do artificial intelligence chatbots address parental concerns effectively? *BMC Oral Health*. 2025;25:736.

18. Alshamrani AS, Alshehri RS, Alqahtani YA, Abdulrahman J, et al. Evaluating understandability and actionability of artificial intelligence in creating a high-quality educational content for type 2 diabetes using predesigned prompts. *Eur J Med Health Res.* 2025;3:227-33.
19. Stephenson-Moe CA, Behers BJ, Gibons RM, Behers BM, et al. Assessing the quality and readability of patient education materials on chemotherapy cardiotoxicity from artificial intelligence chatbots: an observational cross-sectional study. *Medicine (Baltimore).* 2025;104:e42135.
20. Ozdemir ZM, Yapici E. Evaluating the accuracy, reliability, consistency, and readability of different large language models in restorative dentistry. *J Esthet Restor Dent.* 2025;37:1740-1752.
21. Tudor Car L, Dhinakaran DA, Kyaw BM, Kowatsch T, et al. Conversational agents in health care: scoping review and conceptual analysis. *J Med Internet Res.* 2020;22:e17158.
22. Thorat V, Rao P, Joshi N, Talreja P, et al. Role of artificial intelligence (AI) in patient education and communication in dentistry. *Cureus.* 2024;16:e59799.
23. Agrawal R, Khan AM, Moothedath M, V JR, et al. Effect of AI-based chatbots in promoting oral health awareness among rural populations. *Bioinformation.* 2025;21:827-31.
24. Kaneyasu Y, Shigeishi H, Sugiyama M, Ohta K. Effectiveness of e-learning to promote oral health education: a systematic review and meta-analysis. *Medicine (Baltimore).* 2023;102:e36550.