JOMDI

Journal of Medical and Dental Investigations

J Med Dent Invest 2025; 6: e250161
https://doi.org/10.5577/jomdi.e250161

Original Article Open Access

Comparative evaluation of the accuracy and completeness of responses provided by ChatGPT-5, Gemini 2.5 Flash, Grok 4, and DeepSeek-R1 to open-ended questions on cracked and fractured teeth

Makbule Taşyürek[1], Özkan Adıgüzel[1], Suzan Cangül[2], Hatice Ortaç[3]

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

Received: 26 November 2025
Accepted: 20 December 2025
Published: 30 December 2025

Correspondence:

Dr. Makbule TAŞYÜREK
Dicle University, Faculty of Dentistry, Department of Endodontics, Diyarbakır, Türkiye.
E-mail: makbule.tasyurek@dicle.edu.tr

How to cite this article:

Taşyürek M, Adıgüzel Ö, Cangül S, Ortaç H, Comparative evaluation of the accuracy and completeness of responses provided by ChatGPT-5, Gemini 2.5 Flash, Grok 4, and DeepSeek-R1 to open-ended questions on cracked and fractured teeth. J Med Dent Invest 2025;6:e250161. https://doi.org/10.5577/jomdi.e250161

Abstract

Aim: The aim of this study was to evaluate and compare the accuracy and completeness of responses to open-ended questions related to chronic cracked and fractured teeth given by four recently introduced large language models (LLMs): ChatGPT-5, Grok 4, Gemini 2.5 Flash, and DeepSeek-R1-0528.

Methods: A total of 25 open-ended questions related to cracked and fractured teeth were prepared. The responses to these questions given by the LLMs were evaluated in terms of accuracy and completeness by restorative and endodontic specialists with 15 years of experience. Accuracy was evaluated using a 5-point Likert scale, and completeness was evaluated using a 3-point Likert scale. The Intraclass Correlation Coefficient (ICC), Shapiro-Wilk test, Kruskal-Wallis test, Dunn-Bonferroni test, and Spearman correlation analysis were used in the statistical analyses of the study.

Results: The responses given to the questions about the diagnosis, causes, and treatment of cracked and fractured teeth showed significant differences among the four artificial intelligence (AI) models ($p=0.015$). In the subgroup analysis, the mean points of ChatGPT-5 (5±0) were significantly higher than those of DeepSeek-R1-0528 (4.48±0.77) ($p=0.013$). No significant difference was observed in the other paired comparisons ($p>0.05$). The completeness points also showed a difference between the models ($p=0.028$). The mean points of ChatGPT-5 (3±0) were significantly higher than those of DeepSeek-R1-0528 (2.76±0.44) ($p=0.046$), and no significant difference was observed in the other paired comparisons ($p>0.05$).

Conclusion: Differences in terms of accuracy and completeness were observed between the LLMs in the responses given to open-ended questions about cracked and fractured teeth. The lowest mean points were obtained by DeepSeek-R1-0528, followed by Grok 4 and Gemini 2.5 Flash, with the highest points obtained by ChatGPT-5. It must not be forgotten that the results obtained can show variability according to the subject, the form of the question, and the content. Therefore, further studies encompassing both similar and different subjects are needed. It must also not be ignored that misinformation in healthcare services can lead to serious consequences.

Keywords: Artificial intelligence, large language models, ChatGPT, Gemini, Grok, DeepSeek, cracked tooth, fractured tooth, accuracy, completeness

Introduction

Artificial intelligence (AI) is a field of science encompassing various technologies of computer systems that perform tasks generally associated with human intelligence, such as learning, reasoning, interpretation, and data analysis without any direct human intervention (1).

In recent years, Large Language Models (LLMs), which are a revolutionary technology in the field of AI, have rapidly reshaped areas of medical research and clinical practice (2). Owing to their extraordinary abilities in understanding and producing natural language and information-based reasoning, LLMs are not limited to the effective processing and synthesizing of large amounts of medical data but also show significant potential in critical areas, such as the development of clinical decision support systems, restructuring medical education, and accelerating the discovery of scientific knowledge (3). LLMs can take on the role of a digital assistant in dentistry, and their use could potentially provide a shortening of examination and consultation times, and could be expanded to areas that would reduce the workload for dentists and clinical support personnel. For example, they could be used for tasks such as preparing templates for sending information to patients or care instructions after treatment.

In the case of false information production, where LLMs are sure of incorrect summarizing of references or the production of illogical output, this is defined in the literature as “hallucination” (4). Although advanced LLMs can provide correct responses in the field of healthcare, hallucinations in the responses can cause significant clinical problems through the risk of misdirecting clinicians at the stages of diagnosis and treatment.

ChatGPT, which was launched in 2020, is an AI-based LLM trained on multilingual large text data clusters and has the ability to produce human-like responses to text input (5, 6). The name ChatGPT was given by the developers, Open AI (OpenAI, San Francisco, CA, USA), with the meaning of chatbot (a program that can extract meaning and produce a response using a text-based interface) based on "generative pre-trained transformer" (GPT) architecture (7, 8).

Most medical specialists currently agree that ChatGPT can provide potential benefits in clinical settings because of its large database. However, there is a need for research in respect of the capabilities and potential risks of this new technology. Some researchers have suggested that the free-of-charge version of ChatGPT has limited reliability in the diagnosis and management of real clinical cases (9). The same authors stated that ChatGPT could not differentiate the superiority of some additional tests over others and could not identify atypical conditions in patients with complex medical and surgical histories. Other studies have shown that ChatGPT is a good information tool for patients, but is not sufficiently accurate for healthcare professionals, is far removed from playing a reliable supporting role in clinical decision-making processes, and is only a useful complementary tool (10, 11).

ChatGPT-4.0 was introduced for use in March 2024, and ChatGPT-4o in May2024. The new versions have been reported to provide more rapid, less delayed, and multimodal (text, audio, and visual) responses. ChatGPT-4o presents advances in code and text processing, especially in languages other than English, and the database was updated until 2023. Owing to this broad information base and human-like responses, ChatGPT may be a tool that can be used for educational purposes in the healthcare field (12). However, hallucinations that can emerge in ChatGPT can lead to errors in learning in the education process, which can cause serious problems in patient care after education, such as misdiagnosis and the application of inappropriate treatment.

According to the developers, GPT-5, which was presented for use on August 7, 2025, was constructed using a smart and productive basic model. The new system, named GPT-5 Thinking, works together with a deep thinking mode and presents a combined architecture owing to a guiding component that determines, in real time, which mode will come into operation. It has been emphasized that compared to previous versions, it provides valuable improvements in response speed and accuracy rate, reduces the tendency to hallucinations, and improves the ability to comply with instructions. By minimizing unnecessary confirmatory behavior, it has also been said to establish clearer and more productive interactions with the user. Moreover, GPT-5 has been reported to show significant increases in performance in common areas of use such as writing, coding, and healthcare. Thus, it aims to provide a more reliable, rapid, and functional experience for both daily use and professional use (13). In previous versions of ChatGPT, not only the reliability was debatable, but it has also been reported that misleading content was produced which was actually erroneous although it seemed to be true (4). Therefore, clinicians should continue to carefully examine developments that aim to reduce the tendency to produce incorrect information and the potential for use in the field of healthcare.

DeepSeek (Hangzhou DeepSeek Artificial Intelligence Basic Technology Research Co., Ltd., Beijing, China) is a Chinese-centered AI model that emphasizes the principles of transparency and repeatability. This chatbot, launched in December 2024, created a large reaction worldwide because of its low cost and high download rates. The R1 model can perform human-like reasoning and has a structure that is effective for scientific problems. According to test results published in January 2025, it showed a performance similar to that of the Open AI o1 model in areas such as chemistry, mathematics, and coding (14). With improvements to the existing R1 model, the DeepSeek-R1-0528 version was announced on May 28, 2025. In general, the accuracy, consistency, and reliability of the model were increased, and the hallucination (production of incorrect information) rate was decreased (15).

Another noteworthy LLM model is Google Bard, which was developed by Google. Initially, Bard was supported by the Language Model for Dialogue Applications (LaMDA), which is a member of the language model specific to Google, and was later supported by the Pathways Language Model (PaLM) 2 LLM. Google Bard was launched on the market on 21 March 2023, and was then renamed as Gemini in February 2024. It is an AI-supported information acquisition tool and advanced chatbot that uses a "local multimode" model that can effectively analyze and adapt various data formats, such as text, audio, and video. It can be used free of charge, and an unlimited number of questions can be asked or dialogues established. Gemini offers free-of-charge dialogue AI services using a series of deep learning algorithms and can respond to questionnaires (16, 17). The developers have stated that Gemini 2.5, which was introduced on March 26, 2025, is an advanced new-generation thinking model that aims to solve increasingly complex problems. This model family has the ability for thoughts to pass through a logic filter before the response is produced. This approach offers greater accuracy, advanced performance, and superior problem-solving capabilities for complex tasks. Gemini 2.5 was designed to not only meet the user's needs rapidly, but also to respond with solutions based on more reliable and logical foundations (18).

xAI developed Grok in response to the perceived deficiencies of existing chatbots. This intervention was started by xAI with 33 billion parameters in the Grok 0 model, and this number was increased to 314 billion in later versions, such as Grok 1. By offering a high level of reasoning and real-time information processing capacity, this rapid development rendered Grok extremely useful for dynamic and data-focused applications. The Grok 0 model entered the market with 33 billion parameters and showed competence in mathematical reasoning and humorous subjects. With 314 billion parameters, the Grok 1 model is promising, with real-time links to social media platforms and additional improvements (19). Compared with other LLMs, the real-time integration of Grok with the social media platform X renders it unequalled and adds additional interest to the results (20).

The developers of Grok 4, which was introduced on July 1, 2025, stated that it was trained with reinforced learning methods for tool use. This approach can support the thinking process by effectively using additional tools, such as code interpreters or web screeners, even in situations that are generally difficult for LLMs. Especially for real-time information searches or complex research questions, Grok 4 gathers data from the web by forming its own search questions. When necessary, it analyzes this information in depth and thus produces high-quality, reliable, and comprehensive responses (21).

Although some studies have investigated the use of LLMs in producing questions related to dentistry (22, 23), these studies have focused on the well-known and widely used ChatGPT and Gemini. Moreover, owing to the rapid advances in this field, older LLMs are generally less competent than new models that have multimode features (24). To overcome these limitations, the current study used the most recent and advanced LLMs.

Previous studies have reported that the accuracy of LLMs in responding to open-ended questions varies between 52.5% and 71.7%, and sometimes the responses may be incorrect, overly general, outdated, or lacking evidential support (25, 26).

A cracked tooth is defined as a thin, superficial defect, the depth and extent of which are not known, which occurs in the enamel and dentin layers and possibly in the cement (27). Cracked teeth are commonly observed in patients aged >40 years, particularly in the mandibular molars (28-31). Patients with this condition usually present with many signs and symptoms of varying severity, which presents a challenge for dentists in terms of diagnosis and treatment. In the literature, this is referred to as the "cracked tooth puzzle" (32). In patients with cracked teeth, if the condition is not controlled in time, the crack deepens and can cause pulpitis and periodontal lesions. Therefore, prompt and appropriate diagnosis and treatment are of great importance to protect the affected teeth and decrease pain for the patient (33). However, as a cracked tooth has an occult course in the early stage and because of the confusion of symptoms with pulpitis or periodontal lesions, it can be easily confused with other diseases, which can lead to difficulties in clinical diagnosis and treatment. Therefore, the effective diagnosis and treatment of cracked teeth remains a challenging and important subject of interest for many clinicians (34).

Vertical root fractures (VRF) can develop slowly without any evident signs or symptoms, making differential diagnosis more difficult. Early detection and appropriate management of VRFs increase the chance of preserving the affected tooth, which is vital for minimizing the unwanted outcomes of this complication. Timely extraction of teeth with advanced VRF (full fracture or separated tooth) prevents the formation of pain/discomfort and limits periradicular bone loss, which can affect subsequent implant treatment planning (35).

The diagnosis of cracked and fractured teeth is extremely difficult because of the vagueness of symptoms, the complexity of clinical findings, and that they can be easily confused with similar diseases. This has led to a range of treatment approaches and ongoing debates about specific protocols, making diagnosis and treatment a significant clinical challenge in dental practice.

To the best of our knowledge, no study has been performed on cracked and fractured teeth using LLMs. This study aimed to evaluate and compare the responses of ChatGPT-5, Gemini 2.5 Flash, Grok 4, and DeepSeek-R1-0528 to open-ended questions related to the diagnosis, causes, and treatment of cracked and fractured teeth, and thereby determine the evidence-based knowledge levels and potentials of these models.

The null hypothesis of this study was that the accuracy and completeness of the responses to the questions about chronic cracked and fractured teeth would be similar for ChatGPT-5, Grok 4, Gemini 2.5 Flash, and DeepSeek-R1-0528, and there would be no

statistically significant difference between the models with respect to these performance criteria.

Materials and Methods

This study only evaluated data produced by AI that are in the public domain and did not include materials obtained from humans or animals; therefore, ethical approval was not required.

A total of 25 open-ended questions related to cracked and fractured teeth in dentistry were prepared. The prepared questions were arranged as a guide for specialists to evaluate cracks and fractures in teeth. A systematic and structured method was adopted to minimize uncertainties that could arise in the interpretation of AI-based responses. All the questions were formed considering scientific validity and clinical importance, based on "Chronic Cracks and Fractures" (chapter 22) of Cohen's "Pathways of the Pulp" (12th edition) (36).

The questions were designed to focus on clinically important and practice-based subjects, such as the identification, symptoms, diagnostic methods, and treatment options for cracked and fractured teeth. Each question was both scientifically and professionally valid to shed light on the conditions encountered in the daily practice. The scope and applicability of the questions were carefully examined by an endodontist and a restorative dental treatment specialist, each with 15 years of experience, and their expert opinions contributed to the content of the questionnaire (Table 1).

The questions were asked in accounts newly created by a single researcher on the ChatGPT-5, Gemini 2.5 Flash, Grok 4, and DeepSeek-R1-0528 platforms between 10 and August 12, 2025.

Table 1. List of questions asked of the chatbots.

Questions	
1.	According to the 2016 edition of the American Association of Endodontists (AAE) Endodontic Terminology Dictionary, what is the definition of a cracked or fractured tooth?
2.	Why does the diagnosis of cracked and fractured teeth pose a clinical challenge for dentists?
3.	What does the term "split tooth" mean in dentistry?
4.	What does the term 'fractured tooth fragment' mean in dentistry?
5.	What are the typical complaints of patients with fractured tooth fragments, and what findings are typically observed in their medical history?
6.	How are fractured tooth fragments diagnosed?
7.	How should a dentist treat a patient with a cracked tooth?
8.	Under what circumstances is root canal treatment necessary when treating a patient with a cracked tooth?
9.	How should a dentist treat a patient with a broken tooth?
10.	What symptoms are observed in the early stages of a cracked tooth?
11.	What symptoms are seen in advanced stages of a cracked tooth?
12.	What are the advantages and limitations of the methods used to diagnose cracked or broken teeth? Which of these methods may be more effective in which situations?
13.	What are the causes of tooth cracks?
14.	How should treatment planning be performed for cracked tooth?
15.	What are the typical complaints of patients with vertical root fractures who visit the clinic?
16.	What are the symptoms of vertical root fractures in the early stages?
17.	What is the importance of early diagnosis of vertical root fractures in teeth?
18.	What are the differences between periodontal pockets and pockets formed in vertical root fractures in teeth?
19.	Is CBCT recommended for the diagnosis of vertical root fractures in dentistry?
20.	Why is CBCT not recommended for the diagnosis of vertical root fractures in dentistry?
21.	What are the iatrogenic predisposing factors in the development of vertical root fractures in teeth?
22.	What are the typical appearances of bone resorption patterns associated with vertical root fractures in teeth?
23.	Does reduced hydration after endodontic treatment cause vertical root fractures in teeth?
24.	What is the effect of the cross-sectional shape of the root on the occurrence of vertical root fractures in teeth?
25.	When is surgical intervention indicated for the diagnosis of vertical root fractures in teeth?
Total 25 Questions	

To prevent the responses of the AI-based language models from being affected by previous questions and answers, each question was asked in a separate chat session. By clearing the previous chat from the system at the start of each session, the model focused only on the current question. This method aimed to remove the effect of contextual information or guidance on responses that could be obtained by AI in consecutive chats. Thus, it was ensured that each response produced was independent, unbiased, and based only on the current input. Each LLM language model was allowed to provide only one response, and after the first response was produced, it was prevented from correcting or reproducing it.

After recording the responses collected from the four AI models in Microsoft Word format (Microsoft, Redmond, WA, USA), they were anonymized and forwarded to an endodontist and a restorative dental treatment specialist, each with 15 years of experience. The responses of each chatbot were color-coded before presentation to these specialists; thus, the evaluations were blinded to which responses were given by which models.

"Chronic Cracks and Fractures", which is the 22nd chapter of Cohen's Pathways of the Pulp (12th edition), was used as the reference standard in the evaluation process. This guideline text was presented to the expert evaluators together with the anonymized responses, and to perform structured and consistent scoring, each evaluator used a standardized Excel scoring table (Microsoft, Redmond, WA, USA). The study flow is shown in Figure 1.

The responses obtained from the LLMs were evaluated in a double-blind design by two independent researchers. The evaluations were made using Likert scales in two dimensions (37, 38):

- Accuracy (5-point Likert scale)
- Completeness (3-point Likert scale) (Table 2).

Statistical Analysis

The Intraclass Correlation Coefficient (ICC) values were calculated to evaluate the score agreement between the assessors. The conformity of continuous variables to a normal distribution was examined using the Shapiro–Wilk test. Variables showing normal distribution were stated as mean±standard deviation (SD) values, and those not conforming to normal distribution were stated as median (minimum-maximum) values. According to the normality test results, the ANOVA test was applied in the comparisons of more two groups of data showing normal distribution, and the Kruskal-Wallis test was used when normal distribution was not observed. When there was overall significance, subgroup analyses were performed using the Bonferroni test after ANOVA and Dunn-Bonferroni test after the Kruskal-Wallis test.

The relationships between the scores were examined using the Spearman correlation coefficient.

SPSS for Windows software (27.0; IBM Corp., Armonk, NY, USA) was used for all the statistical analyses. Type I error of 5% was accepted, and the level of statistical significance was set at $p<0.05$.

Table 2. Evaluation dimensions and Likert scales.

Scale Dimension	Points	Definition
Accuracy	1	Very poor: the responses are not appropriate to or consistent with the facts; the most important details are missing.
	2	Poor: There are some correct components but important content is missing or uncertain.
	3	Moderate: Partially accurate, including correct, missing, or uncertain information.
	4	Good: Mostly accurate and consistent, with only a few deficiencies.
	5	Excellent: Extremely accurate, comprehensive, and well organized.
Completeness	1	Incomplete: The basic components of the question are not addressed.
	2	Sufficient: All the important components of the question are addressed.
	3	Comprehensive: The response addresses all the necessary components and includes additional related information.

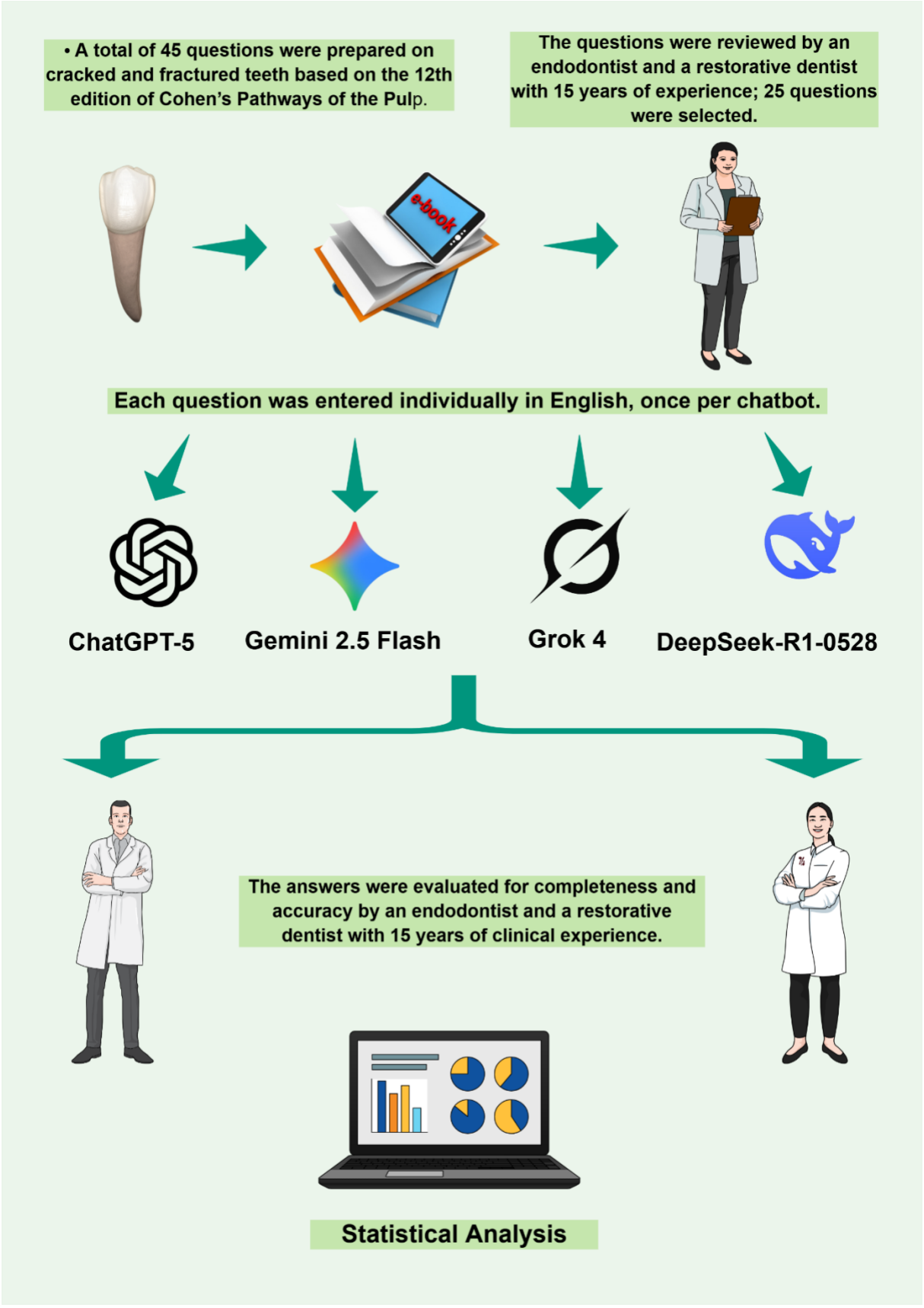

Figure 1. Graphical illustration of the study flow.

Results

The agreement between the evaluators with respect to the accuracy points was examined using the ICC (C,1). The same points were given by both evaluators for all items in ChatGPT-5; therefore, the ICC was not calculated, and it was accepted that there was 100% absolute agreement. For Gemini 2.5 Flash (ICC=0.859), Grok-4 (ICC=0.934), and DeepSeek-R1-0528 (ICC=0.966), agreement was observed to be at a high-excellent level (p<0.001).

The agreement in respect of completeness points was again 100% for ChatGPT-5 and was determined to be excellent for Gemini 2.5 Flash and Grok 4 (ICC=1.00) and high for DeepSeek-R1-0528 (ICC=0.820) (p<0.001).

The agreement between the evaluators was high for all models, and excellent agreement was observed in the ChatGPT-5, Gemini 2.5 Flash, and Grok 4 models in particular.

Statistically significant differences were observed in the accuracy of the responses provided by the four AI models to open-ended questions regarding the diagnosis, causes, and treatments of cracked and fractured teeth (p=0.015). In the subgroup analyses, the mean points of the responses given by ChatGPT-5 (5±0) were seen to be statistically significantly higher than the mean points of DeepSeek-R1-0528 (4.48±0.77) (p=0.013). No statistically significant difference was found in the other paired group comparisons (p>0.05) (Table 3).

Statistically significant differences were observed in the completeness of the responses provided by the four AI models to the open-ended questions regarding the diagnosis, causes, and treatments of cracked and fractured teeth (p=0.028). In the subgroup analyses, the mean points of the responses given by ChatGPT-5 (3±0) were seen to be statistically significantly higher than the mean points of DeepSeek-R1-0528 (2.76±0.44) (p=0.046). No statistically significant difference was found in the other paired group comparisons (p>0.05) (Table 3).

Table 3. Comparison of accuracy and completeness scores for the ChatGPT-5, Gemini 2.5 Flash, Grok-4, and DeepSeek-R1-0528 models.

		Chatbot				
		ChatGPT-5	Gemini 2.5 Flash	Grok-4	DeepSeek-R1-0528	p-value
Accuracy	Mean ± SD	5±0	4.72±0.54	4.56±0.77	4.48±0.77	0.015[a]
	Median (min.-max.)	5(5-5)	5(3-5)	5(3-5)	5(3-5)	
Completeness	Mean ± SD	3±0	2.96±0.2	2.8±0.5	2.76±0.44	0.028[a]
	Median (min.-max.)	3(3-3)	3(2-3)	3(1-3)	3(2-3)	

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

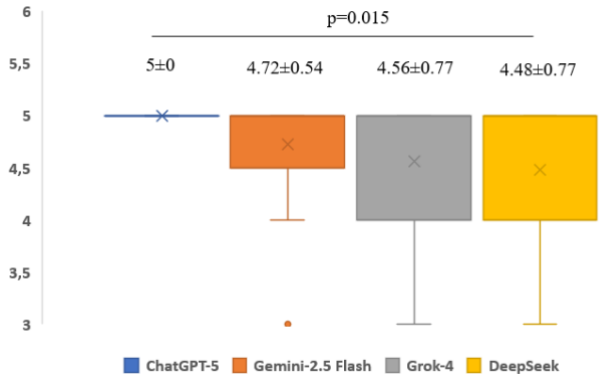

Figure 2. Mean accuracy scores of ChatGPT-5, Gemini 2.5 Flash, Grok-4, and DeepSeek-R1-0528 in response to questions related to chronic cracks and fractures.

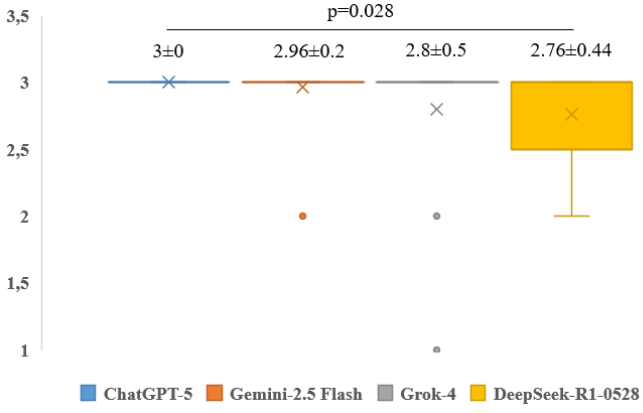

Figure 3. Mean completeness scores of ChatGPT-5, Gemini 2.5 Flash, Grok-4, and DeepSeek-R1-0528 in response to questions related to chronic cracks and fractures.

A statistically significant positive correlation was found between the accuracy and completeness scores (rs=0.73; p<0.001). The findings showed that an increase in the accuracy score was related to an increase in the completeness score (Table 4).

Table 4. Spearman correlation analysis between the accuracy and completeness scores for ChatGPT-5, Gemini 2.5 Flash, Grok-4, and DeepSeek-R1-0528.

	Completeness
Accuracy	
• r_s	0.73
• p-value	<0.001

r_s: Spearman's correlation coefficient

Discussion

This study aimed to evaluate the accuracy and completeness levels of responses to questions about chronic cracked and fractured teeth provided by ChatGPT-5, DeepSeek-R1-0528, Gemini 2.5 Flash, and Grok-4 AI models. According to the analysis results, the null hypothesis was rejected, and significant differences were observed in both the measures.

Yap et al. compared the opinions of general dentists and specialists in the management of cracked teeth and reported that general dentists, prosthodontists, and endodontists showed different approaches to the diagnosis, prognosis, and treatment of cracked teeth. General dentists stated that they referred patients with cracked teeth to specialists because of the difficulties experienced in diagnosis and treatment. It was concluded in that study that there is a need for evidence-based guidelines to be established for the treatment of cracked teeth, and because of the increasing interest in the literature in this field, it was emphasized that dentists should be encouraged to continuously follow new evidence (32). From this starting point, it was thought that general dentists could benefit from LLMs before referring cases to specialists when cracked and fractured teeth are detected. Accordingly, four LLMs were compared in this study to determine how accurately they could guide general dentists, particularly in terms of diagnosis and treatment.

AI is defined as software that can mimic human cognitive functions. By analyzing large data clusters, these systems aim to produce summarized information that can be used in clinical decision processes. In this context, LLMs have attracted increasing interest from researchers in the field of medicine. The rapid decision-making capacity offered by LLMs makes an important contribution. These advances in digital technology present considerable benefits for healthcare professionals, medical science, and patients (39).

In a study by Deborah et al., the completeness, comprehensibility, and accuracy of the responses given by ChatGPT to 243 questions covering different areas of dentistry were examined. As a result of the evaluations, the highest success rate was determined to be in open-ended questions, followed by paired questions, and the lowest performance was seen in multiple-choice questions (40). In parallel with these findings, open-study to be able to make more robust and reliable evaluations. The questions were asked to the LLMs only once, and when the model was insufficient, no additional explanation was provided, and no follow-up question was asked. This method aimed to compare the LLM responses under controlled conditions as much as possible.

In an article published in 1932, Likert defined a simple but strong method based on a group of related questions to measure the attitude of an individual on a certain subject (41). Self-reported Likert scales are an incredibly prominent data source in psychology and the social sciences in general (42). As LLMs are not inherently conscious beings, they cannot have direct bias in open-ended questions about dentistry. In addition, the use of Likert scales in the evaluation of the responses given can be helpful in revealing whether or not a prejudiced or biased understanding is reflected in the output produced by the models. This approach may be important for measuring the neutrality and consistency of model responses to questions involving clinical scenarios, treatment options or patient-physician communication.

Connective Health reported the high efficacy shown by Gemini 2.5 Flash in the processing of free text, including complex medical records. According to the organization's explanations, the institution's basic mission is to strengthen healthcare providers and improve the patient outcomes. In this context, patient trust is evaluated as a critical component. It was emphasized that AI-based solutions are developed in continuous collaboration with healthcare providers to be able to provide accuracy and efficacy. It was also stated that the rapid advances observed in the capabilities of Gemini continuously transformed the way in which clinical insights are presented, allowing the opportunity to discover new practices to improve the quality of life of both patients and healthcare providers (43). The developers of ChatGPT-5 have stated that this is the best model to date for health-related queries. Furthermore, at the beginning of this year, in HealthBench, an evaluation tool based on real scenarios and criteria defined by physicians, it was stated that ChatGPT-5 obtained significantly higher points than all previous models (13). Therefore, because of these strong claims in the field of healthcare, the Gemini 2.5 Flash and ChatGPT-5 versions were compared in the current study to evaluate the accuracy and completeness of responses given to questions related to cracked and fractured teeth.

According to the developers' statement, Grok 4 selects its own search questions when seeking real-time information or answering difficult research questions,

finds information on the web in general, and researches to the depth felt necessary to produce a high-quality response (21). It has been said by the producers of DeepSeek-R1-0528 that compared to previous versions, this upgraded model shows significant improvements in accomplishing complex reasoning tasks (44). Therefore, the Grok 4 and DeepSeek-R1-0528 versions, which are claimed to have deep reasoning competence and the ability to access up-to-date information, were evaluated together with two other LLMs in this study in the context of questions related to cracked and fractured teeth, which is a clinically complex subject requiring detailed reasoning. Furthermore, the most recent versions of the four models which were available at the time of the study were used for the evaluations.

Kaygısız et al. evaluated the responses to clinical case scenarios related to oral pathology provided by two different AI models: DeepSeek-V3 and ChatGPT-4o. The accuracy of the responses was measured using a 5-point Likert scale, and the mean points were determined to be 4.02±0.36 for DeepSeek-V3 and 3.15±0.41 for ChatGPT-4o. These results showed that both models displayed moderate to high performance. In the overall evaluation, it was concluded that DeepSeek-V3 was statistically more successful than ChatGPT-4o (45). Similarly, a 5-point Likert scale was used in this study to evaluate accuracy. The results showed that the accuracy of ChatGPT-5 was higher than that of DeepSeek-R1-0528. In the context of the subject being dealt with, this difference can be attributed to differences in question content or the different versions of the LLMs used.

In a study by Xu et al., a recovery-supported LLM framework for the development of endodontics education was developed and evaluated using measurements taken using two separate 5-point Likert scales for clinical accuracy and completeness of the information. The results showed that ChatGPT-4 (3.68±0.99) obtained higher mean points than DeepSeek-R1 (3.05±1.05) in respect of clinical accuracy and completeness of information (4.09±0.87 vs. 3.45±1.14) (46). Although more recent versions of these models were used in the current study- ChatGPT-5 and DeepSeek-R1-0528-the results obtained also showed better performance of ChatGPT than DeepSeek in terms of both accuracy and completeness.

In a study of clinical consensus and case analysis in dental implantology, four LLMs (ChatGPT-4, Qwen 2.0 72B, Claude 3 Opus, and Gemini Pro 1.5) were compared. In the responses to both simple and complex questions, a statistically significant difference was found between ChatGPT-4 and Gemini 1.5 Pro (47). In another study, Wu et al. compared and evaluated the performance of ChatGPT-o3-mini, DeepSeek-R1, Grok, Gemini 2.0-flash-thinking, and Qwen 2.5-max in clinical scenarios in the field of dental implantology. According to the results obtained, the Gemini 2.0-flash-thinking model was found to be superior to the other models in both responding to professional questions and in the context of case analyses. This superiority, observed especially in case analysis performance, suggests that the advanced information integration of the model is associated with strong clinical reasoning skills and effective decision-making capacity in complex clinical contexts (48). In the current study, no statistically significant difference was observed between ChatGPT-5 and Gemini 2.5 Flash in terms of accuracy and completeness of the responses to questions about cracked and fractured teeth.

Consistent with the statements of the developers (13, 43), both Gemini 2.5 Flash and ChatGPT-5 produced highly accurate and complete responses to questions related to cracked and fractured teeth, which is an important subject in the field of healthcare. However, despite these successful results, it must not be forgotten that LLMs will not replace healthcare specialists and should only be evaluated as a supportive tool in clinical decision-making processes.

In a study by Rivera et al., 12 cases involving different areas of cardiology were presented to GPT-4, Grok, and Gemini without giving any diagnosis and treatment information. The clinical reasoning of the LLMs was compared according to both areas and question types. Although the models showed superiority or inferiority to each other in some areas and questions, no statistically significant difference was determined between the three models in the overall analysis (49). In the current study, questions about diagnosis and treatment were not presented in the form of separate groups, and despite the differences in the subjects and changes in the versions of the LLMs used, no statistically significant difference was determined between ChatGPT-5, Gemini 2.5 Flash, and Grok 4.

The results of the current study revealed a significant positive relationship between the accuracy and completeness scores. This showed that an increase occurring in the accuracy score is associated with an increase in the completeness score. In other words, it can be understood that as the accuracy of the dataset increases so the completeness is strengthened. This highlights that the accuracy and completeness dimensions support each other and the importance of assessing these two markers together to evaluate the system performance holistically. Moreover, it can be suggested that improvements directed at increasing accuracy would be positively reflected not only in that dimension but also in completeness.

The results of the current study showed that the mean scores for the questions about conditions creating difficulties in the diagnosis and treatment of cracked and fractured teeth revealed the advanced-level professional skills of new-generation LLMs and shed light on the evolving development process of these models in medicine. However, there is a need for a thorough and comprehensive confirmatory process to evaluate the performance of LLMs with different question contents related to the same subject.

Hallucinations (unsupported information production) and misinformation can give rise to serious risks in clinical decision-making processes. This problem can originate from the models having been trained on data from articles, books, Wikipedia, news content, blogs, forum discussions, social media sharing, and various online sources, rather than the data obtained

directly from scientific associations. In the current study, although correct and consistent responses were produced in the reference sources, the hallucination problem in LLMs was seen to continue. Therefore, there is a need for further comprehensive research to understand the origin of this tendency and to develop more reliable systems.

Conclusion

In contemporary orthodontic practice, the health and stability of the temporomandibular joint should not be treated as an isolated concern but as an integral component of the treatment planning. The review of occlusal theories, particularly cuspid rise and group function, underlines the importance of force distribution during mandibular movements to prevent unnecessary strain on the TMJ. Moreover, a clear distinction between static and functional occlusion provides a more accurate understanding of how dental alignment affects joint mechanics during rest and function.

Orthodontic goals have evolved from merely achieving esthetic alignment to ensuring functional harmony and long-term stability. Maintaining musculoskeletal equilibrium and preventing relapse through carefully applied forces, post-treatment retention, and individualized treatment protocols is essential. Special attention must be given to diagnosing and managing TMJ disorders before, during, and after orthodontic intervention.

Additionally, recognizing the signs of TMD, differentiating sources of facial pain, and understanding the correlation between malocclusions and joint pathology are critical for accurate diagnosis and effective treatment. Malocclusions, such as Class II, Class III, deep bites, and crossbites, each impose unique mechanical challenges that can compromise TMJ function.

Ultimately, the success of orthodontic therapy is not only measured by dental alignment but also by the functional well-being of the TMJ. A holistic, evidence-based approach incorporating occlusal theory, anatomical understanding, and patient-specific diagnostics will ensure improved outcomes and long-term joint health.

Disclosures

Peer-review: Externally peer-reviewed.

Author Contributions: Concept – M.T.; Design – M.T., Ö.A.; Supervision – Ö.A., S.C.; Materials – M.T.; Data collection &/or processing – S.C.; Analysis and/or interpretation – H.O.; Literature search – M.T.; Writing – M.T.; Critical review – M.T.

Conflict of Interest: There is no any conflict of interest.

Funding: There is no any funding.

References

1. Brandini DA, Sonoda CK, Takamiya AS, Monteiro DR. Use of ChatGPT in dental traumatology. Aust Endod J. 2024;50(2). https://doi.org/10.1111/aej.12845
2. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023;29(8):1930-1940. https://doi.org/10.1038/s41591-023-02448-8
3. Meng X, Yan X, Zhang K, Liu D, Cui X, Yang Y, et al. The application of large language models in medicine: a scoping review. iScience. 2024;27(5):109713. https://doi.org/10.1016/j.isci.2024.109713
4. Nietsch KS, Shrestha N, Ndjonko LCM, Ahmed W, Mejia MR, Zaidat B, et al. Can large language models (LLMs) predict the appropriate treatment of acute hip fractures in older adults? Comparing appropriate use criteria with recommendations from ChatGPT. JAAOS Glob Res Rev. 2024;8(8):e24.00123. https://doi.org/10.5435/JAAOSGlobal-D-24-00123
5. Nielsen JP, von Buchwald C, Grønhøj C. Validity of the large language model ChatGPT (GPT4) as a patient information source in otolaryngology by a variety of doctors in a tertiary otorhinolaryngology department. Acta Otolaryngol. 2023:779-782. https://doi.org/10.1080/00016489.2023.2272078
6. Wahlster W. Understanding computational dialogue understanding. Philos Trans A Math Phys Eng Sci. 2023;381(2251):20220049. https://doi.org/10.1098/rsta.2022.0049
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. Int Dent Res. 2025;15(3) (Advanced Online). https://doi.org/10.5577/intdentres.662
8. Domingos P. The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World. New York: Basic Books; 2015.
9. Lechien JR, Georgescu BM, Hans S, Chiesa-Estomba CM. ChatGPT performance in laryngology and head and neck surgery: a clinical case-series. Eur Arch Otorhinolaryngol. 2024;281(1):319-333. https://doi.org/10.1007/s00405-023-08248-6
10. Vaira LA, Lechien JR, Abbate V, Allevi F, Audino G, Beltramini GA, et al. Accuracy of ChatGPT-generated information on head and neck and oromaxillofacial surgery: a multicenter collaborative analysis. Otolaryngol Head Neck Surg. 2024;170(6):1492-1503. https://doi.org/10.1002/ohn.664
11. Mago J, Sharma M. The potential usefulness of ChatGPT in oral and maxillofacial radiology. Cureus. 2023;15(7):e42133. https://doi.org/10.7759/cureus.42133
12. Ali R, Tang OY, Connolly ID, Sullivan PLZ, Shin JH, Fridley JS, et al. Performance of ChatGPT and GPT-4 on neurosurgery written board examinations. Neurosurgery. 2023;93(6):1353-1365. https://doi.org/10.1227/neu.0000000000002636
13. OpenAI. Introducing GPT-5. 2025.

https://openai.com/tr-TR/index/introducing-gpt-5/
14. Gibney E. China's cheap, open AI model DeepSeek thrills scientists. Nature. 2025;638(8049):13-14. https://doi.org/10.1038/d41586-025-00234-7
15. DeepSeek. DeepSeek-R1-0528 Release. 2025. https://api-docs.deepseek.com/news/news250528
16. Masalkhi M, Ong J, Waisberg E, Lee AG. Google DeepMind's Gemini AI versus ChatGPT: a comparative analysis in ophthalmology. Eye (Lond). 2024;38(8):1412-1417. https://doi.org/10.1038/s41433-024-02958-w
17. Abbas YN, Mahmood YM, Hassan HA, Hamad DQ, Hasan SJ, Omer DA, et al. Role of ChatGPT and Google Bard in the diagnosis of psychiatric disorders: a comparative study. Barw Med J. 2023.
18. Google. Gemini momentum continues with launch of 2.5 Flash-Lite and general availability of 2.5 Flash and Pro on Vertex AI. 2025. https://cloud.google.com/blog/products/ai-machine-learning/gemini-2-5-flash-lite-flash-pro-ga-vertex-ai
19. Wangsa K, Karim S, Gide E, Elkhodr M. A systematic review and comprehensive analysis of pioneering AI chatbot models from education to healthcare: ChatGPT, Bard, Llama, Ernie and Grok. Future Internet. 2024;16(7):219. https://doi.org/10.3390/fi16070219
20. Sozer A, Sahin MC, Sozer B, Erol G, Tufek OY, Nernekli K, et al. Do LLMs have 'the eye' for MRI? Evaluating GPT-4o, Grok, and Gemini on brain MRI performance: first evaluation of Grok in medical imaging and a comparative analysis. Diagnostics (Basel). 2025;15(11):1320. https://doi.org/10.3390/diagnostics15111320
21. xAI. Grok 4. 2025. https://x.ai/news/grok-4
22. Ahmed WM, Azhari AA, Alfaraj A, Alhamadani A, Zhang M, Lu C-T. The quality of AI-generated dental caries multiple choice questions: a comparative analysis of ChatGPT and Google Bard language models. Heliyon. 2024;10(7):e28546. https://doi.org/10.1016/j.heliyon.2024.e28546
23. Kim H-S, Kim G-T. Can a large language model create acceptable dental board-style examination questions? A cross-sectional prospective study. J Dent Sci. 2025;20(2):895-900. https://doi.org/10.1016/j.jds.2024.11.012
24. Chau RCW, Thu KM, Yu OY, Hsung RT-C, Lo ECM, Lam WYH. Performance of generative artificial intelligence in dental licensing examinations. Int Dent J. 2024;74(3):616-621. https://doi.org/10.1016/j.identj.2023.12.007
25. Suárez A, Jiménez J, de Pedro ML, Andreu-Vázquez C, García VD-F, Sánchez MG, et al. Beyond the scalpel: assessing ChatGPT's potential as an auxiliary intelligent virtual assistant in oral surgery. Comput Struct Biotechnol J. 2024;24:46-52. https://doi.org/10.1016/j.csbj.2023.12.008
26. Batool I, Naved N, Kazmi SMR, Umer F. Leveraging large language models in the delivery of post-operative dental care: a comparison between an embedded GPT model and ChatGPT. BDJ Open. 2024;10(1):48. https://doi.org/10.1038/s41405-024-00232-5
27. American Association of Endodontists. Glossary of Endodontic Terms. Chicago: American Association of Endodontists; 2003.
28. Kim S-Y, Kim S-H, Cho S-B, Lee G-O, Yang S-E. Different treatment protocols for different pulpal and periapical diagnoses of 72 cracked teeth. J Endod. 2013;39(4):449-452. https://doi.org/10.1016/j.joen.2012.12.016
29. Kang SH, Kim BS, Kim Y. Cracked teeth: distribution, characteristics, and survival after root canal treatment. J Endod. 2016;42(4):557-562. https://doi.org/10.1016/j.joen.2016.01.014
30. Krell KV, Caplan DJ. 12-month success of cracked teeth treated with orthograde root canal treatment. J Endod. 2018;44(4):543-548. https://doi.org/10.1016/j.joen.2017.12.025
31. Kahler W. The cracked tooth conundrum: terminology, classification, diagnosis, and management. Am J Dent. 2008;21(5):275-282.
32. Yap EXY, Chan PY, Yu VSH, Lui J-N. Management of cracked teeth: perspectives of general dental practitioners and specialists. J Dent. 2021;113: 103770. https://doi.org/10.1016/j.jdent.2021.103770
33. Ellis S. Incomplete tooth fracture—proposal for a new definition. Br Dent J. 2001;190(8):424-428. https://doi.org/10.1038/sj.bdj.4800997
34. Li X, Tang M. The clinical diagnosis and treatment of 78 cracked teeth with pulpitis. Chongqing Med. 2015:3070-3071, 3075.
35. Patel S, Bhuva B, Bose R. Present status and future directions: vertical root fractures in root filled teeth. Int Endod J. 2022;55(Suppl 3):804-826. https://doi.org/10.1111/iej.13737
36. Hargreaves KM. Cohen's Pathways of the Pulp: South Asia Edition E-book. Elsevier Health Sciences; 2020.
37. De Vito A, Colpani A, Moi G, Babudieri S, Calcagno A, Calvino V, et al. Assessing ChatGPT's potential in HIV prevention communication: a comprehensive evaluation of accuracy, completeness, and inclusivity. AIDS Behav. 2024;28(8):2746-2754. https://doi.org/10.1007/s10461-024-04358-9
38. Marshall RF, Mallem K, Xu H, Thorne J, Burkholder B, Chaon B, et al. Investigating the accuracy and completeness of an artificial intelligence large language model about uveitis: an evaluation of ChatGPT. Ocul Immunol Inflamm. 2024;32(9):2052-2055. https://doi.org/10.1080/09273948.2023.2287755
39. Cuocolo R, Caruso M, Perillo T, Ugga L, Petretta M. Machine learning in oncology: a clinical appraisal. Cancer Lett. 2020;481:55-62. https://doi.org/10.1016/j.canlet.2020.03.009
40. Sybil D, Shrivastava P, Rai A, Injety R, Singh S, Jain A, et al. Performance of ChatGPT in dentistry: multi-specialty and multi-centric study. 2023.
41. Batterton KA, Hale KN. The Likert scale what it is and how to use it. Phalanx. 2017;50(2):32-39.
42. Jebb AT, Ng V, Tay L. A review of key Likert scale development advances: 1995-2019. Front Psychol. 2021;12:637547. https://doi.org/10.3389/fpsyg.2021.637547
43. Google. Gemini momentum continues with launch of 2.5 Flash-Lite and general availability of 2.5 Flash and Pro on Vertex AI. 2025. https://cloud.google.com/blog/products/ai-machine-learning/gemini-2-5-flash-lite-flash-pro-ga-vertex-ai
44. DeepSeek. DeepSeek-R1-0528 Release. 2025. https://api-docs.deepseek.com/news/news250528
45. Kaygisiz ÖF, Teke MT. Can DeepSeek and ChatGPT be used in the diagnosis of oral pathologies? BMC Oral Health. 2025;25(1):638. https://doi.org/10.1186/s12903-025-05892-3
46. Xu X, Liu S, Zhu L, Long Y, Zeng Y, Lu X, et al. Development and evaluation of a retrieval-augmented large language model framework for enhancing endodontic education. Int J Med Inform. 2025:106006. https://doi.org/10.1016/j.ijmedinf.2025.106006
47. 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. https://doi.org/10.1016/j.prosdent.2024.12.014
48. Wu X, Cai G, Guo B, Ma L, Shao S, Yu J, et al. A multi-dimensional performance evaluation of large language models in dental implantology: comparison of ChatGPT, DeepSeek, Grok, Gemini and Qwen across diverse clinical scenarios. BMC Oral Health. 2025;25(1):1272. https://doi.org/10.1186/s12903-025-06596-9
49. Reyes-Rivera J, Castro Molina A, Romero-Lorenzo M, Ali S, Gibson C, Saucedo J, et al. Evaluating the Clinical Reasoning of GPT-4, Grok, and Gemini in Different Fields of Cardiology. Circulation. 2024;150(Suppl_1):A4147550-A.