doi: 10.53446/actamednicomedia.1949463

Acta Medica Nicomedia

Cilt: 9 (2026), 1-12 / Vol: 9 (2026), 1-12
https://dergipark.org.tr/tr/pub/actamednicomedia

Research Article | Araştırma Makalesi

ARE AI MODELS READY FOR CLINICAL DECISION-MAKING IN DENTAL BLEACHING? AN EVIDENCE-BASED EVALUATION OF ACCURACY, COMPLETENESS, AND READABILITY OF CHATGPT-5, GEMINI 2.5 PRO, DEEPSEEK V3.2, AND CLAUDE SONNET 4.5

YAPAY ZEKÂ MODELLERİ DİŞ BEYAZLATMADA KLİNİK KARAR VERME SÜREÇLERİNE HAZIR MI? CHATGPT-5, GEMİNİ 2.5 PRO, DEEPSEEK V3.2 VE CLAUDE SONNET 4.5'İN DOĞRULUK, TAMLIK VE OKUNABİLİRLİK AÇISINDAN KANITA DAYALI DEĞERLENDİRİLMESİ

Suzan Cangul[1*], Ozkan Adiguzel[2], Tuba Tunc[1], Makbule Tasyurek[2], Hatice Ortac[3]

[1]Dicle University, Faculty of Dentistry, Department of Restorative Dentistry, Diyarbakir, Türkiye. [2]Dicle University, Faculty of Dentistry, Department of Endodontics, Diyarbakir, Türkiye. [3] Dicle University, Faculty of Medicine, Department of Biostatistics, Diyarbakır, Türkiye.

ABSTRACT

Objective: The aim of this study was to compare the accuracy, completeness, and readability of information related to whitening in dentistry provided by the large language models (LLM) ChatGPT-5, Gemini 2.5 Pro, DeepSeek v3.2, and Claude Sonnet 4.5.

Methods: A total of 25 open-ended questions covering various aspects of dental whitening were prepared and presented to four different LLMs. The responses were recorded and evaluated independently by two restorative dental treatment specialists who were blinded to the source of the responses. Accuracy and completeness were evaluated using 5-point and 3-point Likert scales respectively. Readability was evaluated using the Flesch Reading Ease Score (FRES), the Flesch–Kincaid Grade Level (FKGL), and the Simple Measure of Gobbledygook (SMOG) index. In the statistical analyses, the Intraclass Correlation Coefficient (ICC), Shapiro-Wilk test, Kruskal-Wallis test, Dunn-Bonferroni test, and Spearman correlation analysis were used.

Results: A statistically significant difference was identified among the artificial intelligence models regarding the accuracy of their responses to open-ended questions related to dental whitening ($p<0.05$). The accuracy score of DeepSeek v.3.2 (4.52±0.77) was determined to be statistically signficntly higher than that of ChatGPT-5 (3.88±1.01). The FRES points for readability were found to be statistically significantly higher for DeepSeek v3.2 (42.02±7.63) compared to the mean points obtained for ChatGPT-5 (29.09±9.07), Gemini 2.5 Pro (33.92±9.24), and Claude Sonnet 4.5 (21.65±9.29).

Conclusion: DeepSeek v3.2 exhibited better performance than ChatGPT-5 in terms of accuracy. No statistically significant differences were identified among the models regarding the completeness scores of responses to open-ended questions. However, the FRES results indicated that DeepSeek v3.2 generated texts with higher readability compared to the other chatbots. Overall, these findings suggest that DeepSeek v3.2 demonstrated superior performance in terms of both accuracy and readability.

Keywords: Artificial intelligence, large language models, teeth whitening, dentistry

ÖZ

Amaç: Bu çalışmanın amacı, büyük dil modelleri (LLM) olan ChatGPT-5, Gemini 2.5 Pro, DeepSeek v3.2 ve Claude Sonnet 4.5 tarafından diş hekimliğinde beyazlatma ile ilgili sunulan bilgilerin doğruluk, tamlık ve okunabilirlik açısından karşılaştırılmasıdır.

Yöntem: Diş beyazlatmasının çeşitli yönlerini kapsayan toplam 25 açık uçlu soru hazırlanmış ve dört farklı büyük dil modeline yöneltilmiştir. Elde edilen yanıtlar kaydedilmiş ve kaynakları gizlenmiş şekilde iki restoratif diş tedavisi uzmanı tarafından bağımsız olarak değerlendirilmiştir. Doğruluk ve kapsamlılık sırasıyla 5'li ve 3'lü Likert ölçekleri kullanılarak değerlendirilmiştir. Okunabilirlik ise Flesch Reading Ease Score (FRES), Flesch–Kincaid Grade Level (FKGL) ve Simple Measure of Gobbledygook (SMOG) indeksleri kullanılarak analiz edilmiştir. İstatistiksel analizlerde Sınıf İçi Korelasyon Katsayısı (ICC), Shapiro–Wilk testi, Kruskal–Wallis testi, Dunn–Bonferroni testi ve Spearman korelasyon analizi kullanılmıştır.

Bulgular: Diş beyazlatmaya ilişkin açık uçlu sorulara verilen yanıtların doğruluk puanları açısından yapay zekâ modelleri arasında istatistiksel olarak anlamlı farklılık saptanmıştır ($p<0,05$). DeepSeek v3.2'nin doğruluk puanı (4,52±0,77), ChatGPT-5'in doğruluk puanından (3,88±1,01) istatistiksel olarak anlamlı derecede daha yüksek bulunmuştur. Okunabilirlik açısından değerlendirilen FRES puanlarının ise DeepSeek v3.2'de (42,02±7,63), ChatGPT-5 (29,09±9,07), Gemini 2.5 Pro (33,92±9,24) ve Claude Sonnet 4.5'e (21,65±9,29) kıyasla istatistiksel olarak anlamlı derecede daha yüksek olduğu belirlenmiştir.

Sonuç: DeepSeek v3.2, doğruluk açısından ChatGPT-5'e kıyasla daha iyi performans göstermiştir. Yanıtların tamlık puanları bakımından modeller arasında istatistiksel olarak anlamlı bir farklılık bulunmamıştır. Bununla birlikte, FRES sonuçları DeepSeek v3.2'nin diğer sohbet botlarına göre daha yüksek okunabilirliğe sahip metinler ürettiğini göstermiştir. Genel olarak, bu bulgular DeepSeek v3.2'nin hem doğruluk hem de okunabilirlik açısından üstün performans sergilediğini ortaya koymaktadır.

Anahtar Kelimeler: Yapay zekâ, büyük dil modelleri, diş beyazlatma, diş hekimliği

***Corresponding author/İletişim kurulacak yazar:** Suzan Cangül; Dicle University, Faculty of Dentistry, Department of Restorative Dentistry, Diyarbakir, Türkiye.
Phone/Telefon: +90 (555) 888 61 51, **e-mail/e-posta:**suzanbali@outlook.com
Submitted/Başvuru: 12.05.2026 • **Accepted/Kabul:** 22.06.2026 • **Published Online/Online Yayın:** 24.07.2026

Introduction

Artificial intelligence (AI) is a term that defines machine or software systems that have been developed to perform tasks that require human cognitive skills, such as decision-making, problem-solving, and learning from experience.[1,2] AI is based on large language models (LLM), usually trained with the deep learning technique, and these models form the basis of the Natural Language Processing (NLP) approach, a sub-field of AI that allows machines to understand, analyze, and create texts in the same way as a human. LLMs are neural network-based models that have been intensely trained on Wikipedia, digitized books and articles, and other web pages on the Internet. The main aim of these models is to create consistent, human-like responses to questions or requests.[3]

In recent years, LLMs have attracted the interest of healthcare professionals because of their NLP capabilities. Having become a focus of research in the field of medicine, LLMs present opportunities for significant development. In clinical applications, LLMs can help doctors determine clinical diagnoses and treatments and interpret the results of laboratory tests. In scientific research, LLMs can contribute to the productivity of research by assisting in data analysis and article writing. In addition, LLMs can simulate real patients in medical education and function as virtual teaching assistants by preparing learning programs.[4,5]

Currently, there are different language models with various characteristics.[3] The first ChatGPT model, which was introduced by Open AI in 2022, attracted great attention in both academic and practical areas, and its popularity increased rapidly in a short time.[6] ChatGPT-5 was presented for use on August 7, 2025, and was described by the developers as an AI model characterized by intelligence and processing efficiency. According to the manufacturing firm, ChatGPT-5 is one of the most advanced AI models developed for healthcare. Although positive results have been obtained in studies evaluating the use of ChatGPT-5 in clinical practice, potential negative effects, such as confidentiality, ethics, and discrimination, must not be ignored. The text produced by the model may sometimes contain information that is incorrect or not based on reality. This is known as "hallucination" which means that the model produces fictional or incorrect content that is not based on reality. For ChatGPT to be safely used, a thorough control process and human participation in the workflow are required. It is also of great importance for ChatGPT to conform to standards such as accuracy, reliability, interpretability and explainability.[6-8] Google Gemini, developed by Google DeepMind, is a multimodal LLM that can simultaneously process different data forms, such as text, images, audio, and code.[9,10] Google Gemini contains different LLM models, formed from the three main versions of Gemini Nano, Gemini Pro, and Gemini Ultra. These three versions were developed to accommodate diverse user requirements and usage contexts. Nano was designed to enhance accessible and enable efficient processing on smartphones, whereas Ultra represents the most advanced version, incorporating the full spectrum of Google's artificial intelligence capabilities. However, as Pro is a combined version of the other two, it is the most balanced version in terms of use and AI capabilities.[11,12]

DeepSeek v3.2, developed by DeepSeek AI in 2025, is an advanced LLM built on the previous V3.[13] Owing to its high performance in advanced reasoning abilities and various complex tasks, DeepSeek has attracted great interest. As it can compete with the ChatGPT-4o and GPT-o1 models and shows better performance than these models in some areas, DeepSeek has become a significant turning point, not only in terms of LLM capabilities but also in the race for global AI development. In areas such as logical reasoning, mathematical problem solving, and coding, DeepSeek shows significant advances, leaving many competitors lagging behind in cognitive tasks.[14,15] Claude 3.5 Sonnet, represents the initial release within the Claude 3.5 model series, was launched on the market on June 21, 2024. Surpassing GPT-4, Gemini 1.5, and Claude 3 Opus, this model has shown a significant advance in terms of presenting more rapid responses.[16] Using the Reinforcement Learning Based on Human Feedback (RLHF) approach, Claude Sonnet 3.5 was developed as an AI model based on adaptation to structural reasoning and human values. On September 29, 2025, Claude Sonnet 4.5 was presented for use by Anthropic. This version is distinguished from previous Claude models by significant improvements, especially in coding performance, long-term autonomous task management capacity, and security. The model was also designed to preserve connective consistency, provide continuity in dialogues, and optimize adaptability specific to the field. Owing to these characteristics, it is evaluated as an AI system suitable for medical guidance applications that require high accuracy.[17-19]

Tooth whitening is a procedure frequently requested by the public in current aesthetic dentistry. In accordance with the increasing demand for whiter teeth and an aesthetically perfect smile, various tooth-whitening methods have been developed. These include products that can be applied at home, such as toothpastes and gels, and office-type applications of high-concentration whitening agents used under professional supervision.[20] The success of these different methods shows variability depending on the etiology and type of dental discoloration being treated. Tooth discoloration occurring in teeth can be classified into two groups: internal and external staining. Internal staining may be associated with factors such as genetics, age, antibiotics, and developmental disorders and can start before the teeth erupt. External discoloration may be caused by smoking, pigments in food and drinks, and environmental factors such as iron or copper. Pigments or chromogenic components originating from these sources can cause tooth discoloration by penetrating the tooth surface or enamel layer.[21,22] The elimination of chromophores is performed by the oxidation of organic components using

oxidizing substances such as hydrogen peroxide, which is actively found in many bleaching agents. This agent has a high level of solubility in water, and the solution formed has an acidic pH value. Due to the low molecular weight, it also has the property of penetration of enamel and dentin tissues at a high level. It is generally used at a concentration of 35-38% in clinical applications and–1.5-7.5% in home-type whitening procedures.[23,24] Carbamide peroxide is another agent used in whitening treatment. It is the most frequently used bleaching agent in home-type whitening applications, generally at a concentration in the range of 10%-22%, but can be used at a concentration of up to 35% in office-type whitening procedures.[25]

No study of LLMs on dental whitening treatments could have been found in the dentistry literature. The aim of this study was to evaluate and compare the responses of ChatGPT-5, Gemini 2.5 Pro, DeepSeek v3.2, and Claude Sonnet 4.5 models to open-ended questions related to the diagnosis, etiology, and treatment approaches for dental discoloration, and thereby reveal the evidence-based accuracy, completeness, and readability of the responses of these models. The null hypothesis of this study proposed that there would be no statistically significant differences among ChatGPT-5, Gemini 2.5 Pro, DeepSeek v3.2, and Claude Sonnet 4.5 regarding the accuracy, completeness, and readability of their responses to dentistry-related whitening questions.

Methods

This study was conducted in accordance with the ethical principles of the Helsinki Declaration and did not require Ethics Committee approval. An overview of the study design is presented in Figure 1.

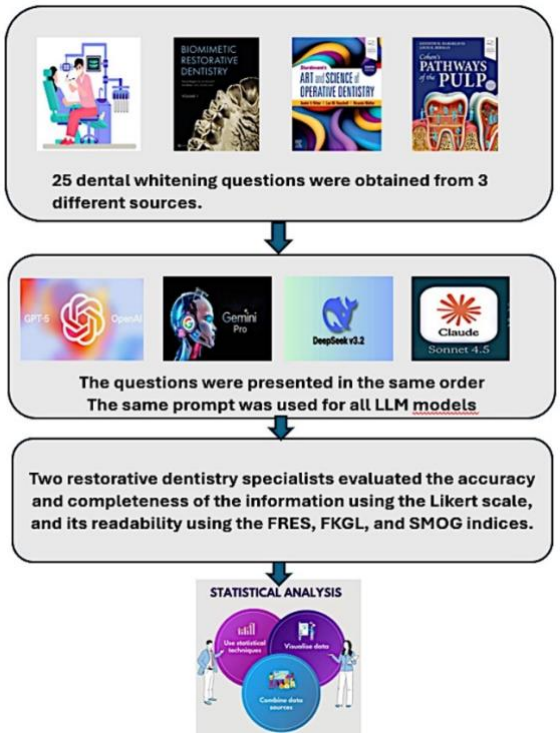

Figure 1. Overview of the research design

A total of 45 open-ended questions related to dental whitening procedures were prepared. The scope and applicability of the questions, which were designed to guide specialists providing whitening treatments, were developed through consultation with two restorative dental treatment specialists. The specialists shared their opinions to provide guidance appropriate to the aim of the study. In accordance with the feedback obtained, some statements were simplified or removed to increase the comprehensibility of the questions. Of the 45 open-ended questions, 25 were selected for the study (Table 1).

Table 1. 25 open-ended questions selected for the study

	Questions
1.	What is the etiology of external and internal tooth discoloration?
2.	What are the vital bleaching techniques used in dentistry, and how are they performed?
3.	Is vital bleaching of teeth safe and appropriate for children and pregnant women?
4.	Which tooth bleaching method is the most cost-effective and easiest to apply for patients?
5.	What is the appointment frequency for in-office vital bleaching in dentistry, and how long does the procedure take?
6.	What is the effect of laser use during non-vital tooth bleaching in dentistry?
7.	How are individual whitening trays used in home bleaching applications prepared in terms of clinical stages and laboratory procedures?
8.	When should treatment be performed on teeth that require restoration after bleaching?
9.	How long does it take for teeth to return to their original color after vital tooth bleaching?
10.	Which tooth color responds best to vital bleaching?
11.	What is the daily application time and total treatment time for at-home vital tooth bleaching?
12.	How should we treat a chemical burn on the gums during dental bleaching?
13.	What precautions should be taken to protect against the adverse effects of the at-home bleaching technique in dentistry?
14.	What are the side effects of the materials used in home bleaching techniques in dentistry?
15.	What is the most effective treatment for tetracycline staining in teeth resistant to dental bleaching?
16.	What should be considered during the restoration phase of a tooth that has undergone non-vital bleaching?
17.	How is the Walking Bleaching technique applied in dentistry?
18.	What factors determine the prognosis of the Walking Bleach technique in dentistry?
19.	What are the therapeutic effects of carbamide peroxide on oral tissues?
20.	What is the function of the urea component in a 10% carbamide peroxide formulation used in dentistry?
21.	What are the differences between bleaching agents containing 10% hydrogen peroxide and agents containing 10% carbamide peroxide used in dentistry?

22.	What materials can be placed in the pulp chamber when bleaching endodontically treated teeth, and what is the most effective mixture?
23.	What can be done to prevent cervical resorption of teeth during nonvital bleaching?
24.	What is calcific metamorphosis and how is it treated?
25.	How is sensitivity in teeth after bleaching treated?
	Total 25 Questions

All questions were developed in accordance with the principles of scientific validity and clinical significance. When determining the content of the questions, guidance was taken from "Ultraconservative Treatment Options" (chapter 3) of Biomimetic Restorative Dentistry QUINTESSENCE PUBLISHING (1st edition), "Additional Conservative Esthetic Procedures" (chapter 9) of Sturdevant's Art and Science of Operative Dentistry ELSEVIER (7th edition), and "Bleaching Procedures" (Chapter 26) of Cohen's Pathway of the Pulp (12th edition).[26-28] The questions were presented to the ChatGPT-5, Gemini 2.5Pro, DeepSeek v.3.2, and Claude Sonnet 4.5 AI models by a volunteer using newly created accounts. To minimize the potential effects of previous responses, a new chat session was started for each question, and only the first responses were analyzed. This approach was adopted because the probabilistic characteristics of large language models may lead to variability in outputs when the same prompt is repeated. Consequently, the "regenerate response" function was not utilized in this study. The responses were recorded in a Microsoft Word file (Microsoft, Redmond, WA, USA) for later analysis. Before presentation to the researchers, the responses of each chatbot were color-coded for differentiation. Thus, the examinations of the responses by the researchers were made blinded with respect to which chatbot they belonged.

The accuracy and completeness of all responses were evaluated separately by two restorative dental treatment specialists. For the evaluation of accuracy, a 5-point Likert scale was used,[29] and for completeness, a 3-point Likert scale was used (Table 2).

Table 2. Likert scales used for accuracy and completeness evaluation

Scale	Score	Level	Definition
Accuracy (5-point Likert scale)	1	Very Poor	The answers are inaccurate or inconsistent with established facts, and important details are missing.
	2	Poor	Some information is accurate; however, fundamental content deficiencies are present.
	3	Fair	The content contains generally accurate elements but has notable shortcomings and lacks sufficient clarity in some aspects.
	4	Good	The content is largely accurate and consistent, with only minor shortcomings.
	5	Excellent	The content demonstrates a high level of accuracy, comprehensive coverage, and a well-structured presentation.
Completeness (3-point Likert scale)	1	Incomplete	The content fails to address the basic components of the topic.
	2	Adequate	The content addresses all essential elements of the topic.
	3	Comprehensive	All essential elements are covered, and additional relevant information is provided.

Accurate and complete responses encompassing all the information related to the subject were evaluated as 5 points, and responses that were incompatible with the literature, contained severe errors, or could be potentially harmful to a patient were evaluated as 1 point. Readability was measured using the Flesch Reading Ease Score (FRES) and the Flesch-Kinkaid Grade Level (FKGL) indexes, which have often been used for this purpose in the literature (Table 3).

Table 3. The FRES and FKGL points obtained by the AI models (30)

FRES	Reading Level	FKGL	Estimated Reading Graduate Level
90–100	Very Easy	5	5th Grade
80–89	Easy	6	6th Grade
70–79	Quite Easy	7	7th Grade
60–69	Standard	8–9	8th–9th Grade
50–59	Quite Difficult	10–12	High school
30–49	Difficult	13–16	University Level
0–29	Very Difficult	>16	University Graduate

The FRES, FKGL, and SMOG indices were selected because they are among the most widely used and validated readability measures for English-language health information. These indices estimate reading difficulty based on sentence length, word length, and syllable structure and have been extensively applied in studies evaluating patient education materials and AI-generated health information. Since all study questions were submitted in English and readability analyses were performed directly on the original English responses generated by the AI models without translation or modification, the use of these readability formulas was considered appropriate for the present study. In the FRES index, scores are distributed on a scale from 0 to 100, where lower values represent greater linguistic complexity. For example, a text with a FRES score in the range of 0-10 points has extremely low readability, and only individuals with an advanced educational level would be able to understand the text. In contrast, a text with a score of 90-100 points can be read extremely easily and can even be comfortably understood by readers with only a primary school education level. The

FKGL is another scale used in the evaluation of text readability, with scores ranging from 0 to 18 points, where each score shows the duration of education required to understand the text.[30-32] All the readability calculations were performed using an online tool, defined in the web address https://goodcalculators.com/flesch-kincaid-calculator.
The FRES points were evaluated using the following formula:
Flesch Reading Ease Score = 206.835 − 1.015 × (Total Words / Total Sentences) − 84.6 × (Total Syllables / Total Words)
The SMOG index is a scale that estimates the approximate education level required to understand a text and focuses especially on the frequency of words of three or more syllables in a standard sample of 30 sentences. High SMOG points indicate that the text is more complex and contains terms at a more technical or specialist level. The evaluation of each response is reported as both the number and percentage of complex words.[30] The calculations were performed using an online tool, defined in the web address https://www.webfx.com/tools/read-able/ using the following formula.
Formula:
1.0430 x sqrt (30 x complexWords/sentences) + 3.1291

Statistical Analysis

Inter-rater reliability was assessed using the intraclass correlation coefficient (ICC). Data normality was evaluated with the Shapiro–Wilk test. Normally distributed variables were expressed as mean ± standard deviation, whereas non-normally distributed data were reported as median (minimum–maximum). Group comparisons were con-ducted using one-way ANOVA or the Kruskal–Wallis test, as appropriate. When overall significance was observed, post hoc analyses were performed using Bonferroni or Dunn–Bonferroni corrections. Correlations between scores were examined using Spearman's correlation coefficient. All statistical analyses were performed using IBM SPSS Statistics (Version 25.0; IBM Corp., Armonk, NY, USA), with a significance level set at p<0.05.

Results

Consistency between the two evaluators in respect of accuracy scores was evaluated using the two-way consistency type single measurement Intraclass Correlation Coefficient (ICC) [ICC (2,1)]. In the ICC analyses of the evaluators' agreement of the four models, the agreement level for all models was found to be statistically significant. The ICC value was calculated to be 0.864 (95% CI: 0.716-0.938, p<0.001) for DeepSeek v3.2, 0.801 (95% CI:0.322-0.929, p=0.002) for ChatGPT-5, 0.892 (95% CI: 0.756-0.952, p<0.001) for Gemini 2.5Pro, and 0.891 (95% CI:0.772-0.950, p<0.001) for Claude Sonnet 4.5. The highest agreement was obtained for Gemini 2.5 Pro and Claude Sonnet 4.5 (ICC: 0.892 and 0.891, respectively), and the lowest was for the ChatGPT-5 model (ICC: 0.801) (Table 4).

Table 4. Inter-rater agreement for accuracy scores assessed using ICC (2,1)

Model	ICC (2,1)	95%CI	p-value
DeepSeek v3.2	0.864	0.716–0.938	<0.001
ChatGPT-5	0.801	0.322–0.929	0.002
Gemini 2.5 Pro	0.892	0.756–0.952	<0.001
Claude Sonnet 4.5	0.891	0.772–0.950	<0.001

The consistency between the two evaluators with respect to the completeness scores was evaluated using the two-way consistency type single measurement Intraclass Correlation Coefficient (ICC) [ICC (2,1)]. In the ICC analyses of the evaluators' agreement of the four models, the agreement level for all the models demonstrated statistically significant (p<0.001). The ICC value was calculated to be 1.000 (95% CI: 1.000-1.000 p<0.001) for DeepSeek v3.2, 0.733 (95% CI:0.396-0.884, p<0.001) for ChatGPT-5, 0.749 (95% CI: 0.495-0.883, p<0.001) for Gemini 2.5Pro, and 0.871 (95% CI:0.724-0.941, p<0.001) for Claude Sonnet 4.5. The highest agreement was obtained for DeepSeek v3.2 (ICC: 1.000) and the lowest for ChatGPT-5 (ICC: 0.733) (Tablo 5).

Table 5. Inter-rater agreement for completeness scores assessed using ICC (2,1)

Model	ICC (2,1)	95%CI	p-value
DeepSeek v3.2	1.000	1.000-1000	<0.001
ChatGPT-5	0.733	0.396–0.884	<0.001
Gemini 2.5 Pro	0.749	0.495–0.883	<0.001
Claude Sonnet 4.5	0.871	0.724–0.941	<0.001

CI: Confidence Interval

In the analysis performed to examine how the AI-based chatbots could enable dentists to respond to questions about tooth whitening from a more knowledgeable and critical perspective, the accuracy points of the responses of the four AI models were compared, and there were determined to be statistically significant differences between these scores (p=0.042). In the subgroup analyses, the median accuracy points of the responses to the questions about tooth whitening were determined to be statistically significantly higher for DeepSeek v.3.2 (5; min:2 -max:5) than for ChatGPT-5 (4; min:1-max:5) (p=0.047). No statistically significant variances were identified across the remaining paired group evaluations (p>0.05) (Figure 2).
In the analysis performed to examine how the AI-based chatbots could enable dentists to respond to the questions about tooth whitening from a more knowledgeable and critical perspective, the completeness points of the responses of the four AI models were compared, and there were determined to

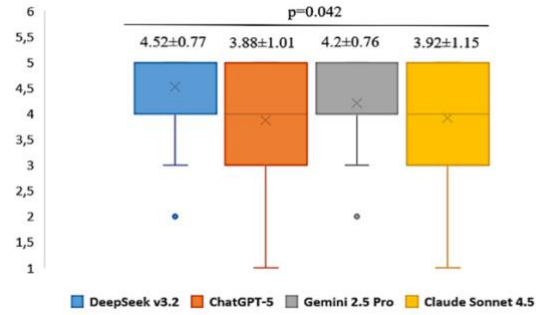

Figure 2. Mean accuracy scores of DeepSeek v3.2, ChatGPT-5, Gemini 2.5 Pro, and Claude Sonnet 4.5

be no statistically significant variances among these data points (p=0.165). In the analysis performed to examine how the AI-based chatbots could enable dentists to respond to the questions about tooth whitening from a more knowledgeable and critical perspective, the FRES points of the responses of the four AI models were compared and there were determined to be statistically significant differences between these scores (p<0.001). In the subgroup analyses, the mean FRES points of DeepSeek v3.2 (42.02±7.63) were determined to be statistically significantly higher than those of ChatGPT-5 (29.09±9.07), Gemini 2.5 Pro (33.92±9.24), and Claude Sonnet 4.5 (21.65±9.29) (p<0.001, p=0.010, p<0.001, respectively). The mean FRES points of ChatGPT-5 (29.09±9.07) and Gemini 2.5 Pro (33.92±9.24) were observed to be significantly greater than those of Claude Sonnet 4.5 (21.65±9.29) (p=0.022, p<0.00, respectively). No significant difference was determined in the other paired group comparisons (p>0.05) (Figure 3).

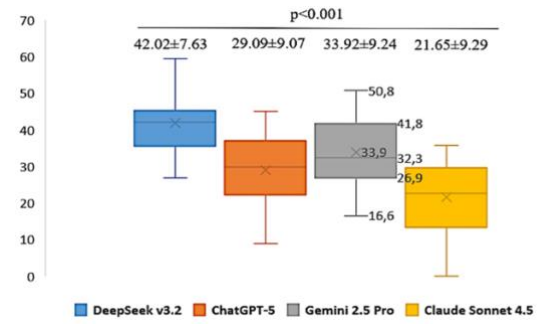

Figure 3. Mean Readability-FRES scores of DeepSeek v3.2, ChatGPT-5, Gemini 2.5 Pro, and Claude Sonnet 4.5.

In the analysis performed to examine how the AI-based chatbots could enable dentists to respond to the questions about tooth whitening from a more knowledgeable and critical perspective, the FKGL points of the responses of the four AI models were compared and there were determined to be statistically significant differences between these scores (p<0.001). In the subgroup analyses, the mean FKGL points of DeepSeek v3.2 (10.21±1.24) were determined to be statistically significantly higher than those of the Gemini 2.5 Pro (12.75±1.84) and Claude Sonnet 4.5 (12.49±1.49) models (p<0.001, p<0.001, respectively). The mean FKGL points of ChatGPT-5 (11.2±1.48) were observed to be significantly higher than those of Gemini 2.5 Pro (12.75±1.84) and Claude Sonnet 4.5 (12.49±1.49) (p=0.003, p=0.021, respectively). No significant difference was determined in the other paired group comparisons (p>0.05) (Figure 4).

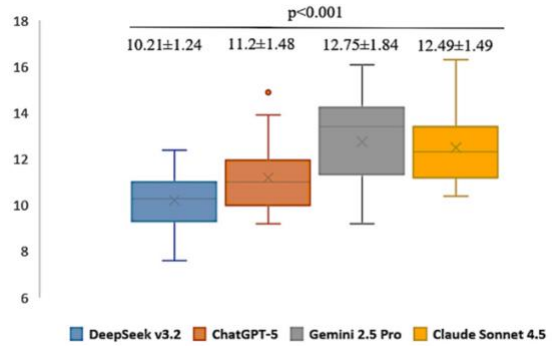

Figure 4. Mean Readability-FKGL scores of DeepSeek v3.2, ChatGPT-5, Gemini 2.5 Pro, and Claude Sonnet 4.5

In the analysis performed to examine how the AI-based chatbots could enable dentists to respond to the questions about tooth whitening from a more knowledgeable and critical perspective, the SMOG points of the responses of the four AI models were compared and there were determined to be statistically significant differences between these scores (p<0.001). In the subgroup analyses, the median SMOG points of Claude Sonnet 4.5 (23.9; min: 12.8-max: 44.3) were determined to be statistically significantly higher than those obtained by DeepSeek-V3.2 (10.9; min: 7.9-max: 14.8), ChatGPT-5 (10.5; min: 8.4-max: 18.5), and Gemini 2.5 Pro (12.6; min: 9.5-max: 15.1) (p<0.001). No significant difference was determined in the other paired group comparisons (p>0.05) (Figure 5) (Table 6).

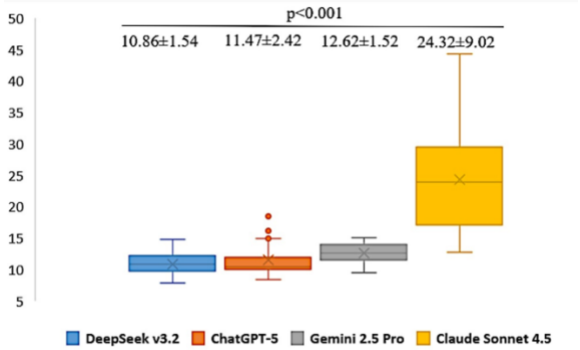

Figure 5. Mean Readability-SMOG scores of DeepSeek v3.2, ChatGPT-5, Gemini 2.5 Pro, and Claude Sonnet 4.5.

Table 6. Comparisons of the accuracy, completeness and readability scores for DeepSeek v3.2, ChatGPT-5, Gemini 2.5 Pro, and Claude Sonnet 4.5

		Chatbot				
		DeepSeek v3.2	**ChatGPT-5**	**Gemini 2.5 Pro**	**Claude Sonnet 4.5**	**p-value**
Accuracy	Mean ± SD	4.52±0.77	3.88±1.01	4.2±0.76	3.92±1.15	**0.042[a]**
	Median (min-max)	5(2-5)	4(1-5)	4(2-5)	4(1-5)	
Completeness	Mean± SD	2.88±0.44	2.64±0.64	2.84±0.47	2.64±0.64	0.165[a]
	Median (min–max)	3 (1-3)	3 (1-3)	3 (1-3)	3 (1-3)	
Readability						
FRES	Mean ± SD	42.02±7.63	29.09±9.07	33.92±9.24	21.65±9.29	**<0.001[b]**
	Median (min–max)	42.2 (26.9-59.6)	29.8 (8.9-45.2)	32.3 (16.6-50.8)	22.6 (0-35.7)	
FKGL	Mean ± SD	10.21±1.24	11.2±1.48	12.75±1.84	12.49±1.49	**<0.001[b]**
	Median (min–max)	10.3 (7.6-12.4)	11 (9.2-14.9)	13.4 (9.2-16.1)	12.3 (10.4-16.3)	
SMOG	Mean ± SD	10.86±1.54	11.47±2.42	12.62±1.52	24.32±9.02	**<0.001[a]**
	Median (min –max)	10.9 (7.9-14.8)	10.5 (8.4-18.5)	12.6 (9.5-15.1)	23.9 (12.8-44.3)	

Data are expressed as mean ± standard deviation and median (minimum: maximum) values. a: Kruskal Wallis test, b: ANOVA test

A statistically significant positive correlation was observed between accuracy and completeness scores (rs= 0.73; p<0.001), indicating that higher accuracy levels were associated with greater completeness (Table 7).
A statistically significant negative correlation was found between the FRES and FKGL readability scores (rs=-0.82; p<0.001). As the FRES points increased, there was a significant decrease in the FKGL values, showing that as the readability of the text increased, the level of reading difficulty decreased. A statistically significant negative correlation was found between the FRES and SMOG readability scores (rs=-0.67; p<0.001). As the readability of the text increased (higher FRES scores), the SMOG values, which refer to the grade-level requirement of the text, decreased. A statistically significant positive correlation was found between the FKGL and SMOG readability scores (rs=0.68; p<0.001). An increase in the FKGL and SMOG values together showed that a higher education level was required to read the text and thus the text had become more difficult to read (Table 8).

Table 7. Spearman correlation analysis between the accuracy, completeness and readability scores for DeepSeek v3.2, ChatGPT-5, Gemini 2.5 Pro, and Claude Sonnet 4.5

		Accuracy	**Completeness**
Completeness			
	r_s	0.73	-
	p value	<0.001	-
FRES			
	r_s	0.04	0.02
	p value	0.664	0.884
FKGL			
	r_s	-0.14	-0.09
	p value	0.174	0.369
SMOG			
	r_s	-0.16	-0.14
	p value	0.106	0.180

FRES: Flesch–Kincaid Reading Ease Score, FKGL: Flesch–Kincaid Grade Level, SMOG: Simple Measure of Gobbledygook.
rs: Spearman's correlation coefficient

Table 8. Spearman correlation analysis between readability scores

		FRES	**FKGL**
FRES			
	r_s	-	-
	p value	-	-
FKGL			
	r	-0.82	-
	p value	<0.001	-
SMOG			
	r_s	-0.67	0.68
	p value	<0.001	<0.001

FRES: Flesch–Kincaid Reading Ease Score, FKGL: Flesch–Kincaid Grade Level, SMOG: Simple Measure of Gobbledygook.
r: Pearson correlation coefficient, rs: Spearman's correlation coefficient

Discussion

The evaluation of AI-based chatbots that aim to provide information about dental whitening treatments plays an important role in the safe use of virtual assistants in responding to questions about diagnosis and treatment. The results of this study demonstrated no difference between ChatGPT-5, Gemini 2.5 Pro, DeepSeek v3.2, and Claude Sonnet 4.5 in terms of the accuracy, completeness, and readability of the responses related to whitening treatments in dentistry; thus, the null hypothesis was partially rejected, because while significant differences were observed between the chatbots when accuracy and readability values were compared, no difference was seen in respect of completeness. AI is an advanced scientific branch that enables computers to perform tasks that have traditionally been performed by humans. Although still in the initial stages of development, LLMs have become an indispensable part of modern life and are used in almost all sectors. Accordingly, LLM applications can benefit the workflow of all specialist areas of dentistry in terms of diagnosis confirmation, treatment planning, and efficiency.[33] Özcivelek et al. asked four chatbots

(DeepSeek-R1, ChatGPT-o1, ChatGPT-4, and Dental GPT) 35 questions about dental prostheses that were most frequently asked by patients. The responses were evaluated in respect of accuracy, quality, comprehensibility, and clinical applicability by two prosthesis specialists using a 5-point Likert scale, Global Quality points, and the Patient Education Materials Evaluation Tool for Printed Materials. While ChatGPT-4 showed the lowest points, DeepSeek-R1 showed the best performance in terms of readability.[34] Freitas et al. evaluated the performance of DeepSeek-R1 in diagnosing oral diseases using text-based case explanations in the "Image Challenge" of the New England Journal of Medicine. It was seen that DeepSeek-R1 reached diagnosis accuracy of 91.6%, slightly exceeding the rate of 88.9% of ChatGPT-4o.[35] In another study, three different LLM models were used, namely ChatGPT-4, Gemini, and DeepSeek-V3, to respond to 194 multiple-choice questions. To evaluate the learning skills of the three AI models over time, the first and last responses were compared. While DeepSeek-V3 showed the highest initial accuracy at 69.6%, the accuracy rates for Gemini were 57.2% and 52.1% for ChatGPT-4. At the end of the study, minimal changes were observed in the responses provided by the chatbots.[36] Although the more recent versions of ChatGPT-5 and Gemini 2.5 Pro of these models were used in the current study, the results were seen to be similar with DeepSeek v.3.2 showing a better accuracy performance than ChatGPT-5. The superior accuracy performance of DeepSeek v3.2 may be associated with its stronger reasoning capabilities and its tendency to generate more focused and direct responses. In contrast, some responses generated by ChatGPT-5 included broader explanations that occasionally contained clinically irrelevant information or recommendations that were not fully aligned with current bleaching protocols. Since the accuracy scoring system prioritized factual correctness and clinical appropriateness rather than response length, concise and evidence-consistent answers may have contributed to the higher scores obtained by DeepSeek v3.2. Moreover, the relatively high performance of DeepSeek v3.2 is also noteworthy from an accessibility perspective. Since free or more easily accessible AI models may be used not only by patients but also by clinicians seeking rapid preliminary support in clinical decision-making, the ability of these models to provide accurate, concise, and evidence-consistent responses is of practical importance. In a study by Zhang et al., the capabilities of four AI models (GPT-4o, Gemini 2.0 Flash, Copilot, and DeepSeek V3) in interpreting periodontal case summaries were evaluated by dentists using a 5-point Likert scale. A total of 258 open-ended questions were used in the study, and each model analyzed the case details and formed responses. DeepSeek V3 obtained the highest accuracy points (0.528), surpassing the other models.[37]

Suarez et al. investigated the accuracy and reliability of the responses generated by ChatGPT-4 in the field of oral and dental surgery. When evaluating the responses to 30 questions about oral surgery, inconsistencies were determined, and an accuracy rate of 71.7% was obtained. These findings show that ChatGPT-4 can play a supportive role in clinical decision-making processes in dentistry. However, it was emphasized that the model cannot replace the accumulated knowledge and experience of specialist clinicians or their clinical intuition; therefore, it should be evaluated as an assistive tool supporting the decision.[38] Consistent with the findings in the literature, it was determined in the current study that the responses given by ChatGPT-5 to some questions could have a negative effect on the diagnosis and treatment process of clinicians, and this was evaluated as 1 point on the Likert scale.

Gürbüz et al. evaluated the accuracy and completeness of ChatGPT-4o responses related to the management of non-carious cervical lesions. Accuracy was assessed using a 6-point Likert scale, while completeness was measured with a 3-point Likert scale. Their findings revealed a statistically significant association between accuracy and informational completeness in the domains of diagnostic assessment and clinical decision-making.[39] The current study results showed no significant difference among the four AI models with respect to the completeness scores of the responses to the questions about tooth whitening. This can be explained by the models having been trained with similar basic dental knowledge areas and that they were able to transfer basic knowledge, such as whitening protocols, indications, and patient management, at a similar level. The absence of significant differences in completeness scores may indicate that contemporary LLMs have access to a comparable core body of dental knowledge. As a result, most models were able to provide sufficient information regarding indications, bleaching protocols, contraindications, and patient management. This finding suggests that completeness alone may not be sufficient to distinguish the clinical utility of AI-generated responses and should be interpreted together with accuracy and readability measures.

One of the questions evaluated in the present study addressed the safety and appropriateness of vital bleaching in children and pregnant women. This topic is particularly important because current pediatric dentistry recommendations advocate a cautious and conservative approach to bleaching procedures in young patients. Factors such as larger pulp chambers, increased pulpal permeability, and a higher risk of post-treatment sensitivity should be considered when treatment decisions are made. Furthermore, contemporary pediatric dentistry emphasizes minimally invasive management strategies, including preventive approaches, monitoring, microabrasion, resin infiltration, and restorative masking procedures before bleaching is considered. Therefore, responses that highlighted individualized treatment planning, professional supervision, and careful risk-benefit assessment were more consistent with current clinical recommendations.[21,40] These findings suggest that the evaluation of AI-generated responses should take into account not only factual accuracy and completeness but

also their consistency with age-specific clinical guidelines and conservative treatment principles.

Many studies conducted using ChatGPT have reported consistent ability to give correct responses about various medical conditions and procedures. It has been shown that the response content of ChatGPT can be adjusted according to a technical or patient-focused approach. Balel et al. analyzed the responses given by ChatGPT to both patient and technical questions and found a higher success rate in the responses to patient questions than to technical questions.[41,42] Differences between the current study results and findings reported in the literature could be considered to be due to the fact that technical and patient questions were not evaluated in separate categories and differences between the subjects and the AI models used.

High FRES points are accepted as an evidence-based marker showing that texts containing shorter sentences and less complex words can be more easily understood by readers, and this has been confirmed in previous studies.[43]

Das et al. compared the responses of DeepSeek-V3, ChatGPT-4o, and Gemini 2.0 to questions about ophthalmic oncology. The accuracy, completeness, readability, and real-world utility of the chatbot responses were compared. All three models obtained similar accuracy points. DeepSeek V3 provided the most detailed and readable responses (FRES points: 38.0), while ChatGPT-4o produced shorter but more clinically reliable responses.[44] In a study conducted by Taşyürek et al., the performance of ChatGPT-5, Gemini 2.5 Flash, and DeepSeek-V3.1 was assessed regarding their ability to provide accurate, understandable, and readable information on endodontic procedures. In terms of both accuracy and FRES readability, DeepSeek V3.1 obtained higher points than the other models.[7] The finding of the current study results were comparable to those in literature as the mean FRES and accuracy scores of the responses given to questions about teeth whitening were found to be higher for DeepSeek v.3.2 (42.02±7.63) than for the other chatbots. The higher readability scores achieved by DeepSeek v3.2 may be attributed to its tendency to generate shorter sentences and use less complex vocabulary. Such characteristics may facilitate comprehension by patients and non-specialist readers. From a clinical perspective, readable health information is important because it may improve patient understanding, treatment adherence, and informed decision-making.

In the comparisons of the FKGL points, the mean 10.21 score of the DeepSeek v.3.2 model was determined to be statistically significantly lower than the 12.75 points obtained for Gemini 2.5 Pro and 12.49 points for Claude Sonnet 4.5. The mean score of 11.2 obtained by ChatGPT-5 was also observed to be lower than the mean scores of Gemini 2.5 Pro (12.75) and Claude Sonnet 4.5 (12.49). According to these results, more easily readable texts were produced by DeepSeek v.3.2 with a required reading level of approximately 10th grade and by ChatGPT-5 at 11th grade. The responses produced by Gemini 2.5 Pro and Claude Sonnet 4.5 were more complex, requiring a high reading level equivalent to the first year of university.

The SMOG points of the responses to the questions about tooth whitening were determined to be higher for Claude Sonnet 4.5 (23.9) than for the DeepSeek-V3.2 (10.9), ChatGPT-5 (10.5), and Gemini 2.5 Pro (12.6) models. When considering that higher points indicate lower readability, it can be seen that Claude Sonnet 4.5 produced texts that were more difficult to read, whereas ChatGPT-5 had a better readability profile. The substantially higher SMOG scores observed for Claude Sonnet 4.5 suggest that the model frequently employed multisyllabic and technically complex terminology. While this may increase the perceived scientific sophistication of the responses, excessive linguistic complexity may reduce accessibility for patients and limit the effectiveness of AI-generated educational content. For clinicians, overly complex responses may increase the cognitive effort required to rapidly interpret AI-generated information, potentially reducing the model's practical utility as a clinical decision-support tool. In clinical settings, where time-efficient and clearly structured information is essential, responses containing unnecessary technical complexity may hinder rapid evaluation and integration into patient management. In the study by Taşyürek et al., evaluating the readability, accuracy, and completeness of LLM responses, ChatGPT-5 was reported to have obtained lower SMOG points than Gemini 2.5 Flash, Grok 4, and Claude Sonnet 4, and thus to have produced more difficult to read texts.[45]

Limitations

This study had some limitations, primarily that the questions were asked only once, and there were no follow-up questions to obtain details. AI models are dynamically updated and developed with continuous updates. Therefore, when evaluating the findings of this study, it must be considered that the results are limited to a certain time interval, and there may be differences in performance in future versions of the models. The study questions were only addressed in English, which may limit the applicability of the findings to other languages. Because the readability formulas used in this study were originally developed and validated for English-language texts, the readability findings may not be directly generalizable to AI-generated responses produced in other languages. This examination of the performance of AI models was based on the evaluation of a small specialist group, which may limit the applicability of the results obtained. In addition, all evaluations were performed by two specialists in restorative dentistry. Although this expertise was appropriate for the assessment of bleaching-related content, one of the study questions specifically addressed the safety and appropriateness of vital bleaching in children and pregnant women. As this topic involves pediatric dentistry-specific considerations and clinical recommendations, the absence of a pediatric dentistry specialist may have limited the

comprehensiveness of the expert evaluation. Finally, it should be taken into consideration that differences in question structure or word selection could affect the responses.

Recommendations for the Future

The clinical diagnosis process requires specialization and experience at a level that cannot be met by LLMs alone. Before the use of information obtained from LLMs in clinical practice, there is a need for confirmation with reliable research.[46] According to a 2024 survey by the American Medical Association (AMA), approximately two-thirds (66%) of clinicians use AI models despite their known drawbacks.[47] The most recent AMA data related to AI in the healthcare sector emphasized the importance of a detailed evaluation system for LLMs used for educational and training purposes.[48] In the framework of this evaluation, there should be a systematic examination of the accuracy level of the model, clinical applicability, and potential effects on patients. For LLMs to be used effectively, there must be multidisciplinary collaboration between healthcare professionals, computer scientists, regulating institutions, and society. For the clinical application of LLMs, some criteria must be met:

• Comprehensive and detailed evaluations should be performed.

• LLMs must be compatible with confirmed and reliable medical knowledge.

• LLMs must consider all dimensions of complex medical conditions and produce treatment approaches appropriate for these clinical scenarios.

They must demonstrate generalizability or consistent performance under different clinical conditions.[49-51]

The responses provided by the AI models did not include any references. This has been confirmed by other studies. De Sousa et al. found that AI responses were reliable and accurate at the rate of 90.63%, but no references were given supporting the responses.[52] Pan et al. performed an analysis of 100 responses produced by four AI models and reported that the models did not state references.[53] In similar study, Jacob et al. evaluated the responses of chatbots to patient questions about third molar extraction and reported that no references or sources were stated in the responses.[41] The absence of citations and references creates a serious concern about the reliability of studies because it is unknown from which sources of information the LLMs benefit most when providing responses. To increase the reliability of future versions of LLMs, they should include references together with responses for patients and physicians to be able to access more sources.

In conclusion, taking the limitations of this study into consideration, the results demonstrated that DeepSeek v 3.2 showed a better performance than the other LLMs in responding to questions about tooth whitening. However, despite advances and increasing capabilities, it is important to be aware that AI models in their current state should not be used indiscriminately. Although AI is currently available in the healthcare sector, research should continue to ensure its safe and effective application in restorative dentistry. AI models have the potential to be an assistive smart virtual assistant in the future with training according to confirmed reference sources and supervision by specialists in restorative dental treatments; however, they will never replace a restorative dental treatment specialist.

Compliance with Ethical Standards

This study was conducted in accordance with the ethical principles of the Helsinki Declaration and did not require Ethics Committee approval.

Conflict of Interest

The authors declared no potential conflict of insterest.

Author Contribution

SC, Conceptualization, Writing – original draft; OA, Review & editing, Validation; TT, Data curation, Writing – review & editing; MT, Investigation, Validation; HO, Statistical analysis.

References

1. Thurzo A, Urbanova W, Novak B, Czako L, Siebert T, Stano P, et al. Where is artificial intelligence applied in dentistry? A systematic review and literature analysis. *Healthcare*. 2022;10:1234. doi:10.3390/healthcare10071269
2. Acar AH. Can natural language processing serve as a consultant in oral surgery? *J Stomatol Oral Maxillofac Surg*. 2023;125:101724. doi:10.1016/j.jormas.2023.101724
3. Dursun D, Bilici Geçer R. Can artificial intelligence models serve as patient information consultants in orthodontics? *BMC Med Inform Decis Mak*. 2024;24:211. doi:10.1186/s12911-024-02619-8
4. Yu E, Chu X, Zhang W, et al. Large language models in medicine: applications, challenges, and future directions. *Int J Med Sci*. 2025;22:2792–2801. doi:10.7150/ijms.111780
5. Xu X, Chen Y, Miao J. Opportunities, challenges, and future directions of large language models, including ChatGPT, in medical education: a systematic scoping review. *J Educ Eval Health Prof*. 2024;21:6. doi:10.3352/jeehp.2024.21.6
6. Eggmann F, Blatz MB. ChatGPT: Chances and challenges for dentistry. *Compend Contin Educ Dent*. 2023;44(4):220–224.
7. Taşyürek M, Adıgüzel Ö, Gündoğar M, Goncharuk-Khomyn M, Ortaç H. Comparative evaluation of the responses from ChatGPT-5, Gemini 2.5 Flash, and DeepSeek-V3.1 chatbots to patient inquiries about endodontic treatment in terms of accuracy, understandability, and readability. *International Dental Research*. 2025;15(3):91–95. doi:10.5577/intdentres.662.
8. İlter Er Ö, Adıgüzel Ö, Taşyürek M, Ortaç H. Comparison of ChatGPT-5, Gemini 2.5 Pro, and DeepSeek-V3.1 chatbot responses to dental avulsion injuries. *Health Sci Monit*. 2025;3:e251201. doi:10.5577/hesmon.e251201
9. Carlà MM, et al. Testing the power of Google DeepMind: Gemini versus ChatGPT-4 facing a European ophthalmology examination. *AJO Int*. 2024;1:100063. doi:10.1016/j.ajoint.2024.100063

10. Masalkhi M, Ong J, Waisberg E, Lee AG. Google DeepMind's Gemini AI versus ChatGPT: a comparative analysis in ophthalmology. *Eye*. 2024;38:1412–1417. doi:10.1038/s41433-024-02958-w
11. Imran M, Almusharraf N. Google Gemini as a next-generation AI educational tool: a review of emerging educational technology. *Smart Learn Environ*. 2024;11:22. doi:10.1186/s40561-024-00310-z
12. Team G, Anil R, Borgeaud S, Wu Y, Alayrac JB, Yu J, et al. *Gemini: a family of highly capable multimodal models*. 2023. doi:10.48550/arXiv.2312.11805
13. DeepSeek AI. DeepSeek-V3 technical report. *arXiv*. 2025; arXiv:2412.19437.
14. Ye J, et al. DeepSeek in healthcare: a survey of capabilities, risks, and clinical applications of open-source large language models. *arXiv*. 2025; arXiv:2506.01257. doi:10.48550/arXiv.2506.01257
15. Deng Z, Han Q, Zhou W, Zhu X, Wen S, Xiang Y. Exploring DeepSeek: advances, applications, challenges, and future directions. *IEEE/CAA J Autom Sin*. 2025;12:872–893. doi:10.1109/JAS.2025.125498
16. Jin H, et al. Comparative study of Claude 3.5 Sonnet and human physicians in generating discharge summaries. *Front Digit Health*. 2024;6:1456911.doi: 10.3389/fdgth.2024.1456911
17. Nasirov R. The role of Claude 3.5 Sonnet and ChatGPT-4 in posterior cervical fusion patient guidance. *World Neurosurg*. 2025;197:123889. doi:10.1016/j.wneu. 2025.123889
18. Anthropic. Introducing Claude Sonnet 4.5 [Internet]. 2025 [cited 2025 Nov 10]. Available from: https://www.anthropic.com
19. Anthropic. Claude Sonnet: Model overview and performance improvements [Internet]. 2025 [cited 2025 Nov 10]. Available from: https://www.anthropic.com
20. Carey CM. Tooth whitening: what we now know. *J Evid Based Dent Pract*. 2014;14:70–76. doi:10.1016/j.jebdp.2014.02.006
21. Epple M, Meyer F, Enax J. A critical review of modern concepts for teeth whitening. *Dent J*. 2019;7:79. doi:10.3390/dj7030079
22. Silva EMD, et al. Can whitening toothpastes maintain optical stability of enamel over time? *J Appl Oral Sci*. 2018;26:e20160460. doi:10.1590/1678-7757-2016-0460
23. Ubaldini AL, et al. Hydrogen peroxide diffusion dynamics in dental tissues. *J Dent Res*. 2013;92:661–665. doi:10.1177/0022034513488893
24. Pauli MC, et al. Current status of whitening agents and enzymes in dentistry. *Braz J Pharm Sci*. 2022;58:e19501. doi:10.1590/s2175979020201000X32e19501
25. Kihn PW. Vital tooth whitening. *Dent Clin North Am*. 2007;51:319–331. doi:10.1016/j.cden.2006.12.001
26. Simonsen RJ. Ultraconservative treatment options. In: *Biomimetic restorative dentistry.* Berlin: Quintessence; 2021. p. 45–67.
27. Hillard JJ, Robinson PB. Additional conservative esthetic procedures. In: *Sturdevant's art and science of operative dentistry.* 7th ed. St. Louis (MO): Elsevier; 2019. p. 235–260.
28. Haywood VB. Bleaching procedures. In: *Cohen's pathways of the pulp.* 12th ed. St. Louis (MO): Elsevier; 2020. p. 2970–3038.
29. Likert R. A technique for the measurement of attitudes. *Arch Psychol.* 1932;22(140):1–55.
30. Flesch R. A new readability yardstick. *J Appl Psychol*. 1948;32(3):221–233. doi:10.1037/h0057532
31. Esmailpour H, et al. Performance of artificial intelligence chatbots in responding to patient FAQs on dental prostheses. *BMC Oral Health*. 2025;25:574. doi:10.1186/s12903-025-05965-9
32. Daraz L, et al. Readability of online health information: a meta-narrative systematic review. *Am J Med Qual*. 2018;33:487–492. doi:10.1177/1062860617751639
33. Bulut AC, Bahadır HS, Ateş G. Artificial intelligence in dental education: can AI-based chatbots compete with general practitioners? *BMC Med Educ*. 2025;25:1319. doi:10.1186/s12909-025-07880-7
34. Özcivelek T, Özcan B. Comparative evaluation of responses from DeepSeek-R1, ChatGPT-o1, ChatGPT-4, and Dental GPT chatbots. *BMC Oral Health*. 2025;25:871. doi:10.1186/s12903-025-06267-w
35. Diniz-Freitas M, Diz-Dios P. DeepSeek: another step forward in the diagnosis of oral lesions. *J Dent Sci*. 2025;20:1904–1907. doi:10.1016/j.jds.2025.02.023
36. Performance of ChatGPT-4, Gemini, and DeepSeek-V3 on dental licensing examination questions. *J Dent Sci*. 2025;20:2154–2162. doi:10.1016/j.jds.2025.07.011
37. Zhang H, et al. DeepSeek performs better than other large language models in dental cases. *arXiv*. 2025;arXiv:2509.02036. doi:10.48550/arXiv.2509.02036
38. Suárez A, et al. Assessing ChatGPT as an intelligent virtual assistant in oral surgery. *Comput Struct Biotechnol J*. 2023;24:46–52. doi:10.1016/j.csbj.2023.11.058
39. Gurbuz E, Tetik B. Accuracy and completeness of ChatGPT-4o in the management of non-carious cervical lesions. *Digit Dent J*. 2025;2:100015. doi:10.1016/j.ddj.2025.100015
40. American Academy of Pediatric Dentistry. Policy on the use of dental bleaching for child and adolescent patients. In: *The Reference Manual of Pediatric Dentistry.* Chicago (IL): American Academy of Pediatric Dentistry; 2025. p. 136-140. https://www.aapd.org/globalassets/media/policies_guidelines/p_bleaching25.pdf?utm_source=chatgpt.com
41. Jacobs T, et al. Is ChatGPT an accurate and readable patient aid for third molar extractions? *J Oral Maxillofac Surg*. 2024;82:1239–1245. doi:10.1016/j.joms.2024. 06.177
42. Balel Y. Can ChatGPT be used in oral and maxillofacial surgery? *J Stomatol Oral Maxillofac Surg*. 2023;124:101471. doi:10.1016/j.jormas.2023.101471
43. Friedman DB, Hoffman-Goetz L. A systematic review of readability and comprehension instruments used for health information. *Health Educ Behav*. 2006;33:352–373. doi:10.1177/1090198105277329
44. Das D, et al. AI chatbots in ocular oncology: a comparative study. *Cureus*. 2025;17:e90773. doi:10.7759/cureus. 90773
45. Taşyürek M, Adıgüzel Ö, Ortaç H. Comparative evaluation of ChatGPT-5, Gemini 2.5 Flash, Grok 4, and Claude Sonnet-4 in endodontic iatrogenic events. *Healthcare*. 2025;13:2615. doi:10.3390/healthcare13202615

46. Meng X, et al. The application of large language models in medicine: a scoping review. *iScience*. 2024;27:109713. doi:10.1016/j.isci.2024.109713
47. Hoyt RE, et al. Evaluating large reasoning model performance on complex medical scenarios. *medRxiv*. 2025;2025.04. doi:10.1101/2025.04.29.25326666
48. American Medical Association. *AMA issues new principles for AI development, deployment & use* [Internet]. Chicago (IL): American Medical Association; 2023 Nov 28 [cited 2026 Jun 28]. https://www.ama-assn.org/press-center/ama-press-releases/ama-issues-new-principles-ai-development-deployment-use?utm_source=chatgpt.com
49. Moreno AC, Bitterman DS. Toward clinical-grade evaluation of large language models. *Int J Radiat Oncol Biol Phys*. 2024;118:916–920. doi:10.1016/j.ijrobp.2023.11.012
50. Tang L, et al. Evaluating large language models on medical evidence summarization. *medRxiv*. 2023. doi:10.1101/2023.04.22.23288967
51. Johri S, et al. Guidelines for rigorous evaluation of clinical LLMs for conversational reasoning. *medRxiv*. 2023. doi:10.1101/2023.09.12.23295399
52. Aguiar de Sousa R, et al. Is ChatGPT a reliable source of scientific information regarding third-molar surgery? J Am Dent Assoc. 2024;155:227–232.e6. doi:10.1016/j.adaj.2023.11.004
53. Pan A, Musheyev D, Bockelman D, Loeb S, Kabarriti AE. Assessment of artificial intelligence chatbot responses to top searched queries about cancer. *JAMA Oncol*. 2023;9:1437–1440. doi:10.1001/jamaoncol.2023.2947