



Do AI Chatbots Tell the Truth About Dentin Hypersensitivity? A Comparative Evaluation of Quality, Accuracy, and Readability

Seval Ceylan Şen^{1,a,*}, Özlem Saraç Atagün^{1,b}, Gülbahar Ustaoglu^{1,c}, Zeynep Hazan Yıldız^{1,d}, Rumeysa Nur Kayacı^{1,e}

¹Department of Periodontology, Gulhane Faculty of Dentistry, University of Health Sciences, Ankara, Türkiye.

*Corresponding author

Research Article

History

Received: 24/12/2025

Accepted: 31/01/2026

ABSTRACT

Objectives: Dentin hypersensitivity (DH) is a common dental complaint, and many patients now seek information from AI chatbots. Yet, the accuracy, reliability, and readability of chatbot-generated DH content remain uncertain.

Materials and Methods: A consensus-based DH question set was presented to three AI chatbots (ChatGPT-4o, DeepSeek, Copilot) in independent, standardized sessions. Three blinded periodontologists evaluated the responses using CLEAR, mGQS, accuracy scores, DISCERN, and readability metrics (FRE, FKGL). Non-parametric tests compared inter-model differences.

Results: Inter-group comparisons revealed statistically significant variations in FKGL ($p = 0.025$), DISCERN ($p = 0.004$), and the length of generated responses ($p < 0.001$). Copilot yielded the highest reliability and quality, DeepSeek produced the most readable content, and ChatGPT showed the most significant variability. Copilot also had the highest proportion of fully accurate, high-quality answers, whereas low-accuracy output occurred only with ChatGPT. Strong correlations were noted among accuracy, completeness, and overall quality.

Conclusions: AI chatbots can generate clinically relevant DH information, but performance varies. Copilot showed the best balance of accuracy and reliability, DeepSeek provided the most accessible language, and ChatGPT demonstrated inconsistent results. Clinician oversight remains essential when using AI-generated content for patient education.

Keywords: Artificial Intelligence, CLEAR, dentin hypersensitivity; DISCERN, modified Global Quality Score, readability

Yapay Zekâ Sohbet Robotları Dentin Hipersensitivitesi Hakkında Doğru Bilgileri Veriyor mu? Kalite, Doğruluk ve Okunabilirliğin Karşılaştırmalı Değerlendirmesi

Araştırma Makalesi

Süreç

Geliş: 24/12/2025

Kabul: 31/01/2026

ÖZ

Amaç: Dentin hassasiyeti (DH), sık karşılaşılan bir dental yakınma olup, günümüzde birçok hasta bu konuda bilgi edinmek amacıyla yapay zekâ (YZ) tabanlı sohbet botlarına başvurmaktadır. Ancak, sohbet botları tarafından üretilen DH içeriklerinin doğruluğu, güvenilirliği ve okunabilirliği henüz netlik kazanmamıştır.

Gereç ve Yöntemler: Uzman görüşüne dayalı olarak oluşturulan DH soru seti, üç farklı YZ sohbet botuna (ChatGPT-4o, DeepSeek ve Copilot) bağımsız ve standartlaştırılmış oturumlarda yöneltilmiştir. Elde edilen yanıtlar, üç körlenmiş periodontolog tarafından CLEAR kriterleri, modifiye Global Kalite Skoru (mGQS), doğruluk puanlaması, DISCERN aracı ve okunabilirlik ölçütleri (Flesch Reading Ease [FRE], Flesch-Kincaid Grade Level [FKGL]) kullanılarak değerlendirilmiştir. Modeller arası karşılaştırmalar için parametrik olmayan istatistiksel testler uygulanmıştır.

Bulgular: FKGL ($p = 0,025$), DISCERN puanları ($p = 0,004$) ve yanıt uzunluğu ($p < 0,001$) açısından YZ modelleri arasında istatistiksel olarak anlamlı farklar saptanmıştır. Copilot, en yüksek güvenilirlik ve kalite düzeyini gösterirken; DeepSeek en okunabilir içeriği üretmiştir. ChatGPT ise en yüksek değişkenliği sergilemiştir. Ayrıca Copilot, tamamen doğru ve yüksek kaliteli yanıtların en yüksek oranına sahipken, düşük doğruluk düzeyindeki yanıtlar yalnızca ChatGPT’de gözlenmiştir. Doğruluk, içerik bütünlüğü ve genel kalite arasında güçlü korelasyonlar tespit edilmiştir.

Sonuç: YZ tabanlı sohbet botları, dentin hassasiyeti hakkında klinik açıdan anlamlı bilgiler üretebilmekle birlikte, performansları modeller arasında farklılık göstermektedir. Copilot, doğruluk ve güvenilirlik açısından en dengeli performansı sergilerken; DeepSeek daha erişilebilir bir dil kullanımı sunmuş, ChatGPT ise tutarsız sonuçlar göstermiştir. Hasta eğitimi amacıyla YZ tarafından üretilen içeriklerin kullanımında klinisyen denetimi hâlen büyük önem taşımaktadır.

Anahtar Kelimeler: Dentin hassasiyeti, CLEAR, DISCERN, modifiye Global Kalite Skoru, okunabilirlik, yapay zekâ

Copyright



This work is licensed under
Creative Commons Attribution 4.0
International License

^a dt.sceylan@hotmail.com

^c gulbaharustaoglu@hotmail.com

^e rumeysakayaci4@gmail.com

^b 0000-0002-3286-7819

^d 0000-0002-4205-861X

^e 0009-0000-8753-5806

^b ozlemsarac2806@hotmail.com

^d zhazany@gmail.com

^b 0000-0002-2964-8244

^d 0009-0004-6222-420X

How to Cite: Şen SC, Saraç Atagün Ö, Ustaoglu G, Yıldız ZH, Kayacı RN. (2026) Do AI Chatbots Tell the Truth About Dentin Hypersensitivity? A Comparative Evaluation of Quality, Accuracy, and Readability. Cumhuriyet Dental Journal, 29(1): 138-147.

Introduction

Dentin hypersensitivity is a widespread clinical condition, typically presenting as short, sharp pain triggered by external stimuli acting on exposed dentin.^{1,2} Although its pathogenesis is multifactorial, the most common etiological contributors include dental caries and

traumatic lesions, followed by enamel and cementum loss due to abrasion, erosion, corrosion, or abfraction.³⁻⁵ Additional predisposing factors—such as improper brushing techniques, periodontal therapy-related root exposure, acidic dietary habits, occlusal stress, and age-

related changes—may further increase dentin vulnerability and symptom severity. Because the clinical presentation of dentin hypersensitivity may resemble caries, pulpitis, or periodontal pathology, accurate diagnosis requires careful history-taking supported by clinical and radiographic examination.⁶⁻⁸

With the increasing integration of artificial intelligence (AI) in dentistry, AI-based chatbots have become widely used for patient education, preliminary information seeking, and support in understanding oral health conditions.⁹⁻¹⁴ As patients increasingly turn to these platforms for guidance, the quality and reliability of chatbot-generated explanations regarding dentin hypersensitivity have become clinically relevant.

Despite the high prevalence of dentin hypersensitivity and the rising use of AI chatbots, no previous study has systematically evaluated AI-generated responses on this condition. To address this gap, the present study developed a structured set of clinically grounded questions based on the consensus framework published by Nardi et al.¹⁵ in *The Decision Tree for Clinical Management of Dentin Hypersensitivity*. Rather than asking patients directly, the questions were formulated by experienced periodontists using evidence-based clinical pathways to ensure they reflected common patient concerns and key diagnostic–therapeutic domains.

Accordingly, this study aimed to comparatively evaluate the accuracy, quality, readability, and reliability of responses generated by three AI chatbots—ChatGPT-4o, DeepSeek, and Copilot—using consensus-driven questions on dentin hypersensitivity. These questions were developed by periodontology specialists using a consensus-based framework to ensure clinical relevance and alignment with established diagnostic and management principles. We hypothesized that AI chatbot performance would differ significantly across platforms and that some models would demonstrate superior accuracy and reliability in providing clinically appropriate information.

Material and Methods

Given that no human participants, identifiable information, or interventions were involved, and that all evaluated AI chatbots were publicly accessible, ethical approval was not required for this study.

Question Development and Expert Assessment

Two periodontology specialists initially developed 41 open-ended questions to ensure scientific accuracy and clinical relevance. The question development process was guided by the key themes, clinical pathways, and diagnostic–therapeutic considerations outlined in the consensus report *“The Decision Tree for Clinical Management of Dentin Hypersensitivity”* by Nardi et al.¹⁵, which provided an evidence-based framework for the etiology, diagnosis, differential diagnosis, and management of dentin hypersensitivity. To refine the clarity and clinical appropriateness of the question set, the same two specialists (SCŞ, GU) subsequently served as an expert panel. They reviewed all items for clarity,

representativeness, and alignment with the diagnostic, etiological, symptomatic, and management-related dimensions emphasized in the 2022 consensus document. Through this process, the panel ensured that the questions accurately reflected the fundamental domains of dentin hypersensitivity and were suitable for patient-oriented assessment.¹⁵

The initial pool of 41 open-ended questions underwent an informal pilot evaluation by the expert panel to enhance content validity. At this stage, no formal content validity index (CVI) or inter-rater reliability coefficients were calculated. Instead, each panel member independently reviewed the questions, followed by a structured yet flexible, consensus-based discussion. Through an iterative review process, the panel assessed the clarity, comprehensiveness, and clinical relevance of each item and collectively agreed on necessary revisions to ensure conceptual coherence and clinical applicability.

This pragmatic, consensus-driven process was deemed sufficient to establish the clinical validity of the question set, thereby rendering formal psychometric testing unnecessary at this stage. Panel members' verbal feedback and agreement ensured that the final questions adequately reflected the fundamental domains of dentin hypersensitivity, including etiology, differential diagnosis, triggers, prevention, and management.

During the review, each question was assessed using a binary classification: “acceptable” (clear and understandable enough to elicit accurate chatbot explanations) or “unacceptable” (insufficient clarity or inability to convey the intended DH-related concept). Following this evaluation, the question pool was refined from 41 open-ended items to 18 questions that were most clinically appropriate and representative, selected for use in subsequent analyses.

AI Platform Interactions

Questions were submitted to three AI conversational agents: CHATGPT-4o, DeepSeek, and Microsoft Copilot. The finalized question list was administered sequentially and independently to each model by three independent investigators (ZHY, RNK, OSA) (Table 1). To minimize personalization bias and historical learning effects, new accounts were created for each platform, and each system was accessed in incognito/private browsing mode using minimal profile information (i.e., no name, no search history, no prior chat history, no linked phone number, and default platform settings only). This ensured that the models generated non-personalized baseline outputs consistent with those of anonymized first-time users. All interactions were conducted on September 15, 2025, between 09:00 and 16:00, on the same device and internet network. All outputs were recorded verbatim for subsequent assessment. Access to the AI platforms was via their standard web interfaces on a computer, and all interactions were designed to reflect usage conditions that most closely resemble real-world clinical information-seeking behavior. Large language models (LLMs) were not explicitly trained on the dentin hypersensitivity literature or on the consensus-based

decision tree used to guide the development of the question set. No pre-testing, prompt optimization strategies, or additional interventions were implemented. Interactions with the AI models were strictly confined to the predetermined, English-language dentin hypersensitivity questions developed for research purposes, with English used consistently for both inputs and outputs. No supplementary instructions, system-level modifications, or alternative interaction modalities were applied. Significantly, the investigators did not alter any generation parameters. Accordingly, all responses were produced under default system settings, thereby reflecting typical real-world usage conditions and minimizing the potential for researcher-induced bias.

To prevent prior interactions from influencing AI language model outputs, each question was entered into a separate chat session. At the beginning of each session, system memory was cleared to ensure that no previous conversation history was retained and that the model

responded exclusively to the current input. This procedure minimized contextual carry-over effects and potential predictive bias associated with sequential prompting. Consequently, each response was generated independently and solely based on the question presented. Furthermore, each language model was permitted to create only a single response, with no corrections, regenerations, or alternative outputs allowed. All generated responses were subsequently documented in Microsoft Word format (Microsoft, Redmond, WA, USA).

The responses generated by the three AI models were anonymized and subsequently evaluated by three experienced periodontology specialists. To ensure complete impartiality during the assessment process, all information identifying the source model was concealed. Before evaluation, the responses were color-coded, enabling the evaluators to assess the content without knowledge of which chatbot had produced each response.

Table 1. The question list

QUESTIONS	
1)	How is tooth sensitivity (dentin hypersensitivity) defined, and what are its characteristic clinical features?
2)	What is the pathophysiology of dentin hypersensitivity, and which etiological factors most commonly contribute to its development?
3)	Is dentin hypersensitivity a permanent condition, or can it resolve over time?
4)	Which foods and beverages are known to exacerbate dentin hypersensitivity by stimulating exposed dentinal tubules?
5)	What are the mechanisms of action of desensitizing toothpaste commonly recommended for the management of dentin hypersensitivity?
6)	Can dentin hypersensitivity be mistaken for dental caries, and how can the two conditions be clinically differentiated?
7)	What evidence-based treatment approaches are recommended for reducing dentin hypersensitivity?
8)	Is hypersensitivity to thermal stimuli (cold or hot) considered physiological, or does it indicate an underlying pathological condition?
9)	What mechanisms explain the transient dentin hypersensitivity experienced after tooth-whitening procedures?
10)	What evidence-supported home-care strategies can help alleviate tooth sensitivity?
11)	In which clinical situations should patients with dentin hypersensitivity seek professional dental evaluation?
12)	Which dental and periodontal conditions may mimic dentin hypersensitivity and lead to diagnostic confusion?
13)	Under what circumstances can dentin hypersensitivity signal a more serious underlying oral pathology?
14)	What is the relationship between age and dentin hypersensitivity, and which age groups are most affected?
15)	What potential clinical consequences may arise if dentin hypersensitivity is left untreated?
16)	What is known from the literature about the association between vitamin or mineral deficiencies and tooth sensitivity?
17)	Can tooth sensitivity be resolved spontaneously without professional intervention or treatment?
18)	Which dental specialty is most appropriate for evaluating and managing dentin hypersensitivity?

Calculation of Sample Size

Sample size estimation was performed using G*Power (version 3.1.9.2).¹⁶ At a 95% confidence level ($\alpha = 0.05$), the standardized effect size was set at 0.46, based on data derived from a comparable study (Table 5; three-group AI comparison for Flesch Reading Ease).¹⁷ Assuming a theoretical power of 0.80, the minimum required sample size was calculated to be 12.

Quality and Reliability Scoring

CLEAR criteria evaluated Completeness, Lack of misinformation, Evidence-based support, Appropriateness, and Relevance, each scored on a 1–5 scale (total 5–25). Scores were categorized as low (5–11), moderate (12–18), or high quality (19–25).¹⁸

Modified Global Quality Score (mGQS) assessed contextual and content quality on a 1–5 scale, ranging from entirely incorrect (score 1) to comprehensive and fully accurate (score 5).¹⁹

Accuracy was also assessed using a 1–5 Likert scale, from completely incorrect (1) to entirely accurate (5).²⁰

Readability Assessment

The readability of AI-generated responses was assessed using the Flesch Reading Ease (FRE) and Flesch–Kincaid Grade Level (FKGL) indices. All readability metrics were calculated with Readable Pro software (version 2024.3; Readable LLC), a validated online readability assessment tool. FRE scores range from 0 to 100, with higher values indicating greater ease of reading, whereas FKGL represents the U.S. school grade level required for text comprehension. In the context of health communication, patient-oriented materials are generally recommended to be written at an eighth grade reading level or lower.^{21,22}

Discern Reliability Assessment

The first eight items of the validated DISCERN tool were used to assess reliability (clarity of objectives, achievement of goals, relevance, citation of sources, currency, balance, supporting resources, and acknowledgment of uncertainties).^{23–25} Each item was scored from 1 (“no”) to 5 (“yes”). Total scores were categorized as poor (8–15), moderate (16–31), or good reliability (32–40).²⁶

Statistical Analysis

To assess agreement among the three evaluators, the Intraclass Correlation Coefficient (ICC) was calculated for both accuracy and completeness scores. For categorical variables (CLEAR, mGQS, accuracy, and DISCERN categories), inter-rater reliability was examined using Cohen’s κ . The analysis demonstrated excellent agreement, with a κ value of 0.840 (95% CI: 0.817–0.863, $p < 0.05$). Given the high level of concordance, the evaluators’ scores were aggregated by calculating the mean for continuous variables and the mode for categorical variables, yielding a single representative score for each chatbot.

All statistical analyses were conducted using non-parametric methods. Data normality was evaluated with the Shapiro–Wilk test, which indicated non-normal distributions for all quantitative variables ($p < 0.05$). Accordingly, inter-model comparisons were performed using the Kruskal–Wallis test for continuous variables and Fisher’s exact test for categorical variables. Relationships between variables were examined using Spearman’s rank correlation coefficient, and the threshold for statistical significance was set at $\alpha = 0.05$.

Results

The distribution and comparison of response characteristics according to AI types are presented in

Table 2. The comparative analysis of the AI models revealed statistically significant differences in several evaluated parameters. While the CLEAR, mGQS, Accuracy, and FRE scores did not show substantial differences among ChatGPT, DeepSeek, and CoPilot ($p > 0.05$), notable disparities emerged in readability and linguistic complexity metrics. Specifically, the Flesch–Kincaid Grade Level (FKGL) was significantly lower for DeepSeek (67.56 ± 38.97) compared to ChatGPT (91.88 ± 45.59) and CoPilot (88.38 ± 26.91) ($p = 0.025$), indicating greater readability difficulty in DeepSeek responses. Similarly, the average number of words per sentence and average syllables per word varied significantly among models ($p = 0.001$ and $p = 0.011$, respectively), with DeepSeek generating shorter and syntactically simpler sentences. Furthermore, statistically significant differences were observed in structural elements of the responses. DeepSeek provided substantially longer answers in terms of sentence count (48 ± 10.80) and word count (355.38 ± 50.62), whereas ChatGPT responses were more concise (12.05 ± 6.19 sentences; 150 ± 55.93 words) ($p < 0.001$ for both variables). CoPilot displayed intermediate values between the two. The total DISCERN score, reflecting the reliability and quality of the information, was highest for CoPilot (26.55 ± 2.38), followed by DeepSeek (25.22 ± 2.69), and ChatGPT (22.44 ± 3.89), with this difference being statistically significant ($p = 0.004$).

Table 2. Distribution and comparison of response characteristics according to AI types

Variable	ChatGPT Mean \pm SD	Median (IQR)	DeepSeek Mean \pm SD	Median (IQR)	CoPilot Mean \pm SD	Median (IQR)	p
CLEAR	18.33 \pm 3.77	21 (15.75–21)	20.55 \pm 1.54	20 (19.75–22)	20.61 \pm 1.57	21 (20–21)	0.244
mGQS	4.16 \pm 0.98	4.5 (3–5)	4.22 \pm 0.73	4 (4–5)	4.60 \pm 0.60	5 (4–5)	0.214
Accuracy	4.22 \pm 0.94	4.5 (3–5)	4.50 \pm 0.51	4.5 (4–5)	4.66 \pm 0.48	5 (4–5)	0.330
FRE	377.37 \pm 190.82	388.5 (311–520)	430.09 \pm 187.60	473.5 (359–538)	457.16 \pm 174.65	472.5 (415–546)	0.283
FKGL	91.88 \pm 45.59 ^a	97 (69–122)	67.56 \pm 38.97	73 (43–85)	88.38 \pm 26.91 ^c	95 (82–99)	0.025
Average words per sentence	108.91 \pm 63.96 ^a	108 (75–144)	69.05 \pm 25.52	74 (62–83)	101.50 \pm 34.07 ^c	102 (86–127)	0.001
Average syllables per word	15.18 \pm 6.58 ^a	17 (15–18)	10.94 \pm 3.02	11 (9–13)	17.55 \pm 1.88 ^c	17.5 (16–18)	0.011
Sentences	12.05 \pm 6.19 ^a	12 (7–15)	48.00 \pm 10.80 ^b	47.5 (40–55)	18.77 \pm 5.18 ^c	17.5 (15–22)	<0.001
Words	150.00 \pm 55.93 ^a	150 (108–167)	355.38 \pm 50.62 ^b	345 (317–389)	188.72 \pm 33.19 ^c	184.5 (163–214)	<0.001
Total DISCERN	22.44 \pm 3.89 ^a	24 (20–25)	25.22 \pm 2.69 ^b	25 (23–27)	26.55 \pm 2.38 ^c	26 (24–29)	0.004

FRE: Flesch Reading Ease FKGL: Flesch–Kincaid Grade Level

^aChatGPT vs. DeepSeek significant difference ($p < 0.05$)

^bChatGPT vs. CoPilot significant difference ($p < 0.05$)

^cDeepSeek vs. CoPilot significant difference ($p < 0.05$)

Table 3 presents the distribution of CLEAR, mGQS, and Accuracy scores across different AI platforms. A statistically significant difference in CLEAR scores was observed among AI types ($p = 0.037$). While all low-quality responses were generated by ChatGPT (5.6%), the majority of responses from DeepSeek (94.4%) and CoPilot (79.6%) were classified as high quality. ChatGPT yielded the highest proportion of moderate-quality reactions (38.9%), indicating greater variability in its performance. Although the mGQS scores did not differ significantly across platforms ($p = 0.196$), CoPilot demonstrated a relatively stronger performance, with 66.7% of its

responses rated as score 5. In contrast, DeepSeek had 38.9% and ChatGPT had 50.0% in the same category. Lower scores (2 or 3) were more frequently associated with ChatGPT. Regarding accuracy, a significant difference was observed between AI types ($p < 0.001$). CoPilot produced the highest proportion of entirely accurate responses (score 5: 66.7%), followed by DeepSeek (50.0%) and ChatGPT (50.0%)—notably, only ChatGPT-generated responses rated with Scores 2 or 3, indicating reduced reliability in specific outputs. Neither DeepSeek nor CoPilot produced any responses rated below a Score 4, suggesting greater consistency in accuracy.

Table 3. Distributions of CLEAR, mGQS, and accuracy scores by AI type and relationships between them

Category	ChatGPT n (%)	DeepSeek n (%)	CoPilot n (%)	p
CLEAR				0.037
Low quality	1 (5.6)	0 (0)	0 (0)	
Moderate quality	7 (38.9)	1 (5.6)	2 (11.1)	
High quality	10 (55.6)	17 (94.4)	16 (88.9)	
mGQS				0.196
Score 2	1 (5.6)	0	0	
Score 3	4 (22.2)	3 (16.7)	1 (5.6)	
Score 4	4 (22.2)	8 (44.4)	5 (27.8)	
Score 5	9 (50.0)	7 (38.9)	12 (66.7)	
Accuracy				<0.001
Score 2	1 (5.6)	0	0	
Score 3	3 (16.7)	0	0	
Score 4	5 (27.8)	9 (50.0)	6 (33.3)	
Score 5	9 (50.0)	9 (50.0)	12 (66.7)	

*p < 0.05, %: Line percentage %AI. It displays the column percentages for AI types and the differences in column ratios by the letters in each row.

Correlation analysis of ChatGPT-generated responses revealed multiple statistically significant associations among the variables evaluated (Table 4). The CLEAR score exhibited strong positive correlations with mGQS ($r = 0.916$, $p < 0.001$), Accuracy ($r = 0.900$, $p < 0.001$), Flesch Reading Ease (FRE) ($r = 0.609$, $p = 0.007$), and Total DISCERN score ($r = 0.659$, $p = 0.003$), along with a significant negative correlation with Flesch–Kincaid Grade Level (FKGL) ($r = -0.607$, $p = 0.008$). These results indicate that greater clarity was associated with higher quality, greater accuracy, improved readability, and greater content reliability.

Similarly, mGQS showed a strong positive correlation with Accuracy ($r = 0.983$, $p < 0.001$) and Total DISCERN ($r =$

0.645 , $p = 0.004$), and a moderate positive correlation with FRE ($r = 0.548$, $p = 0.040$). A significant inverse association was observed between mGQS and FKGL ($r = -0.489$, $p = 0.040$), suggesting that higher-quality responses tended to be more readable. Accuracy was also significantly correlated with Total DISCERN ($r = 0.598$, $p = 0.009$). In addition, FRE demonstrated a significant positive association with Total DISCERN ($r = 0.635$, $p = 0.005$) and a negative association with FKGL ($r = -0.524$, $p = 0.026$). In contrast, no statistically significant correlation was identified between FKGL and Total DISCERN ($p = 0.143$), indicating that content reliability was not directly dependent on reading grade level.

Table 4. Correlation matrix for ChatGPT

	mGQS	Accuracy	Flesch Reading Ease Score	Flesch Reading Grade Level	Total Discern	
Clear	r	0.916**	0.90*	0.609**	-0.607**	0.659**
	p	0.000	0.000	0.007	0.008	0.003
mGQS	r	1.000	0.983**	0.548*	-0.489*	0.645**
	p	.	0.000	0.190	0.040	0.004
Accuracy	r		1.000	0.500*	-0.455	0.598**
	p			0.350	0.058	0.009
Flesch Reading Ease Score	r			1.000	-0.524*	0.635**
	p				0.026	0.005
Flesch-Kincaid Grade Level	r				1.000	-0.360
	p					0.143

* Correlation is significant at the 0.05 level (2-tailed).

** Correlation is significant at the 0.01 level (2-tailed).

The correlation analysis of DeepSeek responses revealed several statistically significant but generally weaker associations than those of ChatGPT (Table 5). The CLEAR score demonstrated substantial positive correlations with mGQS ($r = 0.570$, $p = 0.013$), Accuracy ($r = 0.513$, $p = 0.030$), and Total DISCERN score ($r = 0.543$, $p = 0.020$), suggesting that clarity was moderately aligned with perceived quality, accuracy, and reliability of content. However, no significant relationship was observed between CLEAR and the Flesch Reading Ease Score ($r = 0.256$, $p = 0.306$) or Flesch–Kincaid Grade Level ($r = 0.091$, $p = 0.719$), indicating that clarity did not strongly correspond with readability indices in DeepSeek

responses. The mGQS showed a strong positive correlation with Accuracy ($r = 0.627$, $p = 0.005$) and a moderate correlation with FRE ($r = 0.475$, $p = 0.046$). At the same time, no significant association was found with FKGL ($r = 0.016$, $p = 0.951$) or CLEAR ($r = 0.465$, $p = 0.052$). Notably, the Total DISCERN score was significantly associated only with FKGL ($r = 0.556$, $p = 0.017$), indicating that higher grade-level readability corresponded with greater information reliability. Overall, while DeepSeek showed meaningful correlations among quality, accuracy, and reliability metrics, its relationships with readability indices were weaker and largely non-significant.

Table 5. Correlation matrix for DeepSeek

		mGQS	Accuracy	Flesch Reading Ease Score	Flesch Reading Grade Level	Total Discern
Clear	r	0.570*	0.513*	0.256	0.091	0.543*
	p	0.013	0.030	0.306	0.719	0.020
mGQS	r	1.000	0.627**	0.475*	0.016	0.465
	p		0.005	0.046	0.951	0.052
Accuracy	r		1.000	0.353	0.225	0.443
	p			0.150	0.370	0.66
Flesch Reading Ease Score	r			1.00	-0.005	0.145
	p			.	0.984	0.567
Flesch-Kincaid Grade Level	r				1.00	0.556*
	p					0.017

*Correlation is significant at the 0.05 level (2-tailed).

**Correlation is significant at the 0.01 level (2-tailed).

Correlation analysis for CoPilot responses revealed several statistically significant relationships between clarity, quality, accuracy, and readability metrics (Table 6). The CLEAR score exhibited strong positive correlations with Accuracy ($r = 0.827, p < 0.001$) and a moderate correlation with mGQS ($r = 0.574, p = 0.013$), suggesting that responses with higher clarity were also more accurate and of better perceived quality. However, CLEAR showed no significant association with Flesch Reading Ease Score (FRE) or Flesch-Kincaid Grade Level (FKGL), indicating a limited connection between clarity and textual readability in CoPilot’s outputs. mGQS was significantly associated with Accuracy ($r = 0.755,$

$p < 0.001$), whereas its correlations with FRE and FKGL were weak and not statistically significant. Likewise, although Accuracy correlated moderately with mGQS and CLEAR, it did not show meaningful relationships with either readability index. Notably, a strong negative correlation was observed between FRE and FKGL ($r = -0.820, p < 0.001$), as expected, indicating that texts easier to read (higher FRE) corresponded with lower grade-level demands (FKGL). However, neither FRE nor FKGL demonstrated a significant association with Total DISCERN or CLEAR scores, suggesting that readability features did not substantially affect the perceived reliability or clarity of CoPilot’s responses.

Table 6. Correlation matrix for CoPilot

		mGQS	Accuracy	Flesch Reading Ease Score	Flesch Reading Grade Level	Total Discern
Clear	r	0.574*	0.827**	-0.248	0.351	0.428
	p	0.013	0.000	0.321	0.153	0.760
mGQS	r	1.000	0.755*	0.062	0.030	0.428
	p		0.000	0.806	0.906	0.076
Accuracy	r		1.000	-0.227	0.330	0.390
	p			0.365	0.181	0.110
Flesch Reading Ease Score	r			1.000	-0.820**	-0.032
	p			.	0.000	0.899
Flesch-Kincaid Grade Level	r				1.000	-0.038
	p					-0.038

*Correlation is significant at the 0.05 level (2-tailed).

**Correlation is significant at the 0.01 level (2-tailed).

Table 7 demonstrates apparent performance differences among the AI models across accuracy, CLEAR quality, and mGQS categories. Copilot produced the highest proportion of entirely accurate responses (Score 5: 66.7%), followed by DeepSeek and ChatGPT (both 50%), whereas low-accuracy scores (Score 2–3) occurred only in ChatGPT. CLEAR analysis similarly showed that DeepSeek (94.4%) and Copilot (88.9%) generated predominantly high-quality outputs, whereas ChatGPT displayed greater variability and was the only model with low-quality responses. Consistent with these findings, Copilot achieved the highest proportion of top mGQS scores (66.7%), while DeepSeek performed strongly at intermediate levels and ChatGPT again demonstrated isolated low-quality ratings. Overall, Copilot provided the most consistently accurate and high-quality information; DeepSeek showed stable, generally strong performance;

and ChatGPT exhibited the widest variability across evaluations.

Table 8 shows that CoPilot achieved the highest DISCERN scores, reflecting superior reliability and clinical trustworthiness, followed by DeepSeek; ChatGPT demonstrated moderate reliability, though it was more variable. Readability analyses further reinforced these differences. FRE values showed that CoPilot produced the most fluent and easily comprehensible explanations, with DeepSeek ranking second and ChatGPT generating the least readable outputs. FKGL findings supported this pattern: DeepSeek produced the simplest text in terms of grade-level demand, while ChatGPT and CoPilot required higher reading levels. When considered together, the FRE and FKGL indices confirm that DeepSeek offers the most accessible language level, CoPilot provides the most coherent and well-structured content, and ChatGPT shows the lowest overall readability.

Table 7. Distribution of Accuracy, CLEAR, and mGQS scores by AI model (Row % and Column %)

Category	ChatGPT n	ChatGPT %row	ChatGPT %col	DeepSeek n	DeepSeek %row	DeepSeek %col	CoPilot n	CoPilot %row	CoPilot %col	Total n	Total %
Accuracy											
Score 2	1	100	5.6	0	0	0	0	0	0	1	1.9
Score 3	3	100	16.7	0	0	0	0	0	0	3	5.6
Score 4	5	25.0	27.8	9	45	50	6	30	33.3	20	37.0
Score 5	9	30.0	50.0	9	30	50	12	40	66.7	30	55.6
Total	18	100	100	18	100	100	18	100	100	54	100
CLEAR											
Low quality	1	100	5.6	0	0	0	0	0	0	1	1.9
Moderate	7	70	38.9	1	10	5.6	2	20	11.1	10	18.5
High	10	23.3	55.6	17	39.5	94.4	16	37.2	79.6	43	79.6
Total	18	100	100	18	100	100	18	100	100	54	100
mGQS											
mGQS	ChatGPT n	%col	DeepSeek n	%col	CoPilot n	%col	Total n	Total %			
Score 2	1	5.6	0	0	0	0	1	1.9			
Score 3	4	22.2	3	16.7	1	5.6	8	14.8			
Score 4	4	22.2	8	44.4	5	27.8	17	31.4			
Score 5	9	50.0	7	38.9	12	66.7	28	51.8			
Total	18	100	18	100	18	100	54	100			

Table 8. DISCERN Score, Flesch Reading Ease (FRE), and Flesch–Kincaid Grade Level (FKGL) Comparison Across AI Models

	AI Model	Mean ± SD	Median (IQR)
DISCERN Score	ChatGPT	22.44 ± 3.89	24 (20–25)
	DeepSeek	25.22 ± 2.69	25 (23–27)
	CoPilot	26.55 ± 2.38	26 (24–29)
Flesch Reading Ease (FRE)	ChatGPT	377.37 ± 190.82	388 (311–520)
	DeepSeek	430.09 ± 187.60	473 (359–538)
	CoPilot	457.16 ± 174.65	472 (415–546)
Flesch–Kincaid Grade Level (FKGL)	ChatGPT	91.88 ± 45.59	97 (69–122)
	DeepSeek	67.56 ± 38.97	73 (43–85)
	CoPilot	88.38 ± 26.91	95 (82–99)

Discussion

Dentin hypersensitivity (DH) is a widespread clinical condition, typically manifesting as short, sharp pain triggered by thermal, evaporative, tactile, osmotic, or chemical stimuli acting on exposed dentin, with a notable negative impact on patients’ quality of life.¹ Given its multifactorial etiology and the wide variety of available preventive and therapeutic strategies, patients frequently seek information beyond traditional dental consultations.^{27,28} In recent years, AI-based chatbots have become increasingly popular as accessible sources of health-related information, including dental conditions.²⁹ While these tools can provide rapid, user-friendly, and personalized responses, concerns remain regarding the reliability, accuracy, and completeness of the information they generate.³⁰ Therefore, evaluating the accuracy of AI-generated responses to common patient questions about dentin hypersensitivity is particularly important to ensure safe, evidence-based patient education and decision-making. In line with this perspective, the present study aimed to evaluate the accuracy of AI-based chatbot responses to common questions that patients with dentin hypersensitivity might frequently ask, thereby assessing

the reliability of AI as a supplementary tool for patient education in dentistry.

Several recent studies have explored the accuracy and reliability of AI-based chatbots in addressing patient inquiries across different dental specialties. For example, Alqutaibi et al.²⁸ and Mugri³¹ highlighted the diagnostic potential of AI in radiographic interpretation and in the detection of peri-implant bone loss, suggesting that these tools can augment clinical decision-making but still require human supervision. Similarly, Terzi et al.³² and Çoban and Altay³³ demonstrated that AI chatbots can provide information on dental implants. Yet variability in response accuracy and depth was observed, warranting caution in interpretation. Ibraheem revealed that artificial intelligence models demonstrate high accuracy, often exceeding 90%, in the identification and classification of dental implant systems from radiographs, thereby supporting clinicians in diagnosis and treatment planning.³⁴ Expanding beyond implantology, several investigations examined chatbot performance in prosthodontics, orthodontics, and endodontics. Abu Arqub et al.³⁵ reported that AI-generated responses about clear aligners were informative but occasionally lacked evidence-based justification. Likewise, Esmailpour et al.¹⁷

and Özcivelek and Özcan³⁶ demonstrated that chatbots could answer common patient questions about dental prostheses, though the appropriateness and comprehensiveness of responses varied across platforms. In endodontics, Bükür and Mercan³⁷ noted that AI chatbot responses on root canal retreatment were generally readable and appropriate, but concerns remained about accuracy and consistency. Our findings are partly consistent with previous studies and make unique contributions. First, the absence of statistically significant differences in overall measures (CLEAR, mGQS, Accuracy, and FRE) indicates that AI-based chatbots can perform at a comparable baseline level. However, notable variations were observed in the linguistic and structural characteristics of the responses. For instance, DeepSeek generated longer and more readable sentences, which may enhance patient comprehension and improve communication in clinical settings. Conversely, ChatGPT tended to provide shorter answers, which could be advantageous for quick information retrieval in specific scenarios; however, lower accuracy scores indicate potential limitations when used independently. The highest DISCERN scores observed in CoPilot demonstrate that this model produced more consistent responses in terms of content quality and reliability. This finding supports earlier reports that emphasized variability in response quality across platforms. Moreover, in terms of the Flesch–Kincaid Grade Level (FKGL), DeepSeek produced texts that were more easily readable, an essential advantage for patient-centered communication. Correlation analyses reinforced these differences; while ChatGPT showed strong positive relationships among CLEAR, mGQS, and accuracy scores, such associations were weaker in DeepSeek and CoPilot. These results suggest that each model has distinct strengths and weaknesses, and their selection should depend on the intended use case.

Beyond the model-specific performance differences, several broader implications emerge from our findings. Although Copilot demonstrated the highest reliability and DeepSeek produced the most accessible language level, neither platform consistently balanced clinical accuracy with patient-friendly readability. This disparity highlights a significant limitation of current AI models: highly accurate responses may remain linguistically complex, while more readable explanations may sacrifice clinical precision. Moreover, the dynamic and continuously updated nature of large language models suggests that chatbot performance is not static; therefore, the results represent only a snapshot of model behavior at the time of data collection. Another important consideration is the impact of linguistic context. Because all questions and outputs were evaluated in English, future studies should also assess chatbot performance in other languages, including those spoken by local patient populations, to improve generalizability. Finally, although a consensus-derived question set strengthened the clinical validity of this study, evaluating only single-response outputs without repeated prompt iterations limits the assessment

of response variability. These factors collectively underscore the need for cautious integration of AI chatbots into patient education workflows and emphasize the essential role of clinician oversight.

This study is subject to several limitations that should be acknowledged. First, the evaluation was restricted to three AI-based chatbots, ChatGPT-4o, DeepSeek, and Microsoft Copilot, and was conducted exclusively in the English language. This narrow scope may reduce the generalizability of the findings to other AI platforms or chatbot responses generated in different languages and cultural contexts. Expanding future research to include a wider variety of AI tools and multilingual analyses would therefore provide a more comprehensive understanding. Second, the responses were collected on a single day (September 15, 2025). Given that AI chatbots are dynamic systems that undergo frequent updates and retraining, their outputs may vary over time. Consequently, the responses obtained in this study may not reflect the performance of these models in future iterations or under different temporal conditions. Longitudinal assessments are needed to evaluate the consistency and stability of chatbot-generated content. Third, each patient's question was posed only once, and solely the initial response was analyzed. This approach does not account for the inherent variability in chatbot outputs, as repeated questioning or rephrasing prompts can yield different answers. Incorporating multiple iterations and comparisons could improve the reliability of evaluations. Moreover, although readability of chatbot responses was assessed, comprehensibility and user-friendliness were not formally evaluated using validated assessment tools, such as the Patient Education Materials Assessment Tool (PEMAT). This represents a missed opportunity to better understand the practical utility of AI-generated information from a patient-centered perspective. Finally, emerging evidence indicates that prompt design plays a critical role in shaping the accuracy, bias, and clinical relevance of AI responses. Because different phrasings of the same question may yield divergent outputs, future studies should investigate the impact of prompting strategies. Additionally, more in-depth analyses addressing potential biases, clinical accuracy, and patient comprehension over time will be essential to determine the reliability of AI chatbots as supportive tools in patient education and clinical decision-making.

Conclusions

This study demonstrates that while AI chatbots can serve as valuable supplementary tools for educating patients about dentin hypersensitivity, their performance is far from uniform. Copilot consistently produced the most reliable and clinically aligned responses, DeepSeek offered the most straightforward and readable explanations, and ChatGPT-4o showed notable variability, including occasional inaccuracies. These differences highlight a critical reality: AI chatbots may enhance patient communication, but they cannot yet replace clinician oversight. As patients increasingly turn to digital

platforms for dental advice, ensuring the accuracy, safety, and contextual appropriateness of the information provided is essential. By giving the first consensus-based appraisal of chatbot responses to dentin hypersensitivity, this study fills a significant gap in the literature. It underscores the necessity of continued monitoring, refinement, and validation of AI tools before they can be fully integrated into patient education and clinical workflows. Future research should expand to real-world settings and diverse languages to better define the role of AI in evidence-based dental care.

Acknowledgements

None.

Conflicts of Interest Statement

The authors declare no competing interests.

References

- Holland GR, Narhi MN, Addy M, Gangarosa L, Orchardson R. Guidelines for the design and conduct of clinical trials on dentine hypersensitivity. *J Clin Periodontol* 1997;24(11):808-813.
- Eyuboglu GB, Kalay T. The effects of different desensitizers and their combinations with Er, Cr: Ysgg laser on dentin tubules, and shear bond strength to dentin. *Cumhuriyet Dent J* 2022;25:47-56.
- Addy M. Tooth brushing, tooth wear and dentine hypersensitivity--are they associated? *Int Dent J* 2005;55(4):261-267.
- Gillam D, Orchardson R. Advances in the treatment of root dentin sensitivity: mechanisms and treatment principles. *Endod Topics* 2006;13:13-33.
- Gibson B, Boiko O, Robinson P, Robinson PG, Barlow A, Player T, et al. The everyday impact of dentine sensitivity. *Social Science and Dentistry*. 06/01 2010;1:11-20.
- Davari A, Ataei E, Assarzadeh H. Dentine hypersensitivity: etiology, diagnosis and treatment; a literature review. *J Dent (Shiraz)* 2013;14(3):136-145.
- Liu XX, Tenenbaum HC, Wilder RS, Quock R, Hewlett ER, Ren YF. Pathogenesis, diagnosis and management of dentin hypersensitivity: an evidence-based overview for dental practitioners. *BMC Oral Health* 2020;20(1):220.
- Miglani S, Aggarwal V, Ahuja B. Dentine hypersensitivity: Recent trends in management. *J Conserv Dent* 2010;13(4):218-24.
- Moor J. The Dartmouth College Artificial Intelligence Conference: The Next Fifty Years. *AI Magazine* 2006;27:87-91.
- Mintz Y, Brodie R. Introduction to artificial intelligence in medicine. *Minim Invasive Ther Allied Technol* 2019;28(2):73-81.
- Ding H, Wu J, Zhao W, et al. Artificial intelligence in dentistry- A review. *Front Dent Med* 2023;4:1085251.
- Yilmaz B, Gökkurt Yilmaz B, Ozbey F. Artificial intelligence performance in answering multiple-choice oral pathology questions: a comparative analysis. *BMC Oral Health* 2025;25(1):573.
- Bindra S, Jain R. Artificial intelligence in medical science: a review. *Irish journal of medical science* 2024;93(3):1419-1429
- Tayman M. Validity and reliability of responses to periodontology questions by 4 different artificial intelligence chatbots as public information sources. *Cumhuriyet Dent J* 2025;28:390-396.
- Nardi GM, Sabatini S, Acito G, Colavito A, Chiavistelli L, Campus G. The decision tree for clinical management of dentin hypersensitivity. a consensus report. *Oral Health Prev Dent* 2022;20:27-32.
- Faul F, Erdfelder E, Buchner A, Lang AG. Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. *Behav Res Methods* 2009;41(4):1149-1160.
- Esmailpour H, Rasaie V, Babae Hemmati Y, Falahchai M I. Performance of artificial intelligence chatbots in responding to the frequently asked questions of patients regarding dental prostheses. *BMC Oral Health* 2025;25(1):574.
- Sallam M, Al-Salahat K, Eid H, Egger J, Puladi B. Human versus artificial intelligence: ChatGPT-4 outperforming bing, bard, ChatGPT-3.5 and humans in clinical chemistry multiple-choice questions. *Adv Med Educ Pract* 2024;15:857-871.
- Bernard A, Langille M, Hughes S, Rose C, Leddin D, Veldhuyzen van Zanten S. A systematic review of patient inflammatory bowel disease information resources on the World Wide Web. *Am J Gastroenterol* 2007;102(9):2070-2077.
- Hatia A, Doldo T, Parrini S, Chisci E, Cipriani L, Montagna L, et al. Accuracy and completeness of ChatGPT-Generated Information on interceptive orthodontics: a multicenter collaborative study. *J Clin Med* 2024;13(3):735.
- Helvacioğlu-Yigit D, Demirtürk H, Ali K, Tamimi D, Koenig L, Almashraqi A. Evaluating artificial intelligence chatbots for patient education in oral and maxillofacial radiology. *Oral Surg Oral Med Oral Pathol Oral Radiol* 2025;139(6):750-759.
- Medicine NLo. How to write easy-to-read health materials. Updated November 2020. Accessed December 12, 2022. Available at: https://medlineplus.gov/all_easytoread.html.
- Kilinc DD, Mansız D. Examination of the reliability and readability of Chatbot Generative Pretrained Transformer's (ChatGPT) responses to questions about orthodontics and the evolution of these responses in an updated version. *Am J Orthod Dentofacial Orthop* 2024;165(5):546-555.
- Meade MJ, Dreyer CW. Web-based information on orthodontic clear aligners: a qualitative and readability assessment. *Aust Dent J* 2020;65(3):225-232.
- Stvilia B, Mon L, Yi Y. A Model for Online Consumer Health Information Quality. *JASIST* 2009;60:1781-1791.
- Onder CE, Koc G, Gokbulut P, Taskaldiran I, Kuskonmaz SM. Evaluation of the reliability and readability of ChatGPT-4 responses regarding hypothyroidism during pregnancy. *Sci Rep* 2024;14(1):243.
- Saraç Atagün Ö, Ceylan Şen S, Paksoy T. Analysis of YouTube videos as a source of information about dentin hypersensitivity. *Int J Dent Hyg* 2024;22(2):432-443.
- Alqutaibi AY, Algabri RS, Alamri AS, Alhazmi LS, Almadani SM, Alturkistani AM, et al. Advancements of artificial intelligence algorithms in predicting dental implant prognosis from radiographic images: a systematic review. *J Prosthet Dent* 2025;134(6):2177-2188
- Mertens S, Krois J, Cantu AG, Arsiwala LT, Schwendicke F.I. Artificial intelligence for caries detection: randomized trial. *J Dent* 2021;115:103849.
- Thorat V, Rao P, Joshi N, Talreja P, Shetty AR. Role of Artificial Intelligence (AI) in patient education and communication in dentistry. *Cureus* 2024;16(5):e59799.
- Mugri MH. Accuracy of artificial intelligence models in detecting peri-implant bone loss: a systematic review. *Diagnostics (Basel)* 2025;15(6):655

32. Terzi M, Yavuz MC, Bicer T, Buyuk SK. Evaluation of artificial intelligence robot's knowledge and reliability on dental implants and peri-implant phenotype. *Sci Rep* 2025;15(1):9519.
33. Çoban E, Altay B. ChatGPT may help inform patients in dental implantology. *Int J Oral Maxillofac Implants* 2024;39(5):203-208.
34. Ibraheem WI. Accuracy of artificial intelligence models in dental implant fixture identification and classification from radiographs: a systematic review. *Diagnostics (Basel)* 2024;14(8):806.
35. Abu Arqub S, Al-Moghrabi D, Allareddy V, Upadhyay M, Vaid N, Yadav S. Content analysis of AI-generated (ChatGPT) responses concerning orthodontic clear aligners. *Angle Orthod* 2024;94(3):263-272.
36. 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.
37. Büker M, Mercan G. Readability, accuracy and appropriateness and quality of AI chatbot responses as a patient information source on root canal retreatment: A comparative assessment. *Int J Med Inform.* 2025;201:105948.