



Multidimensional Evaluation of AI-Generated Patient Information on Cone Beam–Computed Tomography

Melisa Öçbe¹ · Hülya Çerçi Akçay^{2,4} · Eda Sır³ · Çağan Taş³ · Rana Erbalta¹

Received: 12 February 2026 / Revised: 8 April 2026 / Accepted: 17 April 2026
© The Author(s) 2026

Abstract

This study aimed to conduct a multidimensional evaluation of artificial intelligence (AI) chatbot–generated patient information regarding cone beam–computed tomography (CBCT) in dentistry, with specific focus on readability, informational quality, reliability, and patient-centered suitability. Twenty frequently asked, patient-oriented questions related to CBCT were systematically identified from a public online forum. Each question was submitted to four large language model–based chatbots (ChatGPT-4o, Gemini Advanced, Claude Sonnet 4, and Microsoft Copilot) under standardized conditions. Generated responses were evaluated using validated instruments, including the DISCERN tool and Global Quality Scale (GQS) for information quality and reliability, as well as Flesch Reading Ease, Flesch–Kincaid Grade Level, and Gunning Fog Index for readability. Patient-centeredness was further assessed using PEMAT–Understandability and PEMAT–Actionability scores. Comparative analyses were performed using linear mixed-effects models. Significant differences were observed among chatbots across all evaluated domains ($p < 0.05$). While advanced models demonstrated higher informational quality and reliability, their responses frequently exceeded recommended health literacy thresholds. Readability, transparency, and actionability varied substantially between platforms. No chatbot consistently met all criteria for optimal patient-directed communication. AI chatbots can provide generally accurate information on CBCT; however, variability in readability, reliability, and educational suitability limits their standalone use for patient education. Careful integration with professional oversight is essential to ensure safe and accessible AI-supported communication in dentomaxillofacial radiology. This study provides the first multidimensional, comparative evaluation of leading AI chatbots in delivering patient-oriented information about cone beam–computed tomography. It shows critical gaps between informational accuracy and health literacy suitability. This study reveals the need for professional oversight when using AI for patient education in dentomaxillofacial radiology.

Keywords Large language models · Artificial intelligence · Cone beam–computed tomography · Oral radiology

Introduction

Large language models (LLMs), a form of generative artificial intelligence (Gen-AI), are trained on vast volumes of textual data and are capable of producing human-like responses, answering questions, and generating contextual recommendations based on user input [1, 2]. AI chatbots practical applications of LLMs are now widely accessible, with platforms such as ChatGPT-4o, Gemini Advanced, and Microsoft Copilot increasingly integrated into everyday life [2]. In parallel, industry-specific, task-oriented LLMs have been developed to improve workflow efficiency in specialized domains. Healthcare has emerged as a major area of interest, and the potential use of LLMs in patient guidance, diagnostics, education, and clinical reporting has been explored across multiple medical specialties, including

✉ Hülya Çerçi Akçay
hulya.akcay@kocaelisaglik.edu.tr; hlycrc@gmail.com

¹ Department of Oral and Maxillofacial Radiology, Faculty of Dentistry, Kocaeli Health and Technology University, Kocaeli 41275, Turkey

² Department of Pedodontics, Faculty of Dentistry, Kocaeli Health and Technology University, Kocaeli 41275, Turkey

³ Faculty of Dentistry, Kocaeli Health and Technology University, Kocaeli 41275, Turkey

⁴ Department of Oral and Maxillofacial Radiology, Faculty of Dentistry, Kocaeli Health and Technology University, Yeniköy Merkez, Ilıca Cd. No:29, Başiskele/Kocaeli 41275, Turkey

radiology [1–7]. Notably, LLM-based systems have shown the ability to triage symptoms, generate draft radiology reports, and even perform at near-clinician levels on certain standardized examinations [4–8]. More recently, multimodal vision–language models have emerged, combining image and text interpretation to strengthen diagnostic decision-making, improve the detection of conditions such as impacted teeth, and deliver clearer, patient-centered explanations of radiological findings and oral diseases, including oral cancer [9–12].

As the popularity and visibility of AI chatbots grow, patients increasingly rely on these tools to obtain medical information before, after, or even instead of consulting healthcare professionals. This trend is particularly relevant in fields where imaging plays a key role in diagnosis and treatment planning. In dentistry, CBCT has become a fundamental three-dimensional imaging modality, providing high-resolution visualization for implant planning, endodontics, orthodontics, temporomandibular joint evaluation, and dentoalveolar pathology assessment [5]. Because CBCT involves ionizing radiation, patients frequently seek reliable information about indications, risks, advantages, radiation dose levels, and comparisons with conventional imaging modalities [1, 2, 12].

However, despite the rapid adoption of LLMs and the increasing availability of AI-generated medical responses, there is limited evidence regarding the reliability, readability, and educational suitability of chatbot-provided information in dentomaxillofacial radiology [2, 4, 12]. While previous studies have compared chatbot responses to professional sources in Oral and Maxillofacial Radiology and found generally acceptable scientific adequacy, concerns remain regarding readability, and interpretability. Ensuring that patient-directed content is both accurate and understandable is critical for informed consent, radiation safety awareness, and patient–clinician communication.

In light of the increasing reliance on artificial intelligence–based chatbots by patients seeking information about CBCT, a critical evaluation of the reliability and educational adequacy of such content has become necessary. Although AI chatbots offer rapid and accessible explanations, the accuracy, readability, and clinical appropriateness of their responses—particularly in relation to radiation exposure, diagnostic indications, and patient risk perception—remain insufficiently explored within dental imaging contexts.

Accordingly, the present study aimed to evaluate the reliability, accuracy, readability, and educational quality of responses generated by large language model–based AI chatbots to patient-oriented questions concerning CBCT in dentistry. In doing so, the study compared chatbot performance under standardized query conditions, with specific attention to informational quality, reference transparency, and linguistic accessibility in clinically sensitive domains.

Based on this framework, it was hypothesized that no statistically significant differences would be observed among AI chatbots with respect to (i) the readability of responses to CBCT-related patient questions, (ii) the educational quality and reliability of the information provided, and (iii) the patient-centered suitability of the responses.

Methodology

Study Design

This cross-sectional analytical study investigated the quality, and readability of responses generated by four widely used LLM-based chatbots to patients questions related to CBCT. Data collection was conducted between 12 and 15 October 2025, and all chatbot outputs were obtained under controlled and standardized conditions.

Selection of Frequently Asked Questions

Frequently asked questions (FAQs) related to CBCT in dentistry were systematically collected from the *r/askdentists* subreddit on Reddit (<https://www.reddit.com/r/askdentists/>), a publicly accessible platform widely used by patients seeking dental advice and explanations regarding diagnostic and treatment-related procedures. This platform was selected due to its high user engagement, open access structure, and frequent patient-oriented discussions on dental imaging.

To minimize algorithmic personalization, user-specific content filtering, and platform-driven bias, data collection was conducted using a newly created Reddit account. This account had no prior browsing or interaction history, was not subscribed to any subreddits, and did not engage in posting, commenting, upvoting, or other user activities. All searches were performed using a standard web browser under default settings, with cookies, cache, and browsing history cleared before each search session.

A structured and predefined keyword-based search strategy was applied using terms including “*CBCT dental*,” “*cone beam CT risks*,” “*CBCT radiation dose*,” “*CBCT vs panoramic*,” and “*CBCT for implant planning*.” Search results were screened manually in a stepwise manner. Duplicate posts, commercially motivated content, off-topic discussions, and posts not reflecting patient-centered CBCT-related questions were excluded. Only text-based questions were included, while posts containing images, radiographs, videos, or external diagnostic material were excluded to ensure that chatbot-generated responses relied exclusively on textual input, thereby standardizing the information provided to the chatbots.

Eligibility screening was guided by explicit inclusion criteria. Questions were included if they:

- explicitly addressed patient concerns regarding the clinical use of CBCT,
- contained sufficient contextual detail to permit unambiguous interpretation,
- were written in English, and
- focused on diagnostic indications, radiation exposure, image quality or accuracy, potential risks, or comparisons with conventional dental radiographic techniques.

For all eligible questions, Reddit post identifiers, publication dates, and associated search keywords were recorded to support methodological transparency and reproducibility. All usernames, hyperlinks, and any potentially identifiable information were removed prior to analysis to ensure full anonymization and compliance with ethical research standards.

From the final dataset, 20 representative questions were selected. This sample size was chosen to ensure comprehensive thematic coverage of the principal domains of patient concern while maintaining methodological consistency and analytical feasibility for detailed comparative evaluation across multiple chatbot platforms and assessment instruments. The selected questions collectively represented key aspects of patient-oriented CBCT information needs, including radiation safety, diagnostic justification, comparative imaging performance, and perceived clinical risks. This approach is consistent with prior AI chatbot studies in healthcare, where the use of 10–20 representative patient-oriented questions has been considered sufficient to ensure thematic coverage and analytical feasibility [13, 14].

Chatbot Query Process

Each of the 20 selected patient-oriented questions was submitted to four LLM-based chatbots: ChatGPT-4o, Gemini Advanced, Claude Sonnet 4, and Microsoft Copilot, using a standardized input protocol (Table 1). All interactions were conducted through the official web interfaces of the respective platforms, and the most recent publicly

accessible versions of each chatbot available during the study period were used.

To reduce potential variability associated with dynamic model updates, server load fluctuations, or temporal instability, chatbot responses were collected over three consecutive days, during the same daily time window (09:00–12:00, GMT + 3). This repeated collection strategy enabled controlled assessment of response consistency while limiting the influence of transient system-level variations.

Prior to each data collection session, browser histories, caches, and cookies were cleared to minimize personalization and tracking effects. ChatGPT-4o, Gemini Advanced, and Claude Sonnet 4 were accessed without logging into user accounts. Microsoft Copilot was accessed via a neutral browser profile without prior activity. No browser extensions, plug-ins, or third-party tools were installed or used throughout the data collection process.

All questions were entered verbatim, without additional contextual information or follow-up prompts, in order to reflect how patients typically formulate questions in real-world online environments. Each chatbot was instructed to “*provide a clear and concise explanation for a patient*,” thereby establishing a consistent communication framework across all models. Responses were recorded immediately upon generation, without regeneration, repetition, or manual modification.

For each chatbot, the three responses generated for each question across the three collection sessions were stored as independent entries to preserve temporal variability. All outputs were exported in plain-text format and archived in a secure, time-stamped database. Associated metadata (including chatbot model, version, date and time of response generation, and session order) were systematically documented.

This structured response-collection protocol ensured consistency across models, minimized external influences on output variability, and enabled the generation of a repeated-measures dataset suitable for later quality, and readability analyses.

Table 1 Comparison of GQS scores according to AI chatbots

GQS score	ChatGPT-4o <i>n</i> (%)	Gemini Advanced <i>n</i> (%)	Microsoft Copilot <i>n</i> (%)	Claude Sonnet 4 <i>n</i> (%)	Test statistic	<i>p</i> -value
1	0 (0.0)	1 (2.3)	1 (2.3)	1 (2.3)	$\chi^2 = 22.42$	0.004
2	0 (0.0)	2 (4.7)	2 (4.7)	3 (7.0)		
3	0 (0.0)	2 (4.7)	4 (9.3)	5 (11.6)		
4	10 (23.3) ^a	19 (44.2) ^b	24 (55.8) ^b	22 (51.2) ^b		
5	33 (76.7) ^a	19 (44.2) ^b	13 (30.2) ^b	12 (27.9) ^b		

*Pearson chi-square test was applied to the overall distribution

^{a-b}Different superscript letters indicate statistically significant differences between AI chatbots after post hoc analysis

Evaluation Criteria

Evaluation of chatbot-generated responses was performed across two main domains: content quality, and readability. Each domain was assessed using validated and standardized instruments to ensure methodological rigor and minimize subjective bias.

Quality Assessment

The overall informational quality and educational value of chatbot responses were assessed using two validated instruments: the DISCERN tool and the Global Quality Scale (GQS).

The DISCERN instrument consists of 16 items that evaluate the reliability, transparency, and completeness of health information, including clarity of aims, citation of sources, and balanced presentation of management options. The GQS, a 5-point Likert scale (1 = poor quality, 5 = excellent quality), was employed to assess the clarity, coherence, and perceived usefulness of responses for patient education.

Prior to formal evaluation, inter-rater calibration sessions were conducted to ensure standardized interpretation and minimize subjective variability among examiners [15]. Both tools were independently applied by two calibrated reviewers. Prior to evaluation, calibration sessions were conducted to standardize scoring criteria. Disagreements were resolved through structured discussion, and when necessary, a third senior reviewer was consulted to reach consensus.

Readability Assessment

The linguistic accessibility of chatbot-generated responses was analyzed using three widely recognized readability metrics: the Flesch Reading Ease (FRE), the Flesch–Kincaid Grade Level (FKGL), and the Gunning Fog Index (GFI). These indices quantify textual complexity by considering average sentence length, number of syllables per word, and proportion of polysyllabic terms.

FRE The FRE score produces a value between 0 and 100, where higher scores indicate greater readability. A score above 60 represents text that is easily comprehensible for the general public [16]. The FRE score was calculated using the following formula:

$$\text{FRE} = 206.835 - (1.015 \times \text{total words}/\text{total sentences}) - (84.6 \times \text{total syllables}/\text{total words})$$

FKGL The FKGL index translates readability into the corresponding U.S. school grade level [17]. For instance, a score

of 8 suggests that the text is understandable by an eighth-grade student. The FKGL score was determined using the following equation:

$$\text{FKGL} = 0.39 \times (\text{total words}/\text{total sentences}) + 11.8 \times (\text{total syllables}/\text{total words}) - 15.59$$

GFI The GFI estimates the years of formal education required for a reader to comprehend the text on the first reading [18]. Values between 7 and 9 are considered appropriate for materials intended for general audiences. The GFI was computed using the following formula:

$$\text{GFI} = 0.4 \times [(\text{total words}/\text{total sentences}) + 100 \times (\text{complex words}/\text{total words})]$$

Collectively, these tools provide complementary and standardized assessments of text complexity. In accordance with the recommendations of the U.S. National Institutes of Health (NIH) and the American Medical Association (AMA), an eighth-grade reading level was adopted as the benchmark threshold for acceptable readability in patient-directed health communication materials [19].

Understandability and Actionability Assessment

The understandability and actionability of chatbot-generated responses were evaluated using the Patient Education Materials Assessment Tool for Printed Materials (PEMAT-P), a validated instrument designed to assess the suitability of patient education materials. The PEMAT-P includes items evaluating both understandability and actionability [20].

As all responses were text-based and did not include visual components, items related to audiovisual content were marked as “not applicable” and excluded from the analysis. All responses were assessed by a single trained reviewer in accordance with the PEMAT-P user manual to ensure methodological consistency across evaluations. Scores were assigned as 0 (“disagree”), 1 (“agree”), or NA (“not applicable”), and final scores were calculated as percentages, with higher values indicating greater understandability and actionability.

Statistical Analysis

Statistical analyses were performed using IBM SPSS Statistics (version 29.0) and R statistical software (version 4.x). The normality of continuous variables (DISCERN, GQS, and readability indices) was assessed using the Shapiro–Wilk test and Q–Q plots. Data are presented as mean $\pm$ standard deviation or median with interquartile range, as appropriate.

Given the repeated-measures design, linear mixed-effects models were used for all comparative analyses, with chatbot

specified as a fixed effect and question and collection session included as random intercepts. Model assumptions, including normality and homoscedasticity, were evaluated using Shapiro–Wilk tests and Q–Q plots. Tukey-adjusted post hoc tests were applied for pairwise comparisons when significant effects were detected. Effect sizes were expressed as partial eta-squared (η^2), and 95% confidence intervals were reported where applicable.

Inter-rater reliability for DISCERN and GQS was assessed using the intraclass correlation coefficient (ICC, two-way random-effects model, absolute agreement). All tests were two-tailed, and a p -value < 0.05 was considered statistically significant. Exact p -values were reported where available.

Results

The readability, quality, reliability, and patient-centered characteristics of AI-generated responses to CBCT-related patient questions were evaluated using multiple validated assessment tools. Readability outcomes are presented in Fig. 1, which illustrates significant differences among AI chatbots across all three indices: Flesch Reading Ease (FRE), Flesch–Kincaid Grade Level (FKGL), and Gunning Fog Index (GFI). FRE scores demonstrated statistically significant variation between chatbots, indicating differences in overall ease of reading (Fig. 1a). FKGL results showed that the required educational grade level of responses differed across models (Fig. 1b), while GFI scores revealed significant variation in linguistic complexity among AI-generated texts (Fig. 1c). Post hoc analyses identified statistically significant pairwise differences between specific chatbot outputs across these readability measures ($p < 0.05$).

The distribution of Global Quality Scale (GQS) scores according to AI chatbot is summarized in Table 1. A statistically significant difference in GQS score distribution was

observed among chatbots ($\chi^2 = 22.42$, $p = 0.004$). The proportion of responses rated at the highest quality level (GQS = 5) varied significantly across models, as reflected by the frequency-based comparison. Differences between chatbots were further supported by post hoc analyses, as indicated by distinct superscript groupings within score categories.

Reliability of information was assessed using DISCERN scores, with comparative results presented in Table 2. Linear mixed-effects model analysis demonstrated a statistically significant effect of chatbot on DISCERN scores ($p < 0.001$). Mean DISCERN scores, accompanied by standard deviations and 95% confidence intervals, differed across AI chatbots, with effect size estimates indicating a moderate to large magnitude of difference (partial $\eta^2 = 0.28$). Post hoc comparisons identified statistically significant differences between multiple chatbot pairs, as indicated by superscript lettering (Fig. 2).

Patient-centered quality outcomes assessed using PEMAT are presented in Tables 3 and 4. PEMAT-Understandability (PEMAT-U) scores differed significantly among AI chatbots ($p < 0.001$), with mean scores and corresponding confidence intervals demonstrating variability

Table 2 Comparison of DISCERN scores according to AI chatbots

AI Chatbot	DISCERN score (mean $\pm$ SD)	95% CI	Post hoc group
Gemini Advanced	72.08 $\pm$ 5.82	71.2–72.9	a
Microsoft Copilot	70.06 $\pm$ 4.76	69.4–70.7	a,b
ChatGPT-4o	68.88 $\pm$ 6.41	68.0–69.7	b
Claude Sonnet 4	64.64 $\pm$ 6.91	63.7–65.6	c

Values are presented as mean $\pm$ standard deviation with 95% confidence intervals. Different superscript letters indicate statistically significant differences between AI chatbots based on linear mixed-effects models with Tukey-adjusted post hoc comparisons. The effect size of the chatbot factor was quantified using partial eta-squared (η^2)

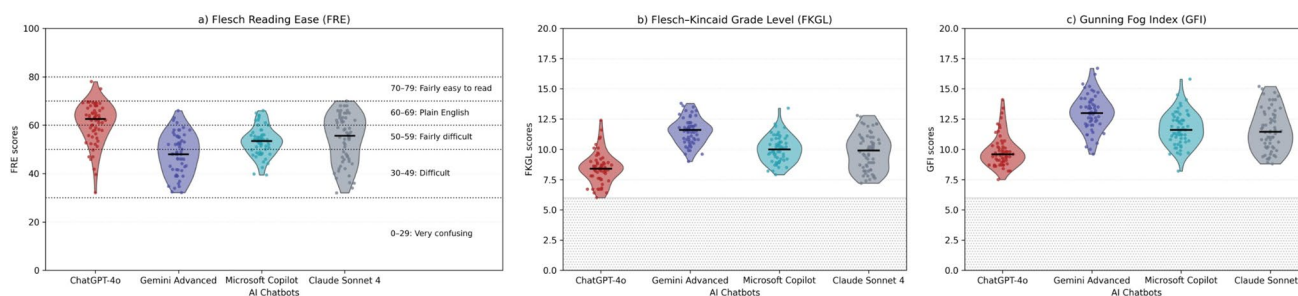


Fig. 1 Readability of AI chatbot responses to CBCT-related patient questions. **a** Flesch Reading Ease (FRE) scores. Higher values indicate easier readability. **b** Flesch–Kincaid Grade Level (FKGL) scores. Lower values indicate lower educational grade levels required for comprehension. **c** Gunning Fog Index (GFI) scores. Lower values

reflect reduced linguistic complexity. Violin plots show score distributions, with dots representing individual responses. Statistical comparisons were conducted using linear mixed-effects models with Tukey-adjusted post hoc tests ($p < 0.05$; $p < 0.01$; ns, not significant)

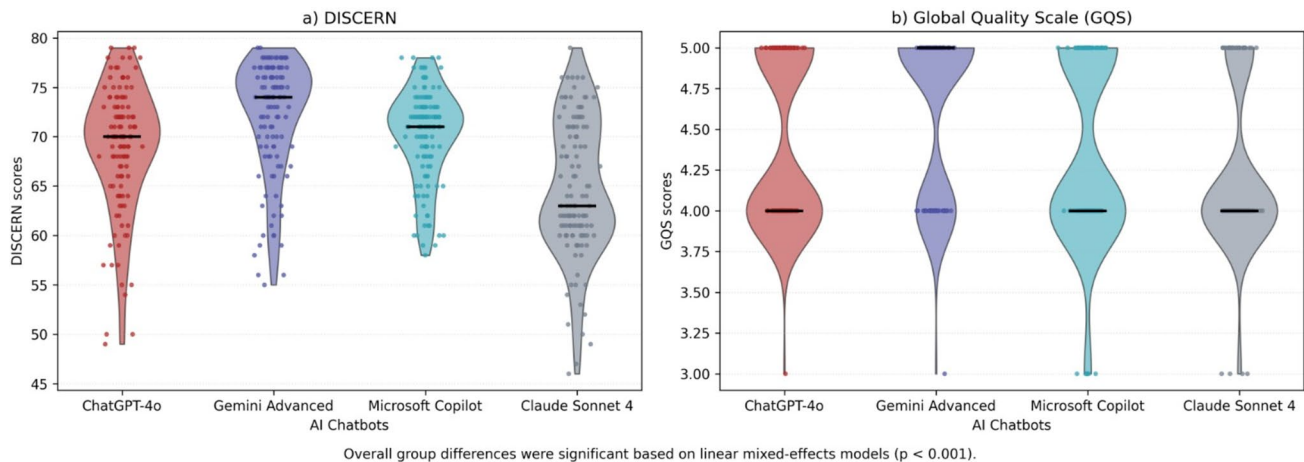


Fig. 2 Quality of AI chatbot responses to CBCT-related patient questions. **a** DISCERN scores assessing information reliability and quality. **b** Global Quality Scale (GQS) scores reflecting overall educa-

tional quality. Violin plots display score distributions with individual responses overlaid. Overall group differences were significant based on linear mixed-effects models ($p < 0.001$)

Table 3 Comparison of PEMAT-Understandability (PEMAT-U) scores according to AI chatbots

AI Chatbot	PEMAT-U score (mean $\pm$ SD)	95% CI	Post hoc group
ChatGPT-4o	86.4 $\pm$ 6.8	85.5–87.3	a
Gemini Advanced	82.1 $\pm$ 7.4	81.1–83.1	b
Microsoft Copilot	80.3 $\pm$ 7.1	79.4–81.3	b
Claude Sonnet 4	76.8 $\pm$ 8.2	75.7–77.9	c

Values are presented as mean $\pm$ standard deviation with 95% confidence intervals. Different superscript letters indicate statistically significant differences between AI chatbots based on linear mixed-effects models with Tukey-adjusted post hoc comparisons. Effect size was expressed as partial eta-squared (η^2)

Table 4 Comparison of PEMAT-Actionability (PEMAT-A) scores according to AI chatbots

AI Chatbot	PEMAT-A score (mean $\pm$ SD)	95% CI	Post hoc group
ChatGPT-4o	71.2 $\pm$ 9.6	69.9–72.5	a
Microsoft Copilot	67.8 $\pm$ 10.1	66.4–69.2	a,b
Gemini Advanced	65.9 $\pm$ 9.8	64.6–67.2	b
Claude Sonnet 4	61.3 $\pm$ 11.2	59.8–62.8	c

Values are presented as mean $\pm$ standard deviation with 95% confidence intervals.. Different superscript letters indicate statistically significant differences between AI chatbots based on linear mixed-effects models with Tukey-adjusted post hoc comparisons. Effect size was expressed as partial eta-squared (η^2)

in the comprehensibility of responses. Similarly, PEMAT-Actionability (PEMAT-A) scores showed statistically significant differences between chatbots ($p < 0.001$), with moderate effect sizes observed for both PEMAT-U ($\eta^2 = 0.22$) and PEMAT-A ($\eta^2 = 0.19$). Pairwise differences were identified through Tukey-adjusted post hoc analyses (Fig. 3).

Inter-rater reliability analysis is summarized in Table 5. Agreement between reviewers was moderate for DISCERN assessments (ICC = 0.45; 95% CI 0.35–0.54) and poor to fair for GQS assessments (ICC = 0.20; 95% CI 0.06–0.32), based on a two-way random-effects model with absolute agreement.

Discussion

The present study provides a comprehensive evaluation of the readability, quality, reliability, and accuracy of responses generated by contemporary AI chatbots to CBCT-related

patient questions, with a particular focus on differences across platforms. Consistent with recent evidence, our findings indicate that although large language models are capable of producing largely accurate and structured responses, substantial variability exists in their suitability for patient-centered education, especially with regard to readability and transparency [12–14].

Readability remains a cornerstone of effective health communication, particularly in dentistry, where information must be accessible to a broad patient population. International health literacy guidelines recommend that patient education materials be written at or below a sixth-grade reading level to ensure comprehension and appropriate decision-making [21]. In the present study, FRE and FKGL analyses demonstrated that ChatGPT-4o generated significantly more complex responses compared with Gemini Advanced and Microsoft Copilot, frequently exceeding recommended readability thresholds. This observation is in line with previous

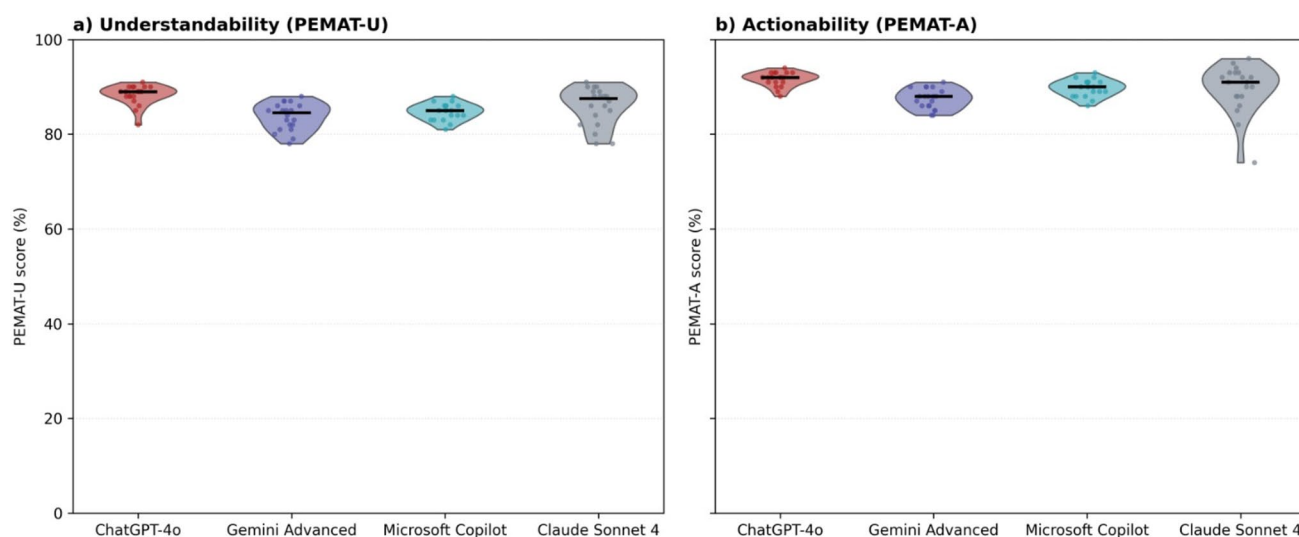


Fig. 3 Patient-centered quality of AI chatbot responses to CBCT-related patient questions. **a** PEMAT-Understandability (PEMAT-U) scores and **b** PEMAT-Actionability (PEMAT-A) scores. Violin plots

Table 5 Inter-observer agreement between reviewers

	ICC (2,1)	95% CI	Interpretation
GQS	0.20	0.06–0.32	Poor to fair agreement
DISCERN	0.45	0.35–0.54	Moderate agreement

Inter-observer agreement was assessed using the intraclass correlation coefficient (ICC), applying a two-way random-effects model with absolute agreement (ICC [1, 2]). Ninety-five percent confidence intervals were obtained using bootstrap resampling. PEMAT-Understandability and PEMAT-Actionability scores were assessed by a single trained reviewer following standardized scoring guidelines. Therefore, inter-rater reliability or internal consistency analyses were not applicable for PEMAT measures

investigations reporting that newer GPT-based models, while more informative, tend to produce linguistically dense outputs that may limit accessibility [13]. Similar readability challenges have been reported across orthodontics, pediatric dentistry, and oral radiology, suggesting that this limitation is not topic-specific but rather inherent to current LLM architectures [22–25].

Despite its lower readability, ChatGPT-4o achieved higher Global Quality Scale scores, indicating superior informational completeness and clinical coherence. This pattern parallels findings reported by Güven et al., who demonstrated that ChatGPT-4o outperformed earlier models and alternative platforms in terms of overall quality and accuracy when responding to dental trauma-related patient queries [20]. Comparable trends have also been observed in studies evaluating AI performance in endodontics and prosthodontics, where depth of explanation and clinical relevance were prioritized over linguistic simplicity [24, 25]. Collectively,

these findings suggest a trade-off between informational richness and readability, with more advanced models excelling in the former but lagging in the latter.

depict score distributions with individual responses overlaid. Higher scores indicate greater understandability and actionability

these findings suggest a trade-off between informational richness and readability, with more advanced models excelling in the former but lagging in the latter. An important dimension highlighted by our results concerns transparency and source attribution. While ChatGPT-4o produced highly accurate responses, its limited provision of explicit references negatively affected DISCERN scores. In contrast, Gemini Advanced consistently achieved higher DISCERN values, largely due to more frequent citation of external sources. This finding aligns with prior reports indicating that Gemini tends to prioritize reference inclusion [20]. However, existing literature also cautions that AI-generated references may be incomplete, unverifiable, or inaccurately attributed, emphasizing that apparent transparency does not necessarily equate to reliability. As such, the inclusion of citations should be interpreted cautiously and does not obviate the need for expert validation.

Microsoft Copilot demonstrated comparatively better readability but lower quality and reliability scores, reflecting a tendency toward brevity and simplified language. Similar observations have been reported in studies assessing AI-generated responses in breastfeeding counseling, refractive surgery, and dental education, where conversational tone improved readability but resulted in omission of clinically relevant details [26, 27]. In clinical practice, where individualized risk communication is essential, such omissions may reduce the practical value of AI-generated information.

The findings further highlight the relative robustness of advanced AI chatbots, particularly in terms of providing structured and generally accurate information. Nevertheless, minor inaccuracies and incomplete explanations

were observed across all platforms, consistent with previous studies indicating that even high-performing AI systems may generate potentially misleading or insufficient guidance when clinical nuance is required [23, 28]. These observations underscore the importance of positioning AI chatbots as adjunctive tools rather than substitutes for professional consultation.

A major strength of this study is the use of a comprehensive, multidimensional evaluation framework combining readability indices, quality and reliability scales, and patient-centered assessment tools. This approach allows a more nuanced appraisal of AI-generated health information beyond accuracy-focused analyses alone. In addition, the repeated-measures design, in which each question was assessed across multiple sessions and platforms, enabled the use of linear mixed-effects modeling to account for within-question variability and enhance statistical robustness. The assessment of inter-rater reliability further strengthens the objectivity of the findings.

Several limitations should be acknowledged. Only widely available, general-purpose AI chatbots were evaluated, and domain-specific dental AI systems were not included. Moreover, chatbot responses were assessed at a single time period and in English only, which may limit generalizability given the dynamic nature of AI models and linguistic diversity. Additionally, the use of a single data source (Reddit) may introduce selection bias and limit external validity, as patient questions may differ from those presented in clinical FAQs, institutional websites, or other online platforms. The relatively low inter-rater agreement for GQS may be attributed to the subjective nature of global quality assessment; however, the use of multiple complementary evaluation tools supports the robustness of the findings. In addition, PEMAT scores were assessed by a single reviewer, which may introduce subjectivity and should be interpreted with caution. The absence of a structured misinformation assessment represents a limitation; inaccurate or incomplete information, particularly regarding radiation exposure, may influence patient perceptions and decision-making. Finally, the behavioral impact of AI-generated responses on patient decision-making was not assessed and warrants further investigation.

Conclusion

This study shows that contemporary AI chatbots are capable of providing largely accurate and structured information in response to CBCT-related patient questions; however, marked differences persist in readability, reliability, and transparency across platforms. Although advanced models offer greater informational depth, their outputs frequently exceed recommended health literacy levels,

limiting their suitability for routine patient education. These findings highlight the need for cautious integration of AI chatbots into oral health communication frameworks. From a policy perspective, the development, regulation, and clinical use of AI-based patient information tools should incorporate standardized readability thresholds, transparent referencing practices, and clinically relevant safeguards to ensure that CBCT-related information is both accurate and accessible.

Acknowledgements None declared.

Author Contribution We confirm that the contents of this manuscript are original and have not been published or submitted elsewhere. All authors have contributed significantly to the conception, drafting, and revision of the manuscript and have approved the final version. They meet the authorship criteria as they planned the work, contributed to writing and reviewing the manuscript, and approved the final version.

Funding Open access funding provided by the Scientific and Technological Research Council of Türkiye (TÜBİTAK).

Data Availability The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Declarations

Informed Consent and Ethical Approval As this study did not involve human participants, patient data, or clinical interventions, ethical approval was not required.

Conflict of interest The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Liu, Z., Nalley, A., Hao, J., et al.: The performance of large language models in dentomaxillofacial radiology: a systematic review. *Dentomaxillofac Radiol.* Published online August 12 (2025). <https://doi.org/10.1093/dmfr/twaf060>
2. 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.* 139(6), 750–759 (2025). <https://doi.org/10.1016/j.oooo.2025.01.001>
3. Mohammad-Rahimi, H., Khoury, Z.H., Alamdari, M.I., et al.: Performance of AI chatbots on controversial topics in oral

- medicine, pathology, and radiology. *Oral Surg Oral Med Oral Pathol Oral Radiol*, 137(5), 508–514 (2024). <https://doi.org/10.1016/j.oooo.2024.01.015>
4. Jeong, H., Han, S.S., Yu, Y., Kim, S., Jeon, K.J.: How well do large language model-based chatbots perform in oral and maxillofacial radiology? *Dentomaxillofac Radiol*, 53(6), 390–395 (2024). <https://doi.org/10.1093/dmfr/twae021>
 5. Katsumata, A.: Deep learning and artificial intelligence in dental diagnostic imaging. *Jpn Dent Sci Rev*, 59, 329–333 (2023). <https://doi.org/10.1016/j.jdsr.2023.09.004>
 6. Stephan, D., Bertsch, A.S., Schumacher, S., et al.: Improving patient communication by simplifying AI-generated dental radiology reports with ChatGPT: comparative study. *J Med Internet Res*, 27, e73337 (2025). <https://doi.org/10.2196/73337>
 7. Tassoker, M.: ChatGPT-4 Omni's superiority in answering multiple-choice oral radiology questions. *BMC Oral Health*, 25(1), 173 (2025). <https://doi.org/10.1186/s12903-025-05554-w>
 8. Arpacı, A., Ozturk, A.U., Okur, I., Sadry, S.: Evaluation of the accuracy of ChatGPT-4 and Gemini's responses to the World Dental Federation's frequently asked questions on oral health. *BMC Oral Health*, 25(1), 1293 (2025). <https://doi.org/10.1186/s12903-025-06624-9>
 9. Çi, T.R.M.: ChatGPT and oral cancer: a study on informational reliability. *BMC Oral Health*, 25(1), 86 (2025). <https://doi.org/10.1186/s12903-025-05479-4>
 10. Stephan, D., Bertsch, A., Burwinkel, M., et al.: AI in dental radiology-improving the efficiency of reporting with ChatGPT: comparative study. *J Med Internet Res*, 26, e60684 (2024). <https://doi.org/10.2196/60684>
 11. Grinberg, N., Ianculovici, C., Whitefield, S., Kleinman, S., Feldman, S., Peleg, O.: Artificial intelligence-assisted decision-making in third molar assessment using ChatGPT: is it really a valid tool? *Oral Radiol*. Published online June 20 (2025). <https://doi.org/10.1007/s11282-025-00836-z>
 12. Russe, M.F., Rau, A., Ermer, M.A., et al.: A content-aware chatbot based on GPT 4 provides trustworthy recommendations for cone-beam CT guidelines in dental imaging. *Dentomaxillofac Radiol*, 53(2), 109–114 (2024). <https://doi.org/10.1093/dmfr/twad015>
 13. Özdemir, Ö.T., Kavan, M.Y., Güven, Y.: Evaluation of the readability, quality, and accuracy of AI chatbot responses to questions about deleterious oral habits. *BMC Oral Health*, 25(1), 1812 (2025). <https://doi.org/10.1186/s12903-025-07298-z>
 14. Aydın, F.O., Aksoy, B.K., Ceylan, A., et al.: Readability and appropriateness of responses generated by ChatGPT 3.5, ChatGPT 4.0, Gemini, and Microsoft Copilot for FAQs in refractive surgery. *Turk J Ophthalmol*, 54(6), 313–317 (2024). <https://doi.org/10.4274/tjo.galenos.2024.28234>
 15. Cronbach, L.J.: Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334 (1951). <https://doi.org/10.1007/BF02310555>
 16. Flesch, R.: A new readability yardstick. *J Appl Psychol*, 32, 221–233 (1948). <https://doi.org/10.1037/h0057532>
 17. Kincaid, J.P., Fishburne, R.P., Rogers, R.L., Chissom, B.S.: Derivation of new readability formulas (automated readability index, fog count, and Flesch reading ease formula) for Navy enlisted personnel. Naval Air Station Memphis (1975). <https://apps.dtic.mil/sti/citations/ADA006655>
 18. Izgi, U., Sezginsoy Seker, B.: Comparing different readability formulas on the examples of science-technology and social science textbooks. *Procedia Soc Behav Sci*, 46, 178–182 (2012). <https://doi.org/10.1016/j.sbspro.2012.05.089>
 19. National Institutes of Health: How to write easy-to-read health materials. MedlinePlus. https://medlineplus.gov/all_easytoread.html
 20. Guven, Y., Ozdemir, O.T., Kavan, M.Y.: Performance of artificial intelligence chatbots in responding to patient queries related to traumatic dental injuries: a comparative study. *Dent Traumatol* (2024). <https://doi.org/10.1111/edt.12976>
 21. Sørensen, K., Pelikan, J.M., Röthlin, F., Ganahl, K., Slonska, Z., Doyle, G., Fullam, J., Kondilis, B., Agraftotis, D., Ueters, E., Falcon, M., Mensing, M., Tchamov, K., van den Broucke, S., Brand, H.: Health literacy in Europe: comparative results of the European health literacy survey (HLS-EU). *Eur J Public Health*, 25(6), 1053–1058 (2015). <https://doi.org/10.1093/eurpub/ckv043>
 22. Camargo, E.S., Quadras, I.C.C., Garanhani, R.R., de Araujo, C.M., Stuginski-Barbosa, J.: A comparative analysis of three large language models on bruxism knowledge. *J Oral Rehabil*, 52(6), 896–903 (2025). <https://doi.org/10.1111/joor.13948>
 23. Dermata, A., Arhakis, A., Makrygiannakis, M.A., Giannakopoulos, K., Kaklamanos, E.G.: Evaluating the evidence-based potential of six large language models in paediatric dentistry: a comparative study on generative artificial intelligence. *Eur Arch Paediatr Dent*, 26(3), 527–535 (2025). <https://doi.org/10.1007/s40368-025-01012-x>
 24. Guven, Y., Ozdemir, O.T., Kavan, M.Y.: Performance of artificial intelligence chatbots in responding to patient queries related to traumatic dental injuries: a comparative study. *Dent Traumatol*, 41(3), 338–347 (2025). <https://doi.org/10.1111/edt.13020>
 25. 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*, 25(1), 871 (2025). <https://doi.org/10.1186/s12903-025-06267-w>
 26. Gheisarifar, M., Shembesh, M., Koseoglu, M., et al.: Evaluating the validity and consistency of artificial intelligence chatbots in responding to patients' frequently asked questions in prosthodontics. *J Prosthet Dent*, 134(1), 199–206 (2025). <https://doi.org/10.1016/j.prosdent.2025.03.009>
 27. Kacer, E.O.: Evaluating AI-based breastfeeding chatbots: quality, readability, and reliability analysis. *PLoS One*, 20(3), e0319782 (2025). <https://doi.org/10.1371/journal.pone.0319782>
 28. Oviedo-Trespalacios, O., Peden, A.E., Cole-Hunter, T., et al.: The risks of using ChatGPT to obtain common safety-related information and advice. *Safety Sci*, 167, 106244 (2023). <https://doi.org/10.1016/j.ssci.2023.106244>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.