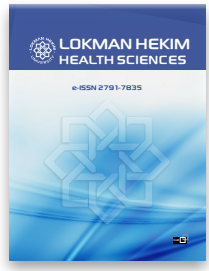
LOKMAN HEKIM HEALTH SCIENCES

DOI: 10.14744/lhhs.2026.25363
Lokman Hekim Health Sci 2026;6(2):246–254



lokmanhekimhs.com

ORIGINAL ARTICLE

Assessment of Artificial Intelligence Chatbots' Information Quality on Home Dental Bleaching: A Comparative Study

Özlem Özişçi, Elif Irmak Gökmen

Department of Prosthodontics, Süleyman Demirel University Faculty of Dentistry, Isparta, Türkiye

Abstract

Introduction: The purpose of this research was to compare how artificial intelligence chatbots (ChatGPT-3.5, ChatGPT-4, Gemini, and DeepSeek) responded to common patient inquiries regarding home dental bleaching, with an emphasis on quality, accuracy, clarity, and practical applicability.

Methods: Forty patient-oriented questions identified using the AlsoAsked tool, which extracts Google "People Also Ask" data, were categorized into seven thematic domains and submitted individually to each chatbot in separate sessions. Responses were independently scored by two evaluators using the global quality scale (GQS), accuracy of ınformation ındex (AOI), and patient education materials assessment tool for printed materials. Response times were recorded in seconds. Statistical analyses included the Kruskal–Wallis test, Bonferroni-adjusted pairwise comparisons, and Spearman correlation ($p<0.05$).

Results: ChatGPT-4 and DeepSeek achieved the highest GQS and AOI scores. DeepSeek had the highest actionability score but the longest response time. ChatGPT-3.5 demonstrated moderate performance, while Gemini had the lowest intelligibility and actionability scores.

Discussion and Conclusion: Advanced artificial intelligence chatbots can provide high-quality and accurate information on at-home dental bleaching. However, unsupervised use may pose patient safety risks; thus, their deployment should be limited to validated, monitored, and task-specific applications.

Keywords: Artificial intelligence; Chatbot; Patient education; Tooth bleaching

Artificial intelligence (AI) includes computational techniques that process large datasets, recognize patterns, and improve task performance over time.[1] These approaches from machine learning to deep-learning models enable automated feature extraction, predictive modeling, and decision support in clinical contexts.[2] In dentistry, AI applications have enhanced diagnostic consistency in radiographic interpretation, caries detection, and treatment planning.[2–5] Although performance varies with task complexity and data quality, validated models can reliably perform well-defined dental tasks under clinician supervision.[1,2]

Chatbots based on large language models (LLMs) extend AI's use by providing scalable, on-demand patient education and guidance.[6–8] Beyond patient communication, LLMs support evidence summarization and administrative tasks,

Cite this article as: *Özişçi Ö, Gökmen EI. Assessment of Artificial Intelligence Chatbots' Information Quality on Home Dental Bleaching: A Comparative Study. Lokman Hekim Health Sci 2026;6(2):246–254.*

Correspondence: Özlem Özişçi, M.D. Süleyman Demirel Üniversitesi, Diş Hekimliği Fakültesi, Protetik Diş Hekimliği Anabilim Dalı, Isparta, Türkiye
E-mail: oslemozisci@gmail.com **Submitted:** 20.10.2025 **Revised:** 28.03.2026 **Accepted:** 16.05.2026 **Available Online:** 02.06.2026

but may also produce inaccurate or fabricated information, posing safety risks if unsupervised.[9–13] Therefore, safe use requires task-specific validation and human oversight.

Dental bleaching is among the most common esthetic procedures.[14,15] Patients frequently request information about efficacy, safety, and risks.[14,16] While dentists conventionally provide such knowledge, AI chatbots are increasingly consulted as alternative sources.[13,17] However, their outputs may lack scientific rigor and completeness.[13,17,18] Given that bleaching outcomes depend on product choice, patient suitability, and post-treatment care, the accuracy of information is critical.[16,19]

Accordingly, a systematic evaluation of AI chatbots' bleaching-related responses is needed. Assessing quality, accuracy, readability, and guideline consistency will clarify current capabilities and limitations.[13,17] In this study, we aimed to comparatively evaluate the quality, accuracy, readability, and guideline consistency of responses generated by four AI chatbots to common patient questions on at-home dental bleaching. The null hypothesis of this study was that there would be no statistically significant differences among the evaluated AI chatbots in terms of information quality (Global Quality Scale [GQS]), accuracy of information index (AOI), intelligibility, actionability, and response duration when answering patient-oriented questions about at-home dental bleaching.

Materials and Methods

Study Place and Design

This study was conducted at the department of prosthodontics, faculty of dentistry. The research process included question identification, chatbot response collection, and evaluator scoring within the same period.

Type of Research

This was a cross-sectional, comparative, and descriptive study designed to evaluate and compare the quality of information provided by four AI chatbots on the topic of home dental bleaching.

Population and Sample

The study population consisted of AI Chatbot-generated responses rather than human participants. Responses from four AI chatbots (ChatGPT-3.5, ChatGPT-4, Gemini, DeepSeek) were compared. A total of 40 patient-oriented questions on at-home bleaching were selected to represent the most frequently searched public inquiries. The questions were generated using the AlsoAsked platform (https://alsoasked.com), which aggregates Google "People Also Ask" queries derived from Google Search results (https://www.google.com). Specifically, queries were retrieved by entering the keywords "home dental bleaching" and related search terms into the Google search engine, and the associated "People Also Ask" suggestions were systematically extracted through AlsoAsked on July 10, 2025. Because Google search results are dynamically updated and may vary by user location, time, and search history, the retrieval date and keyword strategy were documented to enhance reproducibility. All extracted questions were recorded verbatim at the time of data collection.

The four AI chatbots evaluated in this study (ChatGPT-3.5, ChatGPT-4, Gemini, and DeepSeek) were selected based on predefined criteria to ensure representativeness and methodological rigor. First, these platforms were among the most widely accessible and publicly used LLM-based chatbots at the time of data collection, reflecting real-world tools that patients are likely to consult for health-related information. Second, the selected models represent different developers and architectural frameworks (OpenAI, Google, and DeepSeek AI), allowing cross-platform comparison across systems trained on distinct datasets and alignment strategies. Third, prior healthcare and dentistry-related comparative studies have evaluated these models, establishing their relevance in clinical and patient education contexts. Including both earlier generation (ChatGPT-3.5) and more advanced models (ChatGPT-4, Gemini, and DeepSeek) also allowed evaluation of potential performance differences across model generations. Therefore, the selection of these four chatbots was intended to ensure a representative, cross-platform, and clinically relevant comparison of widely used LLM-based systems with differing architectures, generations, and prior healthcare applicability.

Questionnaire Structure

A total of 40 questions on at-home bleaching were categorized into seven domains: General definition and mechanism, eligibility, safety, effectiveness, causes, application, and post-treatment maintenance. Most of the questions were related to safety and effectiveness, highlighting patients' primary concerns regarding the use of whitening products at home, as shown in Table 1.[20,21]

Data Collection

All chatbot responses were collected between July 11 and July 15, 2025, using the publicly accessible

Table 1. Classification of home bleaching questions

Category	Questions
General definition and mechanism	What is home bleaching, and how does it work?
	How does light activation (e.g., LED lights) enhance home bleaching?
	What is the difference between hydrogen peroxide (HP) and carbamide peroxide (CP) in whitening gels?
	What does the percentage on whitening gels mean (e.g., 10%, 16%, 22%)?
	How to make teeth white naturally from yellow?
Eligibility and contraindications	Is home bleaching suitable for everyone?
	Who should avoid it?
	Can children or teenagers use home whitening products?
	Can pregnant or breastfeeding women undergo home bleaching?
	Can people with braces (orthodontic treatment) use whitening products?
	Can you bleach rotten teeth?
Safety and side effects	Is it safe to whiten your teeth at home?
	Is CP safe for teeth?
	Is it safe to put HP directly on your teeth?
	Is it good to brush your teeth with HP every day?
	What percentage of peroxide do dentists use to whiten teeth?
	Is it safe to use high-concentration gels at home?
	Is there a risk of over-bleaching?
	Can teeth whitening crack your teeth?
	Does enamel grow back?
	What are the side effects of household bleach?
	How do I manage gum irritation caused by whitening gel?
Effectiveness and expectations	Is in-office teeth whitening better than take-home?
	Do crest white strips work?
	Do whitening strips contain the same active ingredients as whitening gels?
	How many shades whiter can I expect my teeth to become with home bleaching? Does home bleaching produce the same results for every person?
	Are natural remedies (e.g., baking soda, coconut oil) as effective as home bleaching products?
Causes and pre-treatment conditions	Why are my teeth so yellow?
	What chemical whitens teeth?
Application frequency and duration	How often should I whiten my teeth?
	How many days can you use home bleaching?
	How long does it take to whiten teeth with 6 HP?
	How long does it take to see noticeable results?
	How long does home bleaching last?

web-based interfaces of each platform. All platforms were accessed through their publicly available web interfaces using standard user-level accounts without API access, developer-level permissions, or institutional custom configurations. No premium API integrations, external plugins, or experimental features were enabled during data collection. The versions available on the respective web platforms at the time of access were used without modification. The following model versions were used during the study: ChatGPT-3.5 (OpenAI, GPT-3.5-turbo), ChatGPT-4 (OpenAI, GPT-4), Google Gemini (Google AI, standard publicly available web version), and DeepSeek (DeepSeek LLM, publicly accessible web interface version available at the time of data collection).

All platforms were accessed under default system configurations. No modifications were made to temperature settings, output length parameters, response tone, or formatting options. Each question was entered in a separate new session to minimize contextual carryover effects. No additional prompts, clarifications, or follow-up instructions were provided. Only the first complete response generated by each chatbot was recorded verbatim and included in the analysis without editing, summarization, or structural modification. This approach was adopted to enhance reproducibility and to reflect typical real-world patient interactions with these systems.

Evaluation of Responses

Two independent evaluators assessed chatbot outputs using:

GQS

The GQS is a five-point Likert-type instrument used to evaluate the overall quality, flow, comprehensiveness, and usefulness of health-related information. Scores range from 1 (poor quality, incomplete, and not useful) to 5 (excellent quality, well-structured, and highly useful). Higher scores indicate better overall informational quality.

AOI

A 10-point tool evaluating factual correctness, consistency, and relevance. The AOI assessment was conducted using a predefined evidence-based reference framework. Authoritative professional sources were systematically used as benchmarks, including the American Dental Association guidelines on tooth whitening, the FDI World Dental Federation policy statement on bleaching, and peer-reviewed scientific literature addressing home dental bleaching indications, contraindications, mechanisms, and safety considerations. Before scoring, the evaluators reviewed these reference documents to establish a consensus understanding of accepted clinical standards and evidence-based recommendations. Each chatbot response was then compared against these benchmark sources to determine factual correctness, consistency with established guidelines, clinical appropriateness, and completeness of information. When discrepancies existed between sources, priority was given to the most recent evidence-based clinical guideline. This structured benchmarking approach minimized subjective interpretation and enhanced methodological rigor.

Patient Education Materials Assessment Tool (PEMAT-P)

The PEMAT-P was applied in accordance with the original scoring manual developed by Shoemaker et al.[22] Each item within the understandability and actionability domains was independently evaluated by two trained reviewers. Items were scored dichotomously as "agree" (1 point) or "disagree" (0 points), while items deemed "not applicable" were excluded from the denominator, as recommended in the official PEMAT-P instructions. Percentage scores for understandability and actionability were calculated by dividing the total number of items scored as "agree" by the total number of applicable items and multiplying the result by 100. This standardized percentage approach enabled direct comparison across chatbot responses despite minor variations in structural formatting. Before formal scoring, both evaluators conducted a calibration session using pilot responses to ensure consistent interpretation of PEMAT-P criteria. Discrepancies in scoring were resolved through structured consensus discussions. Inter-rater reliability for PEMAT-P demonstrated good-to-excellent agreement.

Inter-rater reliability for PEMAT-P percentage scores was specifically assessed using the intraclass correlation coefficient (ICC) under a two-way random-effects model with absolute agreement. The ICC values for PEMAT-P ranged between 0.81 and 0.89, indicating good-to-excellent agreement between evaluators. Table 2 presents the evaluation criteria applied to score chatbot responses. The GQS was assessed on a five-point scale, with higher values reflecting better quality, flow, and completeness of the information provided. The AOI included five items (factual accuracy, corroboration, consistency, clarity, and relevance), each scored from 0 to 2, yielding a maximum total of 10 points. These measures enabled a structured and objective assessment of the reliability and accuracy of chatbot-generated content. Inter- and intra-rater reliability were determined using Cronbach's α and ICC, which demonstrated good-to-excellent agreement (α=0.83–0.87; ICC=0.78–0.94), thereby confirming the robustness of the scoring system, as detailed in Table 2.

To ensure methodological transparency and comparability across models, chatbot outputs were not subjected to any artificial constraints or post-processing. No limits were imposed on response length, formatting style, tone, or depth of explanation beyond the default system configurations of each platform. Temperature parameters, response length settings, and stylistic controls were not manually modified. Each question was entered as

Table 2. Modified GQS and AOI

Section	Score/item	Description/definition	Maximum score
GQS (Quality/completeness scale)	1	Poor quality, poor flow of the information, most information missing, not at all useful for clinicians	–
GQS (Quality/completeness scale)	2	Generally poor quality and flow, some information listed but many important topics missing, very limited usefulness for clinicians	–
GQS (Quality/completeness scale)	3	Moderate quality, suboptimal flow, some important information adequately discussed, but others poorly discussed, somewhat useful for clinicians	–
GQS (Quality/completeness scale)	4	Good quality and generally good flow. Most of the relevant information is listed, but some topics are not listed, useful for clinicians	–
GQS (Quality/completeness scale)	5	Excellent quality and flow, very useful for clinicians	–
AOI	Factual accuracy	The response aligns with known facts, data, or established knowledge on the subject	2
AOI	Corroboration	The response is based on evidence from textbooks, studies, or guidelines	2
AOI	Consistency	The response is internally consistent and does not contain contradictory statements	2
AOI	Clarity and specificity	The response is clear and specific, avoiding vague or ambiguous language	2
AOI	Relevance	The response directly addresses and adheres to the question or topic posed	2
AOI	Total AOI score	The sum of all scores	10

GQS: Global Quality Scale; AOI: Accuracy of Information Index.

a standalone prompt in a new session, and no follow-up prompts, clarifications, or iterative refinements were provided. Only the first complete response generated by each model was recorded verbatim. Responses were not edited, summarized, reformatted, or truncated before evaluation. This approach was intentionally adopted to simulate real-world patient interactions, where users typically submit a single question and receive an unmodified response. Although evaluation metrics such as GQS and PEMAT-P may be influenced by response verbosity and structural organization, maintaining default output conditions across all models ensured that comparisons reflected inherent model behavior rather than externally imposed standardization procedures.

Statistical Analysis

Normality was assessed using the Shapiro–Wilk test. As data were not normally distributed, nonparametric tests were applied. Differences among the four AI models were analyzed using the Kruskal–Wallis test, followed by Bonferroni-adjusted pairwise comparisons where appropriate. Associations between quality metrics were evaluated using Spearman correlation analysis. Multiple linear regression analysis was performed to identify independent predictors of overall information quality (GQS). A two-sided $p<0.05$ was considered statistically significant. All analyses were conducted using IBM Statistical Package for the Social Sciences Statistics for Windows, Version 26.0 (IBM Corp., Armonk, NY, USA).

Ethical Approval

Ethics committee approval was not required, as the study analyzed secondary, de-identified chatbot responses and did not involve human participants, personal data, or interventions. This decision aligns with institutional policy and international ethical standards (e.g., Declaration of Helsinki).

Results

An evaluation of 40 prompts by two independent raters revealed statistically significant performance differences among the four AI models across all assessed parameters ($p<0.001$). The comparative results are presented in Table 3 and illustrated in Figure 1. ChatGPT-4 and DeepSeek demonstrated the highest intelligibility scores, whereas Gemini showed the lowest values. Regarding actionability,

Table 3. Descriptive statistics and comparison of mean scores for intelligibility, actionability, duration, GQS, and total AOI across artificial intelligence chatbots

Variables	ChatGPT-3.5 (Mean±SD)	ChatGPT-4 (Mean±SD)	Gemini (Mean±SD)	DeepSeek (Mean±SD)	p
Intelligibility score (%)	86.56±16.84[b]	93.75±6.33[a]	69.06±12.66[c]	94.69±6.26[a]	<0.001*
Actionability (%)	60.42±22.86[b]	56.41±24.67[b]	40.83±33.54[c]	85.84±12.26[a]	<0.001*
Duration (seconds)	3.16±0.99[c]	2.32±0.82[c]	4.35±1.13[b]	8.00±3.83[a]	<0.001*
GQS (1–5)	4.03±0.53[b]	4.55±0.55[a]	4.30±0.56[a]	4.55±0.64[a]	<0.001*
Total AOI score (0–10)	8.27±0.68[b]	9.15±0.62[a]	8.63±1.08[b]	9.18±0.98[a]	<0.001*

a, b, c: Groups sharing the same letter are not significantly different; different letters indicate significant differences between groups (Bonferroni-adjusted pairwise comparisons, p<0.05); SD: Standard deviation; GQS: Global Quality Scale; AOI: Accuracy of Information Index.

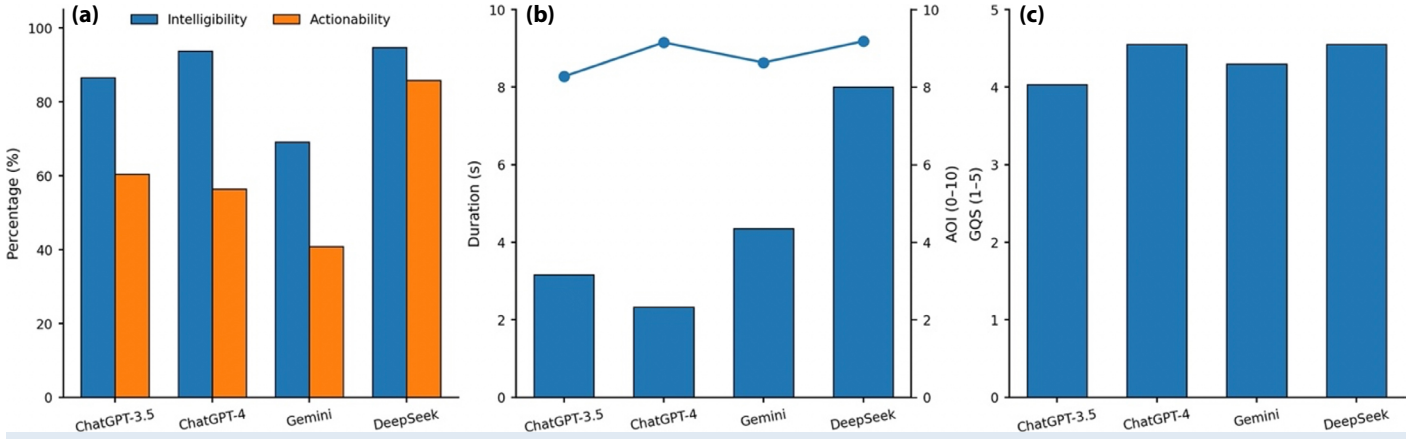


Figure 1. Comparison of artificial intelligence chatbot performance across evaluation metrics. **(a)** Intelligibility and actionability scores (%) presented on a 0–100 scale. **(b)** Duration (seconds) and accuracy of ınformation ındex displayed together using a dual-axis format to improve interpretability. **(c)** Global quality scale scores presented on a 0–5 scale.

DeepSeek achieved the highest mean score, significantly outperforming the other models. ChatGPT-3.5 and ChatGPT-4 exhibited moderate actionability levels, while Gemini had the lowest performance. In terms of response duration, DeepSeek generated significantly longer responses, whereas ChatGPT-4 and ChatGPT-3.5 were the fastest models. For overall quality (GQS), ChatGPT-4 and DeepSeek achieved the highest scores, followed by Gemini, while ChatGPT-3.5 scored comparatively lower. Similarly, for factual accuracy (AOI), ChatGPT-4 and DeepSeek demonstrated superior performance. Overall, ChatGPT-4 and DeepSeek consistently outperformed the other models in intelligibility, overall quality, and factual accuracy. DeepSeek provided the most actionable responses but required longer response times, whereas ChatGPT-4 demonstrated a more balanced profile between quality and efficiency.

Spearman correlation analysis demonstrated a significant positive association between AOI (r=0.39; p<0.001) and GQS (r=0.75; p<0.001) across models. In contrast, response duration was not significantly correlated with overall quality in most comparisons. This finding indicates that higher factual accuracy was consistently associated with higher perceived informational quality. In other words, responses that were more aligned with evidence-based standards tended to receive higher overall quality scores. Conversely, the absence of a significant association between response duration and GQS suggests that longer responses did not necessarily translate into higher informational quality. Multiple linear regression analysis was conducted to identify independent predictors of GQS (Table 4). The regression model was statistically significant (p<0.001) and explained approximately 47% of the variance in perceived quality (R2=0.469). Among the evaluated variables, the AOI emerged as the only significant independent predictor of GQS (β=0.702, p<0.001), whereas understandability, actionability, and response duration did not independently predict overall quality. To further explore the robustness of the regression model, a stepwise regression approach was considered. Given that AOI was the only statistically significant predictor in the full model, it is expected that a stepwise regression model would retain AOI as the sole predictor of GQS. In such a case, a similar coefficient of determination (R^2) would likely be obtained, indicating that

Table 4. Multiple linear regression analysis identifying independent predictors of overall information quality (GQS)

Predictor	B	SE	β	p
Understandability	0.003	0.003	0.087	0.215
Actionability	-0.003	0.002	-0.121	0.112
Response duration (seconds)	0.005	0.012	0.025	0.687
AOI	0.456	0.041	0.702	<0.001

Model summary: R^2=0.469; Adjusted R^2=0.455; F-test p<0.001. GQS: Global Quality Scale; AOI: Accuracy of Information Index

AOI alone explains a substantial proportion of the variance in overall information quality. These findings further support the central role of factual accuracy in determining perceived information quality. Although intelligibility and actionability contribute to the presentation of information, they did not independently predict overall quality when accuracy was controlled for in the regression model. This suggests that evidence-based correctness remains the primary determinant of high-quality patient-oriented information. Our results clearly show that chatbot performance differs across models, likely reflecting differences in architecture and training approaches.

Discussion

The null hypothesis of this study stated that there would be no statistically significant differences among the evaluated AI chatbots in terms of information quality (GQS), factual accuracy (AOI), intelligibility, actionability, and response duration when answering patient-oriented questions about at-home dental bleaching. The present findings demonstrated statistically significant inter-model differences across all assessed parameters (p<0.001). Accordingly, the null hypothesis was rejected. These results indicate that AI chatbot performance is not uniform and varies meaningfully depending on model architecture, training strategy, and alignment mechanisms, underscoring important mechanistic and clinical considerations regarding their use in dental patient education.[17,18]

Performance differences observed among chatbots may be largely attributed to architectural and training-related factors inherent to different model generations. More advanced systems benefit from larger and more diverse training corpora, improved instruction tuning, and refined reinforcement learning with human feedback, which collectively enhance contextual reasoning, internal consistency, and adherence to health-related norms. These mechanisms likely explain the superior balance between factual accuracy and coherence observed in higher-performing models. In contrast, earlier generation models may rely more heavily on surface-level pattern matching, increasing susceptibility to incomplete explanations or clinically ambiguous guidance, particularly in nuanced topics such as dental bleaching safety and contraindications.[13,18]

Another important mechanistic consideration relates to alignment strategies and response generation constraints. Models optimized for conversational fluency may generate longer, more structured responses, which can enhance actionability but also increase the risk of verbosity-driven overgeneralization. Conversely, more conservative alignment approaches may prioritize brevity and factual grounding at the expense of practical guidance. The observed variability in actionability and response duration across models highlights this trade-off and underscores that higher informational quality is not solely dependent on response length but rather on how effectively evidence-based content is translated into patient-understandable recommendations.[16,19]

From a clinical perspective, these findings have direct relevance to dental bleaching, where inappropriate self-directed use of whitening agents can lead to adverse outcomes such as tooth sensitivity, soft-tissue irritation, enamel alterations, or unrealistic esthetic expectations. Accurate guidance regarding patient eligibility, product concentration, treatment duration, and post-bleaching care is essential to ensure safety and efficacy. Although AI chatbots may serve as accessible adjuncts for general patient education, their unsupervised use poses potential risks, particularly when responses lack adequate emphasis on contraindications or the need for professional consultation. Therefore, high-performing chatbots should be viewed as supportive informational tools rather than substitutes for clinician-led decision-making.[23–27]

The strong association observed between factual accuracy and perceived informational quality further reinforces the central role of evidence-based correctness in AI-generated health content. The regression analysis further confirmed this relationship, as AOI emerged as the only independent predictor of GQS, explaining a substantial proportion of the variance in perceived quality. This statistical finding

strengthens the interpretation that accuracy, rather than stylistic features or response length, is the principal driver of quality perception. This finding suggests that improving the reliability of chatbot outputs requires prioritization of validated clinical knowledge rather than optimization for speed or stylistic complexity. Accordingly, future implementations of AI systems in dentistry should adopt a human-in-the-loop framework, where Chatbot-generated information is integrated into dentist-supervised workflows, such as pre-consultation education or post-treatment instruction reinforcement.[28]

Several directions for future research emerge from this study. Longitudinal investigations are needed to evaluate how ongoing model updates influence informational quality over time. Multilingual analyses would further clarify the generalizability of chatbot performance across diverse patient populations. Importantly, experimental studies assessing patient comprehension, behavioral adherence, and clinical outcomes following exposure to AI-generated information would provide critical evidence on real-world effectiveness. In addition, comparative evaluations of domain-specific (fine-tuned dental models vs. general) LLMs may help identify optimal strategies for safely deploying AI tools in clinical dentistry.

Despite these contributions, certain limitations should be acknowledged. The study evaluated chatbot responses at a single time point using standardized prompts, which may not fully capture dynamic conversational interactions. Moreover, outcomes related to patient understanding or behavior were not directly assessed. Nonetheless, by systematically evaluating quality, accuracy, and actionability under controlled conditions, this study provides a robust foundation for future investigations into the responsible integration of AI chatbots into dental patient education.

Conclusion

In this comparative evaluation of four contemporary AI chatbots, statistically significant differences were identified among the models in terms of information quality (GQS), factual accuracy (AOI), intelligibility, actionability, and response duration. ChatGPT-4 and DeepSeek demonstrated the highest performance in overall quality and accuracy. DeepSeek achieved the highest actionability scores but required longer response times, whereas ChatGPT-4 provided a more time-efficient profile with comparable quality. These findings indicate measurable variability in chatbot performance when responding to patient-oriented questions about at-home dental bleaching.

Ethics Committee Approval: Ethics committee approval was not required, as the study analyzed secondary.

Informed Consent: Was not required, as the study analyzed secondary.

Conflict of Interest: None declared.

Financial Disclosure: The author declared that this study has received no financial support.

Use of AI for Writing Assistance: No AI tools were used in the generation, analysis, or writing of the scientific content of this manuscript.

Authorship Contributions: Concept: ÖÖ, EIG; Design: ÖÖ, EIG; Supervision: ÖÖ, EIG; Resource: ÖÖ, EIG; Materials: ÖÖ, EIG; Data collection and/or processing: ÖÖ, EIG; Analysis and/or interpretation: ÖÖ, EIG; Literature review: ÖÖ, EIG; Writing: ÖÖ, EIG; Critical review: ÖÖ, EIG.

Peer-review: Double blind peer-reviewed.

References

1. Topol EJ. Deep medicine: How artificial intelligence can make healthcare human again. New York, NY: Basic Books; 2019.
2. Ding H, Wu J, Zhao W, Matinlinna JP, Burrow MF, Tsoi JKH. Artificial intelligence in dentistry-A review. Front Dent Med 2023;4:1085251. [CrossRef]
3. Shafi I, Fatima A, Afzal H, Díez IT, Lipari V, Breñosa J, Ashraf I. A Comprehensive review of recent advances in artificial intelligence for dentistry e-health. Diagnostics (Basel) 2023;13(13):2196. [CrossRef]
4. Stephan D, Bertsch A, Burwinkel M, Vinayahalingam S, Al-Nawas B, Kämmerer PW, Thiem DG. AI in dental radiology-improving the efficiency of reporting with ChatGPT: comparative study. J Med Internet Res 2024;26:e60684. [CrossRef]
5. Farhadi Nia M, Ahmadi M, Irankhah E. Transforming dental diagnostics with artificial intelligence: advanced integration of ChatGPT and large language models for patient care. Front Dent Med 2025;5:1456208. [CrossRef]
6. Giannakopoulos K, Kavadella A, Aaqel Salim A, Stamatopoulos V, Kaklamanos EG. Evaluation of the performance of generative ai large language models ChatGPT, google bard, and microsoft bing chat in supporting evidence-based dentistry: comparative mixed methods study. J Med Internet Res 2023;25:e51580. [CrossRef]
7. Naik S, Al-Kheraif AA, Vellappally S. Artificial intelligence in dentistry: Assessing the informational quality of YouTube videos. PLoS One 2025;20(1):e0316635. [CrossRef]
8. Eggmann F, Weiger R, Zitzmann NU, Blatz MB. Implications of large language models such as ChatGPT for dental medicine. J Esthet Restor Dent 2023;35(7):1098-102. [CrossRef]
9. Balel Y. Can ChatGPT be used in oral and maxillofacial surgery? J Stomatol Oral Maxillofac Surg 2023;124(5):101471. [CrossRef]
10. Bender EM, Gebru T, McMillan Major A, Mitchell M. On the dangers of stochastic parrots: Can language models be too big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21) 2021:610-23. [CrossRef]

11. Maynez J, Narayan S, Bohnet B, McDonald R. On faithfulness and factuality in abstractive summarization. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020. [CrossRef]
12. Lu X, Zhang R, Wu W, Shang X, Liu M. Relationship between internet health information and patient compliance based on trust: Empirical study. J Med Internet Res 2018;20(8):e253. [CrossRef]
13. Daraqel B, Wafaie K, Mohammed H, Cao L, Mheissen S, Liu Y, et al. The performance of artificial intelligence models in generating responses to general orthodontic questions: ChatGPT vs Google Bard. Am J Orthod Dentofacial Orthop 2024;165(6):652-62. [CrossRef]
14. Akarslan ZZ, Sadik B, Erten H, Karabulut E. Dental esthetic satisfaction, received and desired dental treatments for improvement of esthetics. Indian J Dent Res 2009;20(2):195-200. [CrossRef]
15. Aldaij M, Alshehri T, Alzeer A, Alfayez A, Aldrees F, Almuhaya S, et al. Patient satisfaction with dental appearance and treatment desire to improve esthetics. J Oral Health Comm Dent 2018;12(3):90-5. [CrossRef]
16. Demarco FF, Meireles SS, Masotti AS. Over-the-counter whitening agents: A concise review. Braz Oral Res. 2009;23(Suppl 1):64-70. [CrossRef]
17. 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]
18. Taymour N, Fouda SM, Abdelrahaman HH, Hassan MG. Performance of the ChatGPT-3.5, ChatGPT-4, and Google Gemini large language models in responding to dental implantology inquiries. J Prosthet Dent 2025;134(6):2427-34. [CrossRef]
19. Carey CM. Tooth whitening: what we now know. J Evid Based Dent Pract. 2014;14 Suppl:70-6. [CrossRef]
20. FDI World Dental Federation. FDI policy statement on dental bleaching materials: adopted by the FDI General Assembly: 17 September 2011 - Mexico City, Mexico. Int Dent J 2013;63(1):2-3. [CrossRef]
21. American Dental Association, Council on Scientific Affairs. Tooth whitening/bleaching: Treatment considerations for dentists and their patients. Chicago (IL): American Dental Association; 2010.
22. Shoemaker SJ, Wolf MS, Brach C. Development of the patient education materials assessment tool (PEMAT): a new measure of understandability and actionability for print and audiovisual patient information. Patient Educ Couns. 2014;96(3):395-403. [CrossRef]
23. Alhaidry HM, Fatani B, Alrayes JO, Almana AM, Alfhaed NK. ChatGPT in dentistry: A comprehensive review. Cureus 2023;15(4):e38317. [CrossRef]
24. Azadi A, Gorjinejad F, Mohammad-Rahimi H, Tabrizi R, Alam M, Golkar M. Evaluation of AI-generated responses by different artificial intelligence chatbots to the clinical decision-making case-based questions in oral and maxillofacial surgery. Oral Surg Oral Med Oral Pathol Oral Radiol 2024;137(6):587-93. [CrossRef]
25. Babayiğit O, Tastan Eroglu Z, Ozkan Sen D, Ucan Yarkac F. Potential use of ChatGPT for patient information in periodontology: a descriptive pilot study. Cureus. 2023;15(11):e48518. [CrossRef]
26. Mahdi SS, Battineni G, Khawaja M, Allana R, Siddiqui MK, Agha D. How does artificial intelligence impact digital healthcare initiatives? A review of AI applications in dental healthcare. Int J Inf Manag Data Insights 2023;3(1):100144. [CrossRef]
27. 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]
28. Salmanpour F, Akpınar M. Performance of Chat Generative Pretrained Transformer-4.0 in determining labiolingual localization of maxillary impacted canine and presence of resorption in incisors through panoramic radiographs: A retrospective study. Am J Orthod Dentofacial Orthop 2025;168(2):220-31. [CrossRef]