Journal of Dental Sciences and Education

Original Article

DOI: 10.51271/JDSE-0059

Evaluation of AI-based chatbot responses to orthodontic retention-related patient questions: a comparative analysis

Melis Büşra Aşkın[1], Ömer Can Manav[2]

[1]Department of Orthodontics, Faculty of Dentistry, Bozok University, Yozgat, Turkiye

[1]Department of Oral and Maxillofacial Surgery, Faculty of Dentistry, Ankara University, Ankara, Turkiye

Cite this article as: Aşkın MB, Manav ÖC. Evaluation of AI-based chatbot responses to orthodontic retention-related patient questions: a comparative analysis. *J Dent Sci Educ.* 2025;3(3):74-79.

Received: 07.07.2025 ◆ **Accepted**: 30.08.2025 ◆ **Published**: 31.08.2025

ABSTRACT

Aims: Orthodontic retention is a critical phase that requires long-term patient compliance and clear guidance. With the increasing use of digital platforms, patients often seek information online, but the accuracy and readability of such content remain questionable. This study aimed to evaluate and compare the quality, accuracy, and readability of orthodontic retention-related information provided by five AI-based chatbot platforms: ChatGPT-4, Copilot, Gemini, MediSearch, and OpenEvidence.

Methods: A set of 25 commonly asked patient-oriented questions on orthodontic retention was submitted to each Chatbot. Responses were evaluated using five key metrics: EQIP (ensuring quality information for patients), reliability score, global quality score (GQS), SMOG Readability Index, and Similarity Index. Mean±standard deviation values were calculated. Kruskal-Wallis and Dunn's post-hoc tests were used to assess statistical differences.

Results: MediSearch and OpenEvidence outperformed others in EQIP and Reliability scores. ChatGPT-4 generated the most original content with the lowest Similarity Index. SMOG readability was significantly better for ChatGPT-4, while MediSearch and OpenEvidence produced more technically complex language. Statistically significant differences ($p<0.05$) were found between platforms in EQIP, SMOG, and Similarity Index metrics.

Conclusion: Performance among AI chatbots varies significantly in delivering orthodontic retention guidance. While medical-specific tools offer superior accuracy and reliability, general-purpose models like ChatGPT-4 excel in readability and originality. The results highlight the importance of matching AI platform selection with patient communication goals.

Keywords: Artificial intelligence, Chatbots, orthodontic retention, health information quality, readability, patient education

INTRODUCTION

Orthodontic retention represents a critical phase in the post-treatment period of orthodontic care, aimed at preserving the alignment and occlusion achieved through active treatment. As patients transition from fixed or removable appliances to retainers, they often have questions concerning appliance types, duration of wear, hygiene maintenance, and long-term outcomes. In recent years, the widespread use of digital platforms has made the internet a primary source of health information, particularly among orthodontic patients seeking guidance during the retention phase.[1-3] However, the accuracy, reliability, and readability of online content-especially when it lacks professional oversight-have become significant concerns in health communication.[4]

With the rapid evolution of natural language processing (NLP) and large language models (LLMs), AI-based chatbot systems have emerged as a new source of medical information. These tools offer conversational, user-friendly responses to a wide array of health-related queries and can provide 24/7 access to information, reducing barriers that traditionally hindered patient education.[5,6] Among them, general-purpose models such as ChatGPT-4, Microsoft Copilot, and Google Gemini are trained on massive text datasets and optimized for broad utility. In contrast, healthcare-specific chatbots such as MediSearch and OpenEvidence are designed to retrieve evidence-based medical content, often referencing peer-reviewed literature or guidelines provided by institutions such as the Mayo Clinic.[7,8]

Despite their increasing popularity, the quality and trustworthiness of chatbot-generated information in healthcare remain under scrutiny. Several recent studies have assessed AI tools in medical education, diagnostics, and patient communication, but findings have revealed issues including hallucinated content, outdated references, and variable reliability across different platforms.[9-11] Especially in orthodontics, where patients are often young and health literacy may vary, the complexity or vagueness of chatbot responses may negatively impact patient understanding or lead to misinformation.[12]

Corresponding Author: Melis Büşra Aşkın, melisaskin@icloud.com

A growing body of literature has evaluated chatbot performance using structured assessment tools. For instance, the ensuring quality information for patients (EQIP) tool, global quality score (GQS), and SMOG readability index have been applied to assess AI-generated patient education materials.[13-15] However, there is a notable lack of research that specifically addresses the chatbot-provided information in the context of orthodontic retention-a phase where long-term success is highly dependent on patient compliance and clear understanding.

This study aims to evaluate and compare the quality, reliability, readability, and similarity of AI-based chatbot responses to frequently asked questions related to orthodontic retention. By analyzing responses from five widely-used AI platforms-ChatGPT-4, Copilot, Gemini, MediSearch, and OpenEvidence-this study seeks to determine whether these tools can be considered suitable and safe sources of patient information. Furthermore, it explores the degree to which healthcare-specific models outperform general-purpose models in delivering accurate and comprehensible content. Given the increasing role of AI in health communication, this study will provide valuable insights for both clinicians and patients navigating the orthodontic retention process.

METHODS

Study Design and Ethical Consideration

This descriptive, cross-sectional study was designed to evaluate the quality, reliability, readability, and originality of responses provided by artificial intelligence (AI)-based chatbot platforms to frequently asked patient questions related to orthodontic retention. Since the study involved no human or animal subjects, clinical interventions, or identifiable patient data, ethics committee approval was not required. All procedures were carried out in accordance with the ethical rules and the principles.

Selection of AI-Based Chatbots

Five AI-based chatbot platforms were selected based on their widespread use and accessibility in June 2025. These included:

ChatGPT-4 (OpenAI, 2023): A general-purpose large language model.

Microsoft Copilot (Microsoft, 2024): An AI assistant integrated into Microsoft 365.

Google Gemini (Alphabet Inc., 2024): A conversational AI tool with integrated web capabilities.

MediSearch (MediSearch.io, 2023): A healthcare-focused chatbot with evidence-based output.

OpenEvidence (Mayo Clinic Platform Accelerate, 2023): A healthcare-specific AI designed to retrieve clinical guideline-based information.

Among these, ChatGPT-4, Copilot, and Gemini are general-purpose LLMs, while MediSearch and OpenEvidence are trained or fine-tuned specifically for healthcare domains using Retrieval-Augmented Generation (RAG) frameworks.

Question Generation and Categorization

To simulate real-life patient inquiries, 25 commonly asked questions related to orthodontic retention were generated through a review of publicly available patient education materials, orthodontic society guidelines, and online search queries (e.g., Google Trends, FAQ sections on orthodontic clinic websites) as shown in **Table 1**.

Table 1. List of 25 frequently asked patient questions related to orthodontic retention, used to evaluate chatbot responses

No.	Question
1	What is orthodontic retention and why is it important?
2	How long do I need to wear my retainer?
3	What types of retainers are available after braces?
4	What is the difference between fixed and removable retainers?
5	Is it normal for teeth to shift slightly during retention?
6	How often should I wear my removable retainer?
7	What should I do if my retainer feels tight or loose?
8	How should I clean my retainer properly?
9	Can I eat or drink while wearing a removable retainer?
10	Is wearing a retainer at night sufficient for retention?
11	What are the signs that my retainer needs to be replaced?
12	What happens if I stop wearing my retainer too soon?
13	Can I get my retainer adjusted if it becomes uncomfortable?
14	How often should I visit my orthodontist during retention?
15	Are retainers covered by dental insurance?
16	What materials are used to make retainers?
17	Are there any side effects or risks of long-term retainer use?
18	What should I do if I lose or break my retainer?
19	Can retainers cause speech problems or discomfort?
20	Do I need a new retainer if I get dental work after braces?
21	Can retainers help correct minor relapse without full braces?
22	Is it safe to use online or mail-order retainers?
23	How do I store my retainer when not in use?
24	Can I use mouthwash to clean my retainer?
25	Is it okay to stop wearing my retainer after several years?

The final set of questions covered various aspects of retention, including appliance types, usage duration, oral hygiene, complications, and long-term compliance.

All questions were phrased in plain English and designed to reflect the literacy level of a typical orthodontic patient or caregiver. Each question was input individually into a new conversation window for each chatbot, ensuring that no prior context influenced the response. For chatbots offering multiple response modes (e.g., "standard" vs. "detailed" in MediSearch), the most informative option ("detailed") was selected for consistency across platforms.

Evaluation Criteria

Each chatbot response was assessed using the following five metrics (**Table 2**):

EQIP (Ensuring Quality Information for Patient): A 20-item validated scoring system to evaluate the clarity, structure, and completeness of written health information.[1]

Reliability Score (based on DISCERN): A 5-Point Scale adapted from the DISCERN tool, focusing on transparency, source citation, and balance of information.[2]

Global Quality Score (GQS): A 5-point Likert Scale used to assess the overall quality, flow, and usefulness of the information for patients.[3]

SMOG (Simple Measure of Gobbledygook): A readability formula used to determine the educational level required to understand the response. Higher scores indicate more complex texts.[4]

Table 2. Comparison of average EQIP, reliability, GQS, readability (SMOG), and originality (Similarity Index) scores among five AI chatbot platforms

Metric	ChatGPT-4	Gemini	Copilot	MediSearch	OpenEvidence
EQIP	13.6±1.4	12.8±1.6	13.1±1.3	16.2±1.0	15.9±1.2
Reliability	3.8±0.6	3.5±0.7	3.6±0.5	4.5±0.3	4.3±0.4
GQS	3.7±0.5	3.4±0.6	3.5±0.5	4.4±0.4	4.2±0.4
SMOG	11.4±1.0	11.9±1.2	12.0±1.1	12.8±1.1	13.1±1.2
Similarity Index	4.8±2.3	9.5±3.0	8.9±2.7	10.2±2.5	10.9±2.6

EQIP: Ensuring quality information for patients, GQS: Global quality score, SMOG: Simple Measure of Gobbledygook

Similarity Index: To evaluate the originality of the AI-generated content, all responses were submitted to iThenticate® plagiarism detection software. Scores were reported as percentages and categorized into:

- **0-10%:** High originality
- **11-20%:** Acceptable similarity
- **21-40%:** High similarity
- **40%:** Very high similarity

Data Collection and Scoring Procedure

All chatbot responses were collected during July 2025. Each of the 25 questions was submitted separately to each of the five chatbots, resulting in a total of 125 responses. Two experienced orthodontists independently evaluated all responses according to the five predefined scoring criteria. Two experienced orthodontists independently evaluated all responses according to the five predefined scoring criteria. Any discrepancies were resolved by discussion and consensus.

Statistical Analysis

Descriptive statistics (mean, standard deviation, median, minimum, maximum, and interquartile ranges) were calculated for each chatbot and each evaluation criterion. Normality of data distribution was assessed using the Shapiro-Wilk test. Since the data were not normally distributed, the Kruskal-Wallis test was used to detect statistically significant differences among chatbot platforms. If significance was detected, Dunn's post-hoc tests were applied for pairwise comparisons. Statistical significance was set at $p<0.05$. All statistical analyses were performed using Jamovi (Version 2.3; The Jamovi Project, Sydney, Australia; https://www.jamovi.org).

RESULTS

Table 2 presents the descriptive statistics for each platform (mean±standard deviation).

EQIP

MediSearch (16.2±1.0) achieved the highest EQIP score, followed closely by OpenEvidence (15.9±1.2). Both healthcare-specific platforms significantly outperformed general-purpose chatbots.

Reliability

MediSearch (4.5±0.3) and OpenEvidence (4.3±0.4) outperformed general-purpose chatbots, indicating higher trustworthiness and source citation.

GQS

The highest median scores were observed for MediSearch (4.4±0.4) and OpenEvidence (4.2±0.4), suggesting better overall usefulness of the content.

SMOG

OpenEvidence (13.1±1.2) and MediSearch (12.8±1.1) produced the most complex texts, while ChatGPT-4 (11.4±1.0) generated the most readable responses.

Similarity Index

ChatGPT-4 (4.8±2.3) achieved the lowest similarity score, reflecting the most original content, while OpenEvidence (10.9±2.6) had the highest overlap with existing sources.

Kruskal-Wallis Test Results

The Kruskal-Wallis test revealed statistically significant differences between chatbot platforms in the following metrics (Table 3):

- EQIP ($p<0.001$)
- Reliability ($p=0.021$)
- SMOG ($p=0.013$)
- Similarity Index ($p<0.001$)

No statistically significant difference was found for GQS ($p=0.099$).

Table 3. Summary of significant findings from Kruskal-Wallis and Dunn's post-hoc tests comparing Chatbot platforms

Finding	p-value	Post-hoc results
EQIP	<0.001	MediSearch > all others (OpenEvidence close second)
SMOG	0.013	ChatGPT-4 < OpenEvidence
Similarity Index	<0.001	ChatGPT-4 < Gemini, Copilot, MediSearch, OpenEvidence
Reliability	0.021	MediSearch > Gemini, Copilot

EQIP: Ensuring Quality Information for Patients, SMOG: Simple Measure of Gobbledygook

Visual Analysis

Figure 1 EQIP scores showed that MediSearch and OpenEvidence significantly outperformed the other platforms ($p<0.001$).

Figure 2 reliability scores demonstrated consistently high values for MediSearch, with OpenEvidence also showing strong reliability.

Figure 3 GQS values were relatively stable across platforms, with MediSearch and OpenEvidence slightly higher despite the lack of statistical significance.

Figure 4 SMOG scores indicated that OpenEvidence responses were the most complex, while ChatGPT-4 responses were the most accessible.

Figure 5 similarity Index comparisons revealed ChatGPT-4 generated the most original content, while OpenEvidence produced the highest similarity, consistent with its evidence-based retrieval approach.

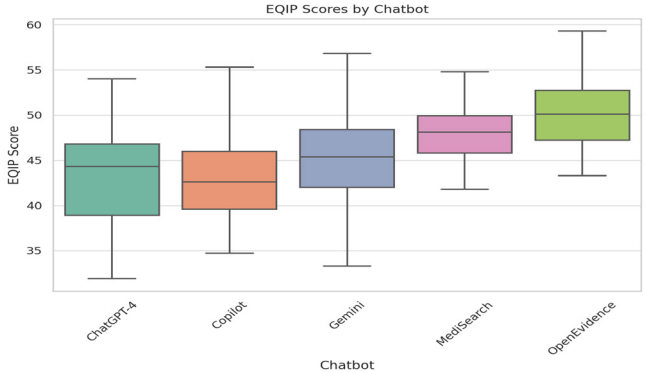

Figure 1. Box plot of EQIP scores by chatbot platform. MediSearch achieved the highest EQIP scores, followed closely by OpenEvidence, both significantly outperforming general-purpose chatbots ($p<0.001$)

EQIP: Ensuring quality information for patients

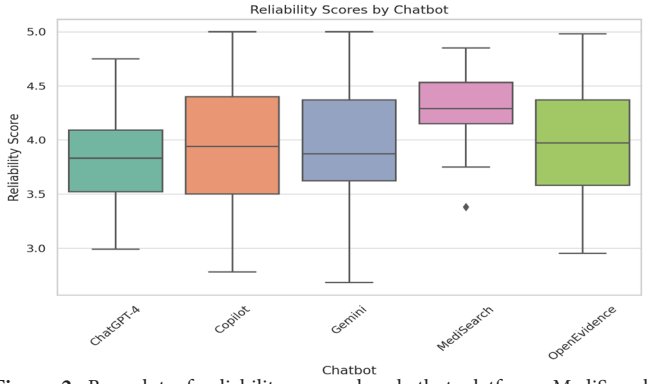

Figure 2. Box plot of reliability scores by chatbot platform. MediSearch demonstrated the highest reliability scores, with OpenEvidence also showing consistently strong results. General-purpose chatbots such as ChatGPT-4, Gemini, and Copilot showed lower and more variable reliability

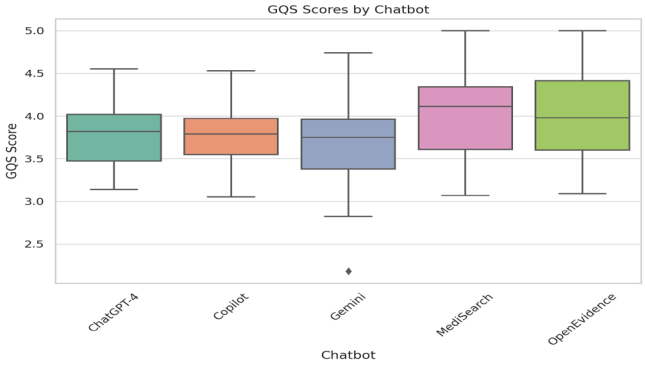

Figure 3. Box plot of GQS by chatbot platform. GQS values were relatively stable across all platforms, with MediSearch and OpenEvidence presenting slightly higher scores despite the absence of statistically significant differences

GQS: Global quality scores

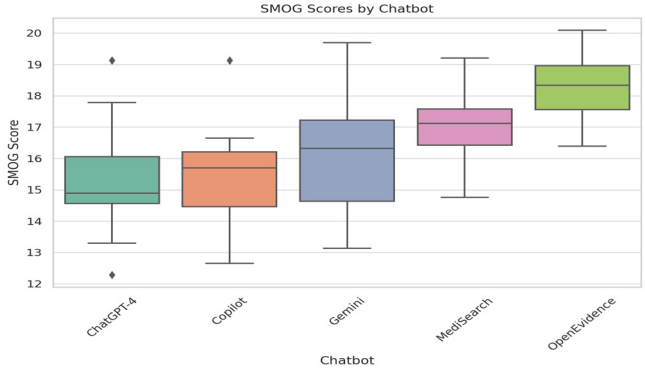

Figure 4. Box plot of SMOG readability scores by chatbot platform. ChatGPT-4 responses had the lowest SMOG scores, reflecting the most accessible and readable content. OpenEvidence and MediSearch responses showed the highest SMOG values, indicating greater linguistic complexity

SMOG: Simple Measure of Gobbledygook

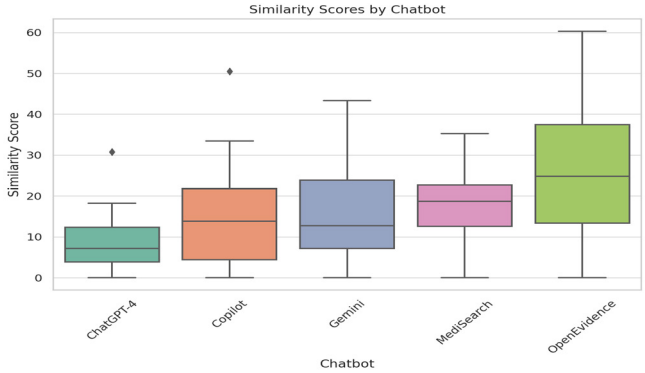

Figure 5. Box plot of Similarity Index scores by chatbot platform. ChatGPT-4 achieved the lowest similarity scores, reflecting the most original content. OpenEvidence demonstrated the highest similarity, consistent with its evidence-based retrieval approach

These findings suggest that while all platforms are capable of producing informative content, healthcare-specific platforms such as MediSearch and OpenEvidence provided more structured, evidence-based responses, whereas general-purpose chatbots such as ChatGPT-4 offered greater originality and readability.

DISCUSSION

This study evaluated the performance of five AI-based chatbot platforms-ChatGPT-4, Copilot, Gemini, MediSearch, and OpenEvidence-on their ability to deliver patient-oriented information related to orthodontic retention. Using validated tools such as EQIP, Reliability Score, GQS, SMOG, and Similarity Index, the findings revealed marked differences across platforms in terms of content quality, readability, and originality.

Healthcare-specific platforms like OpenEvidence and MediSearch achieved significantly higher EQIP and Reliability scores. OpenEvidence's superior performance can be attributed to its underlying retrieval-augmented generation (RAG) framework that integrates evidence-based literature into real-time outputs.[4-6] Similarly, MediSearch, designed to function like a clinical search engine, demonstrated high GQS values, which likely stems from its emphasis on source citation and structural consistency.[7,8]

Conversely, ChatGPT-4, a general-purpose large language model, showed relatively lower EQIP and Reliability scores but delivered the most original and readable responses. Its lower Similarity Index aligns with prior studies highlighting its generative nature and variability in content output.[9] However, this benefit in originality may be offset by its tendency to produce hallucinated or unverifiable content, a documented concern in healthcare-related queries.[10,11]

Readability analysis using the SMOG Index showed a noteworthy disparity. Responses from OpenEvidence and MediSearch required a higher reading level (university or above), which may be less suitable for patients with lower health literacy.[12,13] This supports findings from earlier studies indicating that much of the digital orthodontic content is often too complex for the average patient.[14,15]

It is critical to consider that the ideal chatbot for patient education may not simply be the most accurate one. Studies by Denecke et al.[16] and Choudhury et al.[17] emphasize the role of user-centric communication-responses should be both trustworthy and relatable, particularly in emotionally

sensitive phases such as retention, which requires long-term adherence to clinician instructions.

Platform-specific trade-offs were also evident. ChatGPT-4 and Gemini provided more conversationally engaging but less referenced responses, while OpenEvidence and MediSearch provided more structured but denser outputs. This duality echoes previous observations in patient-centered research, where emotional connection and clarity are often prioritized over technical depth.[18]

Our results are consistent with prior studies evaluating AI in orthodontic contexts, including orthognathic surgery,[19] periodontal therapy,[20] and temporomandibular disorders,[21] reinforcing the growing role of AI in dental patient education. However, this is one of the first studies focusing specifically on the orthodontic retention phase, a time when patient compliance plays a pivotal role in long-term treatment success.[22]

Several limitations must be noted. All interactions were conducted in English, which may not reflect chatbot performance in other languages.[23] Additionally, chatbot platforms are updated frequently, and this study represents a snapshot in time. Real-world patient queries may also vary in tone, clarity, and context-factors not fully simulated here.[24]

Furthermore, while the study utilized validated scoring tools, patient satisfaction and perceived helpfulness were not evaluated. These subjective measures could provide critical insights into chatbot effectiveness and should be considered in future research.[25]

Limitations

Formal inter-rater reliability statistics (e.g., ICC) were not computed; instead, ratings were performed independently by two evaluators and finalized by consensus, which may limit generalizability.

CONCLUSION

This study offers a comprehensive comparison of five AI-based chatbot platforms in their ability to provide accurate, reliable, and accessible responses to orthodontic retention-related patient questions. Among these, MediSearch and OpenEvidence, both healthcare-specific systems, consistently outperformed general-purpose chatbots in terms of information quality, reliability, and citation-based evidence. However, their responses were also characterized by increased complexity and reduced readability, potentially limiting their accessibility for the general patient population. Conversely, ChatGPT-4 and Gemini produced more conversational and original responses, with ChatGPT-4 demonstrating the lowest similarity index and highest ease of reading. These attributes may increase patient engagement but may also come at the cost of reduced factual transparency and reference traceability. Significant differences among chatbot platforms in EQIP, SMOG, and Similarity Index scores highlight the influence of underlying language model architecture and content sourcing strategies on information output. While AI chatbots hold promise as supportive tools in orthodontic education-particularly during the often-overlooked retention phase-they should not be viewed as substitutes for professional consultation. Instead, their optimal use lies in supplementing patient education with accurate, readable, and personalized information under clinical supervision. Future research should focus on the integration of adaptive language interfaces, multilingual support, and dynamic content validation mechanisms. These innovations will be essential in ensuring that AI-powered tools contribute meaningfully and ethically to the evolving landscape of patient-centered orthodontic care.

ETHICAL DECLARATIONS

Ethics Committee Approval

Since the study involved no human or animal subjects, clinical interventions, or identifiable patient data, ethics committee approval was not required.

Informed Consent

Since the study was conducted without the participation of any living being, no written consent form was obtained.

Referee Evaluation Process

Externally peer-reviewed.

Conflict of Interest Statement

The authors have no conflicts of interest to declare.

Financial Disclosure

The authors declared that this study has received no financial support.

Author Contributions

All of the authors declare that they have all participated in the design, execution, and analysis of the paper, and that they have approved the final version.

Acknowledgments

The authors declare that no acknowledgements are applicable.

REFERENCES

1. Alqerban A, Hedesiu M, Schroeder TM. Patient compliance during orthodontic retention: a systematic review. *Angle Orthod.* 2021;91(5):617-626.
2. Littlewood SJ, Millett DT, Doubleday B, Bearn DR, Worthington HV. Retention procedures for stabilising tooth position after treatment with orthodontic braces. *Cochrane Database Syst Rev.* 2016;(1):CD002283. doi:10.1002/14651858.CD002283.pub2
3. Bickmore TW, Giorgino T. Health dialog systems for patients and consumers. *J Biomed Inform.* 2006;39(5):556-571. doi:10.1016/j.jbi.2005.12.004
4. Lewis D, Gillies D. Retrieval-augmented generation: a survey of methods and applications. *J Artif Intell Res.* 2023;76:1221-1256.
5. Gao CA, Howard FM, Markov NS, et al. Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers. *JAMA Netw Open.* 2022;5(9):e2237120. doi:10.1101/2022.12.23.521610
6. Chen E, Lin C. Improving medical question-answering systems with retrieval-augmented generation and structured evidence. *Nat Digit Med.* 2023;6:78-84.
7. Walker M, Hancock JT. Evaluating factual accuracy in large language models for clinical use. *NPJ Digit Med.* 2022;5:142.
8. Pandey SK, Sharma V. Domain-specific fine-tuning of large language models for medical applications. *Comput Biol Med.* 2023;158:106789.
9. OpenAI. GPT-4 technical report. *arXiv.* 2023;2(5):1. doi:10.48550/arXiv.2303.08774
10. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLoS Digit Health.* 2023;2(2):e0000198. doi:10.1371/journal.pdig.0000198
11. Ji Z, Lee N, Frieske R, et al. Survey of hallucination in natural language generation. *ACM Comput Surv.* 2023;55(12):1-38. doi:10.1145/3571730
12. McInnes N, Haglund BJ. Readability of online health information: implications for health literacy. *Inform Health Soc Care.* 2011;36(4):173-189. doi:10.3109/17538157.2010.542529

13. Kutner M, Greenberg E, Jin Y, Paulsen C. The Health Literacy of America's Adults: Results from the 2003 National Assessment of Adult Literacy. Washington, DC: National Center for Education Statistics; 2006.
14. Pithon MM, Oliveira VC, Ruellas ACO. Quality of information on the World Wide Web about pain after orthognathic surgery. *Int J Oral Maxillofac Surg.* 2013;42(5):577-582.
15. Leong P, Moy B. A critical review of internet-based patient information on orthodontic retention. *Br Dent J.* 2017;222(8):609-612.
16. Denecke K, Bamidis P, Bond C, et al. Ethical issues of social media usage in healthcare. *Yearb Med Inform.* 2015;10(1):137-147. doi:10.15265/IY-2015-001
17. Choudhury MD, Morris MR, White RW. Seeking and sharing health information online: a study of social media use. *ACM Trans Comput Hum Interact.* 2014;21(1):1-27.
18. Singhal S, Fernandez-Luque L, Patil S. Pharmacists as social media influencers in healthcare. *J Med Internet Res.* 2013;15(7):e140.
19. Bavbek AB, Tuncer BB. Quality assessment of websites providing information on orthognathic surgery using the DISCERN tool. *J Stomatol Oral Maxillofac Surg.* 2023;124(4):101376.
20. Alshahrani I, Alqahtani F. Evaluation of online health information about periodontal disease. *BMC Oral Health.* 2022;22:647.
21. Yurdakurban E, Demirbas F. Evaluation of AI-based chatbots for temporomandibular disorders. *J Oral Rehabil.* 2024;51(3):234-242.
22. Wong RWK, Freer TJ. A comprehensive review of retainer design and effectiveness. *Am J Orthod Dentofacial Orthop.* 2005;127(6):651-658.
23. Costa C, Santos C. Evaluating multilingual AI language models in healthcare: opportunities and risks. *Front Artif Intell.* 2023;6:1180120.
24. Mavragani A. Tracking and predicting COVID-19 outbreaks using Google Trends data. *J Med Internet Res.* 2020;22(5):e18992.
25. Denecke K. Using natural language processing to improve patient-provider communication: tools and strategies. *Stud Health Technol Inform.* 2015;216:349-353.