NORTHWESTERN MEDICAL JOURNAL

Northwestern Med J. 2026;6(3):192-202
DOI: 10.54307/NWMJ.2026.229

RESEARCH ARTICLE

Evaluation of artificial intelligence chatbots in informing patients about CBCT: A comparative study

Savaş Özarslantürk[1], Seval Ceylan Şen[2], Özlem Saraç Atagün[2], Selin Gaş[3], Tuğçe Paksoy[4], Şeyma Çardakcı Bahar[2]

[1]Department of Oral and Maxillofacial Radiology, Gülhane Faculty of Dentistry, University of Health Sciences, Ankara, Türkiye
[2]Department of Periodontology, Gülhane Faculty of Dentistry, University of Health Sciences, Ankara, Türkiye
[3]Department of Oral and Maxillofacial Surgery, Faculty of Dentistry, İstanbul Gelişim University, İstanbul, Türkiye
[4]Department of Periodontology, Hamidiye Faculty of Dentistry, University of Health Sciences, İstanbul, Türkiye

Cite as: Özarslantürk S, Ceylan Şen S, Saraç Atagün Ö, Gaş S, Paksoy T, Çardakcı Bahar Ş. Evaluation of artificial intelligence chatbots in informing patients about CBCT: A comparative study. Northwestern Med J. 2026;6(3):192-202.

ABSTRACT

Objective: This study aimed to explore common patient inquiries about Cone-Beam Computed Tomography (CBCT) and systematically assess the accuracy, quality, usefulness, and readability of outputs from four AI-based conversational platforms (ChatGPT-3.5, DeepSeek-V3.1, Gemini 1.5 Flash, and Microsoft Copilot Free).

Materials and Methods: Common questions about CBCT were gathered from online sources and expert contributions and submitted to ChatGPT-3.5, DeepSeek, Gemini, and Copilot under standardized conditions. Outputs were independently evaluated by three specialists using CLEAR, mGQS, accuracy, usefulness, DISCERN, and readability metrics (FRE and FKGL).

Results: Significant differences were observed among AI platforms regarding CLEAR scores ($p < 0.05$), with Gemini showing higher values than ChatGPT ($p = 0.016$). Across all four platforms, strong to a very strong positive correlations were found among CLEAR, mGQS, and Accuracy scores ($r \geq 0.77$, $p < 0.001$), while these measures were strongly to very strongly and inversely correlated with Usefulness ($r \leq -0.77$, $p < 0.001$). Flesch Reading Ease and Flesch-Kincaid Grade Level demonstrated a very strong negative correlations across platforms (r ranging from −0.85 to −0.93, $p < 0.05$). Within the Gemini group, Flesch-Kincaid Grade Level showed a moderate positive association with DISCERN scores ($r = 0.496$, $p = 0.043$).

Conclusions: Gemini and DeepSeek produced more accurate and clearer responses. The readability of all AI systems was below the recommended level for patient education. These systems may support patient education but require expert validation.

Keywords: CBCT, artificial intelligence, chatbots, patient education, readability

Corresponding author: Savaş Özarslantürk **E-mail:** savasozarslanturk@yahoo.com
Received: 10.11.2025 **Accepted:** 03.07.2026 **Published:** 31.07.2026

INTRODUCTION

Cone-beam computed tomography (CBCT) is a three-dimensional (3D) imaging technique that offers high spatial resolution at a lower radiation dose relative to standard computed tomography (CT) scans. These features form the basis for its widespread use in dentistry and maxillofacial imaging. It enables detailed evaluation of maxillofacial bone structures, facilitates accurate diagnosis and staging, and supports lesion follow-up (1).

The increasing use of CBCT has heightened patient demand for information about imaging procedures. As access to healthcare professionals may be limited due to increasing workloads, patients increasingly seek information from digital resources, including artificial intelligence (AI)-based chat systems (2).

Broadly speaking, AI encompasses computational systems built to approximate specific human cognitive abilities, drawing on subfields such as Natural Language Processing (NLP), Machine Learning, and Deep Learning to do so. Recent advances have enabled AI models to analyze complex information, interpret, perform reasoning, and generate predictions (3). ChatGPT, introduced by OpenAI in 2022, attracted broad interest with its advanced text-generation capabilities (4). In 2024, Google released Gemini, an AI platform built with multimodal capabilities that allow it to work across text, visual, and graphical content types (5). DeepSeek, developed in China in the same year, was primarily designed for academic and professional applications in the fields of mathematics, engineering, science, and technology (6). Microsoft Copilot, built on advanced GPT-based models, integrates with Microsoft 365 and Windows 11 and has potential applications in optimizing healthcare workflows (7).

Patients frequently inquire about the indications, applications, benefits, and radiation risks of CBCT. Therefore, evaluating the quality of AI-derived information is increasingly important. Despite these advances, the extent to which AI-generated answers can be considered accurate, dependable, clear, and user-friendly remains a subject of ongoing debate (8).

Four widely used AI platforms were evaluated in this study based on how they responded to patient questions commonly raised about CBCT, with comparisons drawn across accuracy, explanatory quality, understandability, readability, reliability, and usability. By doing so, the study seeks to clarify both the strengths and shortcomings of current AI technologies in a healthcare context—information that may prove useful to clinicians and developers alike.

MATERIALS AND METHODS

The study followed the ethical standards of the Declaration of Helsinki, and approval from an ethics committee was not deemed necessary. Common patient questions about CBCT, covering its indications, applications, procedure, diagnostic value, radiation risks, interpretation of findings, comparison with alternative imaging methods, and insurance-related issues, were submitted to ChatGPT-3.5, DeepSeek-V3.1, Gemini 1.5 Flash, and Microsoft Copilot Free (GPT-4o–based).

AI platform evaluation and data extraction protocol

A four-stage protocol (identification, selection, refinement, and finalization) was used to generate a standardized dataset of patient questions regarding CBCT. During the identification phase, potential questions were gathered from public health forums, frequently asked questions (FAQs) on dental imaging websites, online search trends, and clinical inquiries developed by an expert panel of dental specialists. These sources were selected to reflect both common patient concerns and clinically relevant topics related to CBCT examinations. In the selection phase, all questions were compiled into a master database (Microsoft Excel), and two authors (S.C.Ş., Ö.S.A.) independently screened entries to remove duplicates and semantically overlapping items, resolving discrepancies by consensus. During refinement, grammatical errors were corrected and technical terminology was adapted to patient-friendly language to reflect queries from individuals with average health literacy. In the finalization phase, all questions were formatted as open-ended prompts without additional context or instructions to ensure consistency across platforms.

The final dataset consisted of 17 open-ended questions designed to represent the most common patient concerns regarding CBCT examinations. The questions covered multiple domains, including general information about CBCT, clinical indications, procedural aspects, diagnostic applications, radiation safety, vulnerable patient populations (such as children and pregnant individuals), interpretation of findings, comparison with alternative imaging modalities, and insurance-related issues. This broad thematic coverage was intended to reflect real-world patient information needs and simulate typical patient–provider discussions regarding CBCT.

ChatGPT-3.5 (OpenAI; https://chat.openai.com/), Gemini 1.5 Flash (Google; https://gemini.google.com/), DeepSeek-V3.1 (https://chat.deepseek.com/), and Microsoft Copilot Free (https://copilot.microsoft.com/) were accessed in incognito mode using newly created email accounts and default settings. No prompt engineering, plugins, parameter adjustments, or response regeneration were used. Questions were entered sequentially in separate "New Chat" sessions and submitted once to each chatbot on August 24, 2025, between 09:00 and 16:00 (S.G.). Responses were recorded verbatim in Microsoft Word and analyzed in their original form. All data generated for this study are included in the Supplementary Materials to facilitate transparency and reproducibility.

Outcome measures and expert assessment criteria

Three independent experts (T.P., G.U., and S.Ö.), none of whom participated in the question development process, evaluated all responses produced by the AI models. Assessments were performed using the CLEAR criteria, Usefulness scale, modified Global Quality Score (mGQS), accuracy ratings, DISCERN instrument, and readability indices (FRE and FKGL).

Response reliability and overall quality were gauged through the CLEAR criteria, which rate five domains—completeness, misinformation, evidence-based support, appropriateness, and relevance—on a 5-point scale ranging from 1 (very poor) to 5 (excellent). Summed scores can range from 5 to 25 and are grouped into low- (5-11), moderate- (12-18), or high-quality (19-25) categories (9).

Content quality was quantified via the modified Global Quality Score (mGQS), a 5-point scale spanning from 1 (strongly disagree; inaccurate or irrelevant response) to 5 (strongly agree; comprehensive and fully accurate response) (10). Meanwhile, Accuracy was captured using a distinct 5-point Likert scale, where higher ratings indicated a greater degree of factual correctness (11).

A 4-point scale—very useful, useful, partially useful, or not useful—was used to evaluate Usefulness, with classification based on the completeness and accuracy of the information conveyed (12).

Readability metrics were evaluated via two indices, FRE and FKGL, generated through Readable Pro software (https://app.readable.com). The FRE scale runs from 0 to 100, with higher scores denoting greater readability, while the FKGL reflects the grade level a reader would need to comprehend the text. Sentence length and word difficulty underlie both measures. Given accessibility concerns, patient-facing health content is typically advised to stay at or below an eighth-grade level (13).

Statistical analysis

Descriptive statistics for continuous variables were reported as means with standard deviations (SDs), or as medians with 25th–75th percentiles, depending on distribution. The Shapiro-Wilk test was used to assess normality. The Kruskal-Wallis test was applied for between-group comparisons; where significant differences emerged, Bonferroni-adjusted post hoc analyses were subsequently conducted. The Intraclass Correlation Coefficient (ICC) was employed to evaluate agreement across the three raters, while relationships among continuous variables were assessed via Spearman's rank correlation. All analyses were conducted with IBM SPSS Statistics version 27. Given the exploratory nature of this study, which sought to compare the performance of commonly used large language models in responding to CBCT-related patient questions, no a priori sample size or power calculation was carried out.

RESULTS

Inter-rater agreement demonstrated high to excellent reliability across all subjective assessment tools. Single-measure ICC values for CLEAR, mGQS, Accuracy, Usefulness, and DISCERN ranged from 0.82 to 0.94 (all $p < 0.001$). These values demonstrate strong consistency among the three evaluators.

Response characteristics according to AI platform are presented in Table 1. Significant between-platform variation in CLEAR scores was detected using the Kruskal-Wallis test ($p < 0.05$). When Bonferroni-adjusted post hoc comparisons were applied, Gemini was found to score significantly higher than ChatGPT ($p = 0.016$). None of the other pairwise comparisons showed statistical significance (Table 1).

Spearman correlation analysis of ChatGPT responses demonstrated significant positive correlations between CLEAR scores and both mGQS and Accuracy ($r = 0.768$ for both, $p < 0.001$), whereas CLEAR was negatively correlated with Usefulness ($r = -0.768$, $p < 0.001$). mGQS showed a very strong positive correlation with Accuracy ($r = 1.000$, $p < 0.001$) and a very strong negative correlation with Usefulness ($r = -1.000$, $p < 0.001$) and FKGL ($r = -0.522$, $p = 0.010$). Similarly, Accuracy was very strongly negatively correlated with Usefulness ($r = -1.000$, $p < 0.001$) and moderately negatively correlated with FKGL ($r = -0.522$, $p = 0.010$), whereas Usefulness was moderately positively correlated with FKGL ($r = 0.522$, $p = 0.010$). In addition, FRE and FKGL were very strongly and negatively correlated ($r = -0.929$, $p < 0.001$) (Table 2).

Correlation patterns observed for Copilot closely mirrored those identified for ChatGPT. CLEAR scores were positively correlated with both mGQS and Accuracy ($r = 0.960$ for both, $p < 0.001$) and negatively correlated with Usefulness ($r = -0.960$, $p < 0.001$). A very strong positive correlation was found between mGQS and Accuracy ($r = 1.000$, $p < 0.001$), while both measures showed a very strong negative correlations with Usefulness ($r = -1.000$, $p < 0.001$). Readability indices also demonstrated a very strong inverse association between FRE and FKGL ($r = -0.927$; $p = 0.004$). Additionally, moderate positive correlation was observed between FKGL and DISCERN ($r = 0.488$; $p = 0.047$) (Table 3).

Table 1. Distribution and comparison of features of responses according to AI types

	ChatGPT		CoPilot		DeepSeek		Gemini			
	Mean ± SD	M. (%25-%75Q.)	Mean ± SD.	M. (%25-%75Q.)	Mean ± SD	M. (%25-%75Q.)	Mean ± SD.	M. (%25-%75Q.)	Test Statistics	p
Clear	17.47±8.85	24(11-25)	24.06±1.48	25(24-25)	22±4.14	25(19-25)	24.65±0.86	25(25-25)	9.394	0.024*
GQS	3.59±1.66	4(2-5)	4.59±0.51	5(4-5)	4.24±1.03	5(3-5)	4.82±0.39	5(5-5)	7.160	0.067
Accuracy	3.59±1.66	4(2-5)	4.59±0.51	5(4-5)	4.24±1.03	5(3-5)	4.82±0.39	5(5-5)	7.160	0.067
Usefulness	2.41±1.66	2(1-4)	1.41±0.51	1(1-2)	1.76±1.03	1(1-3)	1.18±0.39	1(1-1)	7.160	0.067
Flesch Reading Ease Score	59,19±25,7	63,1(42,6-77,3)	51,48±22,34	55,9(38,5-68,8)	54,2±18,74	52(44,4-68,4)	51,8±14,36	52,1(45,4-58,8)	1,426	0,699
Flesch Reading Grade Level	7,88±3,4	6,7(5,2-11,1)	15,76±24,74	9,6(8,2-11,5)	8,34±2,78	8(6,2-11,1)	9,64±1,84	10(8,4-11)	4,940	0,176
DISCERN	25.65±2.06	25(25-26)	27.88±2.39	28(27-29)	27.29±3.42	28(24-30)	27±3.1	28(24-28)	4.088	0.252

*p<0.05

Table 2. Correlations among ChatGPT-related measurements

		GQS	Accuracy	Usefulness	Flesch Reading Ease Score	Flesch Reading Grade Level	DISCERN
Clear	r	0.768	0.768	-0.768	0.440	-0.456	0.060
	p	<0.001*	<0.001*	<0.001*	0.077	0.066	0.819
GQS	r		1.000	-1.000	0.471	-0.522	0.305
	p			<0.001*	0.056	0.032*	0.234
Accuracy	r			-1.000	0.471	-0.522	0.305
	p			<0.001*	0.056	0.032*	0.234
Usefulness	r				-0.471	0.522	-0.305
	p				0.056	0.032*	0.234
Flesch Reading Ease Score	r					-0.929	-0.181
	p					<0.001*	0.486
Flesch-Kincaid Grade Level	r						-0.170
	p						0.515

*p<0.05, and r: correlation coefficient

Table 3. Correlations among Copilot-related measurements

		GQS	Accuracy	Usefulness	Flesch Reading Ease Score	Flesch Reading Grade Level	DISCERN
Clear	r	0.960	0.960	-0.960	-0.202	0.244	0.231
	p	<0.001*	<0.001*	<0.001*	0.436	0.345	0.373
GQS	r		1.000	-1.000	-0.244	0.256	0.178
	p		<0.001*	<0.001*	0.345	0.321	0.495
Accuracy	r			-1.000	-0.244	0.256	0.178
	p			<0.001*	0.345	0.321	0.495
Usefulness	r				0.244	-0.256	-0.178
	p				0.345	0.321	0.495
Flesch Reading Ease Score	r					-0.927	-0.373
	p					<0.001*	0.140
Flesch-Kincaid Grade Level	r						0.488
	p						0.047*

*p<0.05 ,and r: correlation coefficient

Similar findings were observed for DeepSeek responses. CLEAR scores were strongly positively associated with both mGQS and Accuracy (r = 0.980 for both, p < 0.001) and negatively associated with Usefulness (r = −0.980, p < 0.001). A very strong positive correlation was identified between mGQS and Accuracy (r = 1.000, p < 0.001), whereas both measures showed a very strong negative correlations with Usefulness (r = −1.000, p < 0.001). In addition, FRE and FKGL demonstrated a very strong inverse correlation (r = −0.923, p = 0.028) (Table 4).

Gemini responses exhibited the strongest overall correlation pattern among the evaluated platforms.

CLEAR scores were strongly positively associated with both mGQS and Accuracy (r = 0.994 for both, p < 0.001) and negatively associated with Usefulness (r = −0.994, p < 0.001). A very strong positive correlation was observed between mGQS and Accuracy (r = 1.000, p < 0.001), whereas both measures demonstrated a very strong negative correlations with Usefulness (r = −1.000, p < 0.001). Furthermore, FRE and FKGL were very strongly and inversely correlated (r = −0.849, p < 0.001) (Table 5).

Table 4. Correlations among DeepSeek-related measurements

		GQS	Accuracy	Usefulness	Flesch Reading Ease Score	Flesch Reading Grade Level	DISCERN
Clear	r	0.980	0.980	-0.980	0.104	0.079	-0.099
	p	<0.001*	<0.001*	<0.001*	0.690	0.763	0.706
GQS	r		1.000	-1.000	-0.025	0.190	-0.086
	p		<0.001*	<0.001*	0.924	0.466	0.742
Accuracy	r			-1.000	-0.025	0.190	-0.086
	p			<0.001*	0.924	0.466	0.742
Usefulness	r				0.025	-0.190	0.086
	p				0.924	0.466	0.742
Flesch Reading Ease Score	r				1.000	-0.923	0.072
	p					<0.001*	0.783
Flesch-Kincaid Grade Level	r						-0.139
	p						0.595

*p<0.05, and r: correlation coefficient

Table 5. Correlations among Gemini-related measurements

		GQS	Accuracy	Usefulness	Flesch Reading Ease Score	Flesch Reading Grade Level	DISCERN
Clear	r	0.994	0.994	-0.994	0.356	-0.094	0.078
	p	<0.001*	<0.001*	<0.001*	0.161	0.719	0.766
GQS	r		1.000	-1.000	0.346	-0.063	0.115
	p		<0.001*	<0.001*	0.173	0.810	0.659
Accuracy	r			-1.000	0.346	-0.063	0.115
	p			<0.001*	0.173	0.810	0.659
Usefulness	r				-0.346	0.063	-0.115
	p				0.173	0.810	0.659
Flesch Reading Ease Score	r				1.000	-0.849	-0.310
	p					<0.001*	0.225
Flesch-Kincaid Grade Level	r						0.335
	p						0.189

*p<0.05, and r: correlation coefficient

DISCUSSION

Diagnostic decision-making is based on clinical history, physical examination, together with laboratory and imaging findings, to identify the underlying pathology. In dentistry, CBCT has become an essential imaging modality that enhances diagnostic accuracy by providing precise three-dimensional visualization of oral and maxillofacial structures (14). Considering these factors, ChatGPT cannot independently diagnose medical or dental conditions. However, it may serve as an assisting tool for clinicians. While it cannot replace medical practitioners who rely on communication, observation, and questioning to obtain patient history, ChatGPT can help generate clinical notes, summaries, and documentation, thereby saving time and reducing the risk of human error. Furthermore, it may offer therapy recommendations based on patient data (15). With technological advancement, patients increasingly seek information from AI tools, which may provide quick but sometimes inaccurate responses. Therefore, caution is needed when using such technologies in medical contexts. The empathy and human connection between physician and patient remain irreplaceable aspects of healthcare that artificial systems cannot replicate (16). Artificial intelligence chatbots can process a wide range of query formats, including multiple-choice, open-ended, and yes/no questions (17). The accuracy, quality, usefulness, and readability of responses produced by four AI-powered conversational platforms—ChatGPT-3.5, DeepSeek, Gemini, and Copilot—were comprehensively evaluated in this study, alongside an effort to identify the most commonly asked patient questions concerning CBCT. In our study design, questions were asked in an open-ended manner because this better reflects the conversation between the patient and the doctor.

To thoroughly evaluate the clinical safety and educational efficacy of AI-generated responses regarding CBCT, a multi-dimensional assessment framework was intentionally constructed. Each selected evaluation instrument was justified by its ability to capture a distinct and non-overlapping dimension of information quality. While the Accuracy Scale strictly audited the medical and technical correctness of the dental data provided, the modified Global Quality Score (mGQS) focused on the structural integrity, flow, and contextual completeness of the information. To prevent clinical misalignment, the Usefulness Scale was integrated to assess the pragmatic value of responses from a clinician's perspective, distinguishing between completely accurate answers that are clinically helpful and potentially vague to a patient. Furthermore, the CLEAR criteria provided an objective, domain-specific evaluation of reliability across five specific criteria, including evidence-based support and relevance, whereas the DISCERN instrument uniquely quantified the broader systemic reliability and quality of the text as an educational consumer health resource. Finally, because technically accurate and structurally high-quality text can remain inaccessible to the lay public, readability indices (FRE and FKGL) were utilized to provide an objective, mathematical computation of linguistic complexity based on syllable count and sentence length. Together, this comprehensive matrix of metrics ensured that the chatbots were not merely judged on superficial correctness but were rigorously screened for structural quality, clinical utility, trustworthiness, and actual patient accessibility.

In our study, ChatGPT demonstrated strong, positive correlations between the Clear score and both mGQS and Accuracy, with a negative correlation with Usefulness. CoPilot, DeepSeek, and Gemini exhibited progressively stronger positive correlations with mGQS and Accuracy, alongside negative correlations with Usefulness, with Gemini showing the strongest overall associations. The findings of our study are broadly consistent with the available literature on AI chatbot performance in healthcare. Hassona et al.(18) reported that ChatGPT achieved strong results in oral cancer detection based on the CLEAR assessment. Similarly, Sallam et al.(9) found that ChatGPT-4 and Microsoft Bing outperformed ChatGPT-3 and Google Bard (Gemini), which received comparatively lower evaluation scores. Esmailpour et al.(19) observed comparable GQS values across DeepSeek-R1, ChatGPT-o1, ChatGPT-4, and Dental GPT, indicating robust performance. Similarly, a recent comparative study on dental implant failure found that Gemini achieved the highest accuracy, ChatGPT-4 the highest reliability, and DeepSeek-R1 the best readability, underscoring performance differences among AI models (20). Güven et al.(21) also reported that ChatGPT-4 and Google Gemini outperformed

ChatGPT-3.5 in traumatic dental injury scenarios, with ChatGPT-4 providing the most comprehensive treatment recommendations. Hassona et al.(18) further noted high GQS scores for ChatGPT, confirming the reliability of its outputs. Some studies reported differing results. Dursun et al.(22) found CoPilot achieved the highest GQS scores, outperforming ChatGPT-4 and ChatGPT-3.5. Mohammad-Rahimi et al.(23) observed that Claude obtained the highest mean GQS, followed by GPT-4, while Bing consistently underperformed; they also highlighted that a notable proportion of chatbot-generated citations were fabricated. Ozcivelek et al.(24) and Kinikoğlu (3) emphasized that specialized models, such as Dental GPT, ChatGPT-o1, and Gemini 2.0, achieved superior accuracy compared to general-purpose models. Yilmaz et al.(25) similarly found that ChatGPT-o1 had the highest accuracy, followed by Claude, Gemini 2, and DeepSeek, while CoPilot ranked lowest. Çitir (26) reported that ChatGPT-3 and 3.5 frequently produced unverifiable citations, reflecting limitations in real-time access to data.

Singal et al.(27) compared the performance of Google Gemini 2.0 Flash, Google Gemini 2.5 Flash, ChatGPT-4o, and DeepSeek V3 on multiple-choice anatomy questions and reported an overall accuracy rate of 95.9%. Among the models evaluated, Google Gemini 2.5 Flash achieved the highest overall accuracy rate, followed in descending order by ChatGPT-4o, DeepSeek V3, and Gemini 2.0 Flash. These findings further support evidence indicating that successive generations of large language models continue to demonstrate improvements in performance across healthcare-related knowledge and reasoning tasks.

Overall, our results and the literature indicate that newer AI models offer high clarity, accuracy, and response quality, but caution is warranted due to occasional inaccuracies and unverifiable outputs, underscoring the continued need for expert oversight.

From a dental practice perspective, these findings have important implications for patient communication and education regarding CBCT examinations. As patients increasingly seek information from AI-based platforms before or after dental consultations, chatbots may serve as supplementary educational resources that help explain the purpose, indications, benefits, and limitations of CBCT imaging. In particular, they may assist patients in understanding how CBCT contributes to the evaluation of endodontic conditions, cystic lesions, impacted teeth, and maxillofacial trauma. However, the variability observed among chatbot responses, together with the presence of occasional inaccuracies and readability challenges, suggests that AI-generated information should be interpreted as a supportive resource rather than a replacement for dentist-led clinical evaluation. Dentists should therefore be aware that patients may arrive with information obtained from AI chatbots and should be prepared to verify, clarify, and contextualize this information within the framework of individualized clinical decision-making.

In our study, ChatGPT demonstrated strong relationships among readability, quality, and usefulness metrics. A moderate negative correlation was observed between GQS and Flesch Reading Grade Level, and Accuracy showed a very strong negative correlation with Usefulness. Additionally, Accuracy negatively correlated with the Flesch Reading Grade Level, while Usefulness positively correlated with it. A very strong negative correlation was also found between Flesch Reading Ease and Flesch-Kincaid Grade Level. Similarly, CoPilot, DeepSeek, and Gemini displayed a very strong negative correlation between Accuracy and Usefulness. These negative correlations between Accuracy and Usefulness should be interpreted in light of the inverse scoring structure of the Usefulness scale, where lower scores indicate greater usefulness. Therefore, higher Accuracy scores was associated with lower Usefulness scores.

A very strong inverse correlation between Flesch Reading Ease and Flesch-Kincaid Grade Level scores was evident across CoPilot, DeepSeek, and Gemini alike ($r = -0.927$, -0.923, and -0.849, respectively). For CoPilot specifically, Flesch-Kincaid Grade Level and DISCERN scores were also moderately and positively correlated, pointing to a possible link between higher reading grade levels and greater content reliability.

These findings are consistent with previous studies evaluating AI-generated readability. Hassona et al.(18) reported that ChatGPT's responses on oral cancer

detection were difficult to understand, with Flesch-Kincaid scores indicating a higher reading grade level. According to Şahin et al.(28), patients with low health literacy may find it difficult to understand content generated by ChatGPT-4, as the responses often exceed a Year 6 reading level. A comparable pattern emerged in Helvacıoğlu-Yiğit et al.'s (13) work, where none of the three models—ChatGPT-3.5, Gemini, or Copilot—produced text below an eighth-grade reading level. Güven et al.(21) also noted that Gemini had slightly higher readability than ChatGPT models, while Özcivelek et al.(24) observed that DeepSeek-R1 achieved the highest Flesch Reading Ease scores, and Dental GPT performed the worst among the tested models. Esmailpour et al.(19) further confirmed that ChatGPT responses were more complex than Gemini or Copilot, with Gemini providing the most patient-friendly outputs.

Sezer et al.(29) reported that DeepSeek achieved the highest Flesch Reading Ease Score (FRES) among the evaluated models. DeepSeek's output was nevertheless rated as "difficult" in terms of readability, placing it at a high school reading level. In addition, they reported that most of the chatbot outputs did not align with the 6th–8th grade readability standard widely advised for patient-facing educational content. Similarly, Taşyürek et al.(30) compared different question types and versions of large language models (LLMs) and found that DeepSeek-V3.1 produced the highest FRES. Despite this, none of the evaluated chatbots generated text at the recommended 6th–8th grade readability level. Nonetheless, DeepSeek-V3.1 was identified as the model that produced results closest to these thresholds. The authors suggested that this finding may be attributed to DeepSeek-V3.1 being a free, publicly accessible platform, which is likely trained on a broader, more diverse dataset encompassing various question formats and narrative styles.

This research is subject to a number of limitations. Notably, even with standardized phrasing and evaluation criteria, factors such as query wording, timing, and context may still lead to variation in AI-generated responses. Second, the full range of diversity, complexity, and information needs across all patient populations may not be captured by the selected questions, even though these were designed to reflect common patient inquiries about CBCT. As such, the scope of the included question set should be taken into account when interpreting these findings. Third, the reproducibility and generalizability of these findings may be limited, given that the chatbots assessed here undergo continuous updates that can alter their performance and response quality over time. Fourth, the evaluation was based on simulated patient inquiries rather than real-world patient interactions. Consequently, the actual comprehensibility, usefulness, and educational value of the responses from a patient perspective could not be directly assessed. Although inter-rater agreement was high, several assessment tools, including Accuracy, Usefulness, mGQS, CLEAR, and DISCERN, involve expert judgment and may therefore retain an inherent degree of subjectivity. A further limitation lies in the readability indices themselves. Metrics like the Flesch Reading Ease and Flesch–Kincaid Grade Level standardize the measurement of textual complexity, yet they do not account for factors such as health literacy, conceptual understanding, cultural background, or actual comprehension by patients. Additionally, this analysis was restricted to English-language material; readability outcomes and interpretive nuances of AI-generated content are likely to differ in other linguistic and cultural contexts.

CONCLUSION

Within the scope of this study, the evaluated AI-based conversational platforms generally achieved favorable scores across measures of accuracy, quality, readability, and educational value when responding to patient-oriented questions about CBCT. Performance varied across the platforms examined; in particular, more recently released models such as Gemini and DeepSeek-V3.1 tended to score higher across multiple assessment criteria. Nevertheless, no chatbot reliably generated content that fell within the 6th–8th grade readability band advised for patient-facing education materials, suggesting this may pose a challenge for individuals with lower health literacy.

A number of methodological limitations should be taken into account when interpreting these findings—namely, the specific question set used, reliance on expert-based assessment tools, the lack of testing with

real patients, and the fast-changing nature of large language models. While AI chatbots hold potential as supportive tools in patient education, information provision, and clinical assistance, these findings do not suggest that such systems are capable of delivering reliable clinical guidance on their own or standing in for professional expertise. Maintaining accuracy, appropriateness, and ethical standards in AI-generated health information still requires human review. Future research should incorporate real-world patient interactions, multilingual evaluations, and longitudinal assessments of evolving AI systems to better define their role in healthcare communication and education.

Generative AI statement

AI-assisted technologies (ChatGPT-3.5, OpenAI; Claude, Anthropic) were used solely for language editing and translation support during manuscript preparation; all scientific content, analysis, and interpretation remain the sole responsibility of the authors.

Ethical approval

This study, which did not involve human subjects, personal information, or animal experiments, was conducted entirely through AI-based data analysis using open and publicly available sources. Therefore, the authors state that ethical exemption was granted for this work, with the study carried out in line with the Declaration of Helsinki (2013).

Author contribution

Concept: SÖ, SCŞ; Design: SÖ, ÖSA; Data Collection or Processing: SÖ, SG, TP, ŞÇP; Analysis or Interpretation: SÖ, SCŞ, ÖSA; Literature Search: SÖ, SCŞ, ÖSA, SG, TP, ŞÇB; Writing: SÖ. All authors reviewed the results and approved the final version of the article.

Source of funding

The authors declare the study received no funding.

Conflict of interest

The authors declare that there is no conflict of interest.

REFERENCES

1. Ko YY, Yang WF, Leung YY. The role of Cone Beam Computed Tomography (CBCT) in the diagnosis and clinical management of Medication-Related Osteonecrosis of the Jaw (MRONJ). Diagnostics (Basel). 2024; 14(16): 1700. **[Crossref]**
2. Holmgren AJ, Byron ME, Grouse CK, Adler-Milstein J. Association between billing patient portal messages as e-visits and patient messaging volume. JAMA. 2023; 329(4): 339-42. **[Crossref]**
3. Kinikoglu I. Evaluating ChatGPT and Google Gemini performance and implications in Turkish dental education. Cureus. 2025; 17(1): e77292. **[Crossref]**
4. Miah ASM, Tusher MMR, Hossain MM, et al. ChatGPT in research and education: a SWOT analysis of its academic impact. CMES - Computer Modeling in Engineering and Sciences. 2025; 143(3): 2573-614. **[Crossref]**
5. Carlà MM, Gambini G, Baldascino A, et al. Large language models as assistance for glaucoma surgical cases: a ChatGPT vs. Google Gemini comparison. Graefes Arch Clin Exp Ophthalmol. 2024; 262(9): 2945-59. **[Crossref]**
6. Diniz-Freitas M, Diz-Dios P. DeepSeek: another step forward in the diagnosis of oral lesions. J Dent Sci. 2025; 20(3): 1904-7. **[Crossref]**
7. Semeraro F, Gamberini L, Carmona F, Monsieurs KG. Clinical questions on advanced life support answered by artificial intelligence. A comparison between ChatGPT, Google Bard and Microsoft Copilot. Resuscitation. 2024; 195: 110114. **[Crossref]**
8. Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. NeurIPS. 2020: 9459–9474.
9. Sallam M, Barakat M, Sallam M. Pilot testing of a tool to standardize the assessment of the quality of health information generated by artificial intelligence-based models. Cureus. 2023; 15(11): e49373. **[Crossref]**
10. 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): 2070-7. **[Crossref]**
11. Hatia A, Doldo T, Parrini S, et al. Accuracy and completeness of ChatGPT-Generated information on interceptive orthodontics: a multicenter collaborative study. J Clin Med. 2024; 13(3): 735. **[Crossref]**
12. Yeo YH, Samaan JS, Ng WH, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. Clin Mol Hepatol. 2023; 29(3): 721-32. **[Crossref]**
13. Helvacioglu-Yigit D, Demirturk 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): 750-9. **[Crossref]**

14. Hampton JR, Harrison MJ, Mitchell JR, Prichard JS, Seymour C. Relative contributions of history-taking, physical examination, and laboratory investigation to diagnosis and management of medical outpatients. Br Med J. 1975; 2(5969): 486-9. **[Crossref]**

15. Khan RA, Jawaid M, Khan AR, Sajjad M. ChatGPT - Reshaping medical education and clinical management. Pak J Med Sci. 2023; 39(2): 605-7. **[Crossref]**

16. Qaiser N, Arshad R. Pandora box of possibilities and limitations: unleashing the power of chatgpt in healthcare. Pak J Phsyiol. 2025; 21(3): 58-63. **[Crossref]**

17. Bashah A, Salem A, Al-Waqeerah A, et al. Evaluation of deepseek, gemini, ChatGPT-4o, and perplexity in responding to salivary gland cancer. BMC Oral Health. 2025; 25(1): 1358. **[Crossref]**

18. Hassona Y, Alqaisi D, Al-Haddad A, et al. How good is ChatGPT at answering patients' questions related to early detection of oral (mouth) cancer? Oral Surg Oral Med Oral Pathol Oral Radiol. 2024; 138(2): 269-78. **[Crossref]**

19. Esmailpour H, Rasaie V, Babaee 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. **[Crossref]**

20. Ceylan Şen S, Çardakci Bahar Ş, Saraç Atagün Ö, Ustaoğlu G, Yildiz ZH, Toker H. Quality and reliability of ai information on dental implant failure: a comparative multi-model analysis. J Craniofac Surg. 2026; 37(3-4): 675-80. **[Crossref]**

21. 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(3): 338-47. **[Crossref]**

22. Dursun D, Bilici Geçer R. Can artificial intelligence models serve as patient information consultants in orthodontics? BMC Med Inform Decis Mak. 2024; 24(1): 211. **[Crossref]**

23. Mohammad-Rahimi H, Khoury ZH, Alamdari MI, et al. Performance of AI chatbots on controversial topics in oral medicine, pathology, and radiology. Oral Surg Oral Med Oral Pathol Oral Radiol. 2024; 137(5): 508-14. **[Crossref]**

24. 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. **[Crossref]**

25. Yilmaz BE, Gokkurt Yilmaz BN, Ozbey F. Artificial intelligence performance in answering multiple-choice oral pathology questions: a comparative analysis. BMC Oral Health. 2025; 25(1): 573. **[Crossref]**

26. Çi Ti R M. ChatGPT and oral cancer: a study on informational reliability. BMC Oral Health. 2025; 25(1): 86. **[Crossref]**

27. Singal A, Goyal S. Comparative evaluation of AI platforms "Google Gemini 2.5 Flash, Google Gemini 2.0 Flash, DeepSeek V3 and ChatGPT 4o" in solving multiple-choice questions from different subtopics of anatomy. Surg Radiol Anat. 2025; 47(1): 193. **[Crossref]**

28. Sahin S, Erkmen B, Duymaz YK, Bayram F, Tekin AM, Topsakal V. Evaluating ChatGPT-4's performance as a digital health advisor for otosclerosis surgery. Front Surg. 2024; 11: 1373843. **[Crossref]**

29. Sezer B, Aydoğdu T. Performance of advanced artificial intelligence models in traumatic dental injuries in primary dentition: a comparative evaluation of ChatGPT-4 Omni, DeepSeek, Gemini Advanced, and Claude 3.7 in terms of accuracy, completeness, response time, and readability. Appl Sci. 2025; 15(14): 7778. **[Crossref]**

30. Taşyürek M, Adıgüzel Ö, Gündoğan M, Goncharuk-Khomyn M, Ortaç H. Comparative evaluation of the responses from ChatGPT-5, Gemini 2.5 Flash, and DeepSeek-V3.1 chatbots to patient inquiries about endodontic treatment in terms of accuracy, understandability, and readability. Int Dent Res. 2025; 15(3). **[Crossref]**