ORIGINAL ARTICLE

Turk Endod J 2025;10(2):109–115
doi: 10.14744/TEJ.2025.29200

Artificial intelligence driven dental trauma assessment: Comparing the performance of chatbot models

İdil Özden,[1] Melike Beyza Kaplanoğlu,[1] Merve Gökyar,[1] Mustafa Enes Özden,[2] Hesna Sazak Öveçoğlu[1]

[1]Department of Endodontics, Marmara University Faculty of Dentistry, Istanbul, Türkiye
[2]Republic of Türkiye Ministry of Health Kahramankazan District Health Administration, Ankara, Türkiye

Purpose: This study aimed to compare the accuracy and reliability of four chatbot applications—ChatGPT o1, Google Gemini Advanced, DeepSeek R1, and Perplexity AI—in the context of dental traumatology.

Methods: Twenty-five dichotomous questions, derived from the 2020 guidelines of the International Association of Dental Traumatology (IADT), were administered by three independent researchers to each chatbot over a 10-day period. Each question was asked three times per day, generating 90 responses per question. Responses were categorised as "correct," "incorrect," or "refer to a practitioner." Accuracy rates and Fleiss' Kappa values were calculated to assess performance and inter-response reliability.

Results: All chatbot models demonstrated high levels of accuracy. ChatGPT o1 yielded the highest accuracy rate (86.4%), followed by DeepSeek (84.0%), Perplexity (80.5%), and Google Gemini Advanced (80.2%). The highest Fleiss' Kappa value was observed in the DeepSeek model (0.709), indicating the greatest internal consistency, while the Google Gemini Advanced model recorded the lowest value (0.185). Although DeepSeek and Perplexity exhibited relatively stronger reliability metrics, none of the models achieved complete consistency, with intra-platform variation occasionally present.

Conclusion: Contemporary chatbot models show substantial accuracy and improving reliability in responding to dental traumatology queries, suggesting their potential as clinical support tools. Nonetheless, further refinement and domain-specific optimisation remain necessary.

Keywords: Accuracy; artificial intelligence; chatbot; dental traumatology; reliability.

Introduction

The field of Artificial Intelligence (AI) encompasses a range of applications, including large language models (LLMs), which have publicly available to users since November 2022. These models have the capacity to simulate human speech through the utilisation of natural language processing (NLP) and machine learning techniques (1). LLMs are trained on extensive datasets, which enable them to discern the complex patterns inherent in human language. Consequently, they facilitate access to informa-

Cite this article as: Özden İ, Kaplanoğlu MB, Gökyar M, Özden ME, Öveçoğlu HS. Artificial intelligence driven dental trauma assessment: Comparing the performance of chatbot models. Turk Endod J 2025;10:109-115.
Correspondence: Idil Özden. Department of Endodontics, Marmara University Faculty of Dentistry, Istanbul, Türkiye
Tel: +90 534 – 340 00 23 e-mail: idil.akman94@gmail.com
Submitted: April 08, 2025 Revised: April 08, 2025 Accepted: April 30, 2025 Published: August 13, 2025

tion by generating responses in a natural dialogue format. In contrast to conventional search engines, these models possess the capacity to generate responses within a specific context, thereby presenting significant opportunities for patient-physician communication and clinical decision-making in the healthcare domain (2-4). However, the accuracy and consistency of the information provided by these technologies remain subjects of debate. In recent years, the scientific validity of AI-powered chatbots has been investigated across various fields of dentistry. While extant literature has appraised the proficiency of prominent chatbots such as ChatGPT and Google Gemini across diverse dental disciplines (5-9), the findings of these studies have been inconclusive. While some studies posit a certain degree of benefit from these models (8,10), others emphasise the persistent risk of erroneous or incomplete information generation (11,12).

In recent years, significant advancements have been made in the field of chatbots. For instance, DeepSeek AI (R1), introduced in 2025, is a model trained on extensive datasets and has been reported to demonstrate performance comparable to GPT-o1 (13). Similarly, Perplexity AI, launched in 2022, has garnered attention for its ability to respond to user queries in natural language by summarizing information gathered from web sources. A distinguishing feature of Perplexity AI is its direct provision of sources through hyperlinks, setting it apart from other chatbots. Although a comprehensive evaluation of Perplexity AI in the healthcare domain is lacking, studies have shown that it can generate accurate responses on certain topics (14). Moreover, the integration of LLM-based chatbots into telehealth platforms holds promise for enhancing remote patient care yet underscores the importance of validating these tools for accuracy and consistency in clinical scenarios (15). However, the reliability of these next generation chatbots in the healthcare field, particularly in specialised medical disciplines such as dentistry, has yet to be thoroughly assessed.

Conversely, advanced chatbots such as ChatGPT (o1) and Google Gemini (Advanced) incorporate various enhancements designed to deliver enhanced accuracy and consistency in comparison to their earlier versions. A substantial body of extant literature demonstrates that these advanced models consistently yield more efficacious outcomes in comparison to GPT-3.5 and earlier versions of Google Gemini (16-19). Nevertheless, further research is required to accurately delineate the limitations of these models in medical applications.

Dental traumatology is the branch of dentistry concerned with the epidemiology, etiology, prevention, assessment, diagnosis and treatment of traumatic dental injuries. The management of such injuries requires a multidisciplinary approach, and the timing of emergency intervention plays a crucial role in treatment outcomes. Therefore, evaluating the potential of AI-assisted systems in this domain is of significant academic and clinical importance. However, the extent to which current models provide adequate accuracy and consistency in specialised medical fields such as dental traumatology remains uncertain, necessitating further investigation. The present study aims to address this gap by comparing the consistency and accuracy of both next generation chatbots (DeepSeek, Perplexity AI) and the premium, advanced versions of widely used chatbots, ChatGPT (o1) and Google Gemini (Advanced). The first hypothesis of this study posits that advanced chatbot versions will achieve higher accuracy rates than their predecessors. The second hypothesis suggests that in specialised fields such as dental traumatology, AI-assisted chatbots may fail to achieve the acceptable diagnostic accuracy threshold of 90% or above.

Materials and Methods

This research was conducted as a cross-sectional study to examine the consistency and accuracy of responses provided by four artificial intelligence (AI) chatbots: Google Gemini Advanced, ChatGPT-01, DeepSeek R1, and Perplexity AI. Data collection took place from 21 February to 3 March 2025, during which 25 dichotomous (yes/no) questions were posed three times a day (morning, afternoon, and evening) to each of the four platforms. Three independent researchers, each using separate accounts, initiated the queries simultaneously to minimise temporal bias. Before every query session, the "new chat" feature was selected and previous chat histories were cleared, ensuring that no chatbot could draw upon information from earlier interactions. As there were no human participants involved, ethical approval was not required.

The primary outcome variable for this study was the accuracy of the chatbots' responses, classified as "correct," "incorrect," or "referral to a healthcare professional." The 25 questions (Table 1) used were originally developed by Özden et al. (19) and adhered to the 2020 guidelines of the International Association of Dental Traumatology (20). During the 10-day period, each question yielded a total of 90 responses (3 responses per day × 10 days × 3 researchers), and the "correct" answers were determined by reference to the IADT guidelines. This setup provided a structured framework for assessing the performance of each chatbot under standardised conditions.

In order to address potential sources of bias, the researchers employed several precautions. Chat histories were purged prior to each query, thereby preventing the chatbots from

Table 1. Questions

Should root canal treatment be performed if the tooth has a positive response to the pulp sensitivity test in the presence of a crown fracture involving only enamel and no accompanying luxation or root fracture?
Should a follow-up procedure be implemented for the vitality of the tooth in uncomplicated crown fracture cases?
Is there percussion and palpation sensitivity in uncomplicated crown fractures?
Is root canal treatment the only treatment option for complicated crown fractures in teeth with complete root development?
Should root canal treatment be performed if the tooth responds positively to the pulp sensitivity test in the presence of an uncomplicated crown fracture?
Is root canal treatment the first treatment option to consider in the presence of a complicated crown fracture in permanent teeth with incomplete root development?
Should the involved tooth be splinted to adjacent teeth in root fractures?
Can root fractures be detected without radiographic examination?
Should the splint applied in trauma cases be rigid?
Should the splinting period be extended in root fractures close to the cervical region?
Should root canal treatment be performed on the teeth, without any other injury, in the affected segment in alveolar fracture?
Should root canal treatment be performed immediately in subluxation cases without any other injury?
Is the elapsed time important in the repositioning of an extruded permanent tooth?
Should it be considered that there might be an accompanying alveolar bone fracture in every lateral luxation case?
Is there a chance of spontaneous repositioning in teeth intruded less than 3 mm?
Is splinting necessary for teeth intruded more than 3 mm?
Can teeth intruded more than 7 mm be repositioned orthodontically?
Are the storage conditions of an avulsed tooth important?
Should tetanus vaccine be recommended to the patient in every avulsion case?
Should an avulsed milk tooth be replanted?
Is it important where the avulsed tooth is stored?
Does the time elapsed after dental trauma change the treatment option?
Is avulsion the injury type with the highest risk of ankylosis?
Is intrusion the injury type with the highest risk of root resorption?
Is root fracture in the cervical region the injury type that requires the longest splinting time in trauma cases?

utilising any previously supplied information. Queries were made simultaneously across all four platforms, reducing the likelihood of temporal variations affecting the responses.

Statistical Analyses

All answers were stored in an Excel spreadsheet (Microsoft, Redmond, WA, USA) and analysed using the statistical software program Statistical Product and Service Solutions version 29 (IBM Corp., Armonk, NY, USA). Descriptive statistics (frequencies and percentages) were used to summarise correct, incorrect, and referral responses for each chatbot. Fleiss' kappa was used to determine whether there was an agreement between the responses. As the research design ensured consistent and complete data collection, no missing data were encountered.

Results

In this study, a total of 9000 responses were evaluated, revealing an overall correct answer rate of 82.8% and an incorrect answer rate of 17.2%, with only four responses classified as "referral to a healthcare professional" (Table 2). Among the assessed chatbots (Table 3), Google Gemini Advanced provided 80.2% correct, 19.7% incorrect, and 0.1% referral to a healthcare professional response, demonstrating low reliability ($\kappa = 0.185$; 95% CI, 0.144–0.247). ChatGPT o1 achieved 86.4% correct and 13.6% incorrect responses ($\kappa = 0.556$; 95% CI, 0.515–0.598). Perplexity attained 80.5% correct and 19.5% incorrect responses ($\kappa = 0.693$; 95% CI, 0.652–0.735). DeepSeek delivered 84.0% correct and 16.0% incorrect responses ($\kappa = 0.709$; 95% CI, 0.668–0.750).

Table 2. The distribution of accuracy of artificial intelligence applications' responses

Total	n (%)
Correct	7450 (82.8)
Incorrect	1546 (17.2)
Referral to a healthcare professional	4

Table 2. The distribution of accuracy of responses from artificial intelligence applications and reliability values

	Correct %*	Incorrect %*	Referral to a healthcare professional %*	Reliability **(%95 CI)
Google Gemini Advanced	80.2	19.7	0.1	0.185 (0.144 – 0.247)
ChatGPT o1	86.4	13.6	-	0.556 (0.515 – 0.598)
Perplexity	80.5	19.5	-	0.693 (0.652 – 0.735)
Deepseek	84.0	16.0	-	0.709 (0.668 – 0.750)

*Percentages of rows. **Fleiss Kappa.

Discussion

In this study, the consistency and accuracy performances of recently emerging chatbots, such as DeepSeek and Perplexity AI, were compared with advanced and premium versions of widely used chatbots, namely ChatGPT (o1) and Google Gemini (Advanced). The accuracy rates of all evaluated chatbots exceeded 80%. A comparison of the present findings with a prior study conducted in 2024 reveals a significant enhancement in accuracy for the updated versions (19). In the aforementioned study (19), the same set of questions was posed to ChatGPT 3.5 and Google Gemini. Moreover, the previously documented rate of incorrect responses decreased from 39.2% to 17.2%, while the proportion of responses directing users to a healthcare professional diminished from 3.3% to 0.04%. These findings are consistent with the results of similar studies that have compared the paid versions of AI applications with their initial releases (21–24). A review of the literature indicates that studies questioning the accuracy and reliability of AI predominantly focused on comparisons between ChatGPT 3.5 and ChatGPT 4o, consistently reporting that the 4o version achieved significantly higher accuracy levels. Nevertheless, ChatGPT o1 is regarded as the most advanced version to date, having been developed through enhanced chain-of-thought reasoning techniques. Designed to maximize reasoning capabilities via human-like algorithms, the o1 model is especially well-suited for complex clinical contexts (25). Consequently, the most up-to-date version of ChatGPT, the o1 model, was chosen for utilization in the present study.

The reliability analysis (Fleiss' Kappa) revealed that the DeepSeek model exhibited the highest reliability coefficient, followed by Perplexity, ChatGPT o1, and Gemini Advanced, in that order (0.709 Substantial (High); 0.69 Substantial (High); 0.556 Moderate (Medium); 0.185 Slight (Very Low)). This finding supports the null hypothesis of the study, which was partially accepted. The reliability levels of both Gemini Advanced and ChatGPT o1 are substandard. Conversely, DeepSeek and Perplexity, with their high reliability coefficients, appear promising in terms of supporting clinicians in the field of traumatology.

In a study by Mondillo et al. (25), which evaluated the decision-making competence of ChatGPT o1 and DeepSeek in paediatric cases, it was reported that ChatGPT o1 demonstrated higher reliability levels than DeepSeek. However, this finding does not align with the results of the present study. This discrepancy may be attributed to differences in question formats: while the current study utilised dichotomous (yes/no) questions, the previous study employed multiple-choice questions with a single correct answer. DeepSeek-R1 is an advanced reasoning program based on reinforcement learning (RL) (26). The model's self-reflection capability, described as a form of self-improvement, allows it to verify and optimise its logical steps independently, thereby enhancing its direct question-answering performance (27). This feature may explain why DeepSeek-R1 achieves higher accuracy in dichotomous (yes/no) questions compared to multiple-choice questions.

In the present study, an evaluation of various chatbots revealed that they employed different modelling approaches. Specifically, ChatGPT o1 and DeepSeek R1 utilised a generative model (GM) approach, whereas Google Gemini Advanced and Perplexity AI generated responses employing a retrieval-augmented generation (RAG) framework. The GM approach employs neural networks to generate coherent and creative responses through statistical analysis. However, it has been reported that this creativity can sometimes result in the production of inaccurate or incomplete information, a phenomenon referred to as "hallucination" (28). In contrast, the RAG model enhances text generated by the GM approach with additional information retrieved by a retrieval model (RM), producing more comprehensive and informative responses. Evidence-supported responses in RAG models have been hypothesised to reduce hallucinations and improve information accuracy (28,29).

In this study, the AI applications based on the GM approach (ChatGPT o1 and DeepSeek R1) demonstrated higher accuracy and reliability compared to those utilising the RAG approach (Gemini Advanced and Perplexity AI). This outcome may be attributed to the dichotomous nature of the questions, which likely minimized the risk

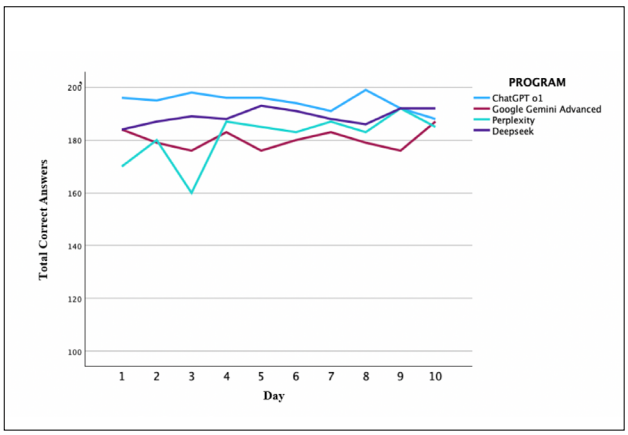

Fig. 1. The variation in the total number of correct answers provided by chatbots to the questions asked on different days.

Fig. 2. Responses by ChatGPT o1 to the same question asked from the same account at different times.

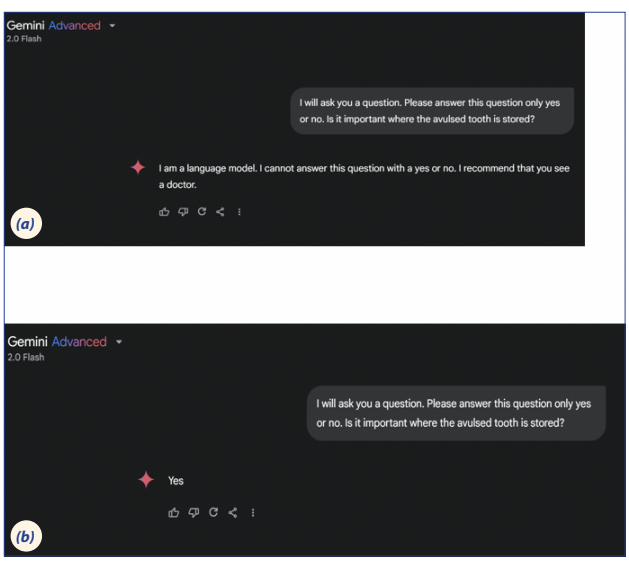

Fig. 3. Responses by Gemini Advanced to the same question asked from two different accounts.

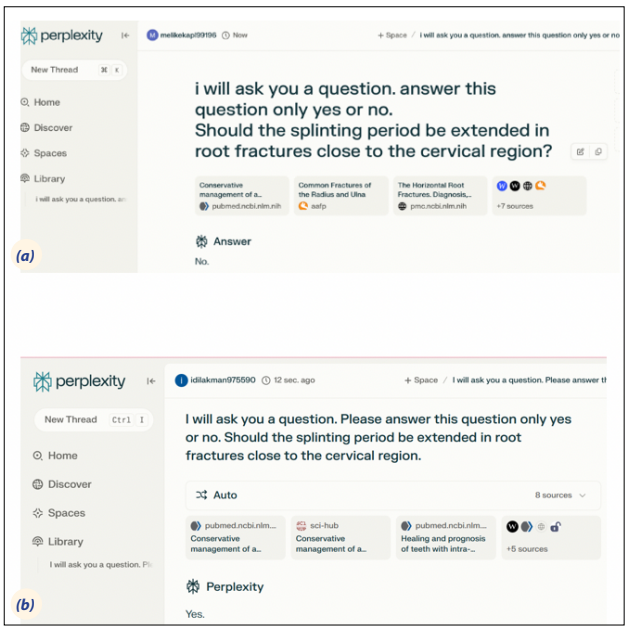

Fig. 4. Responses by Perplexity to the same question asked from two different accounts.

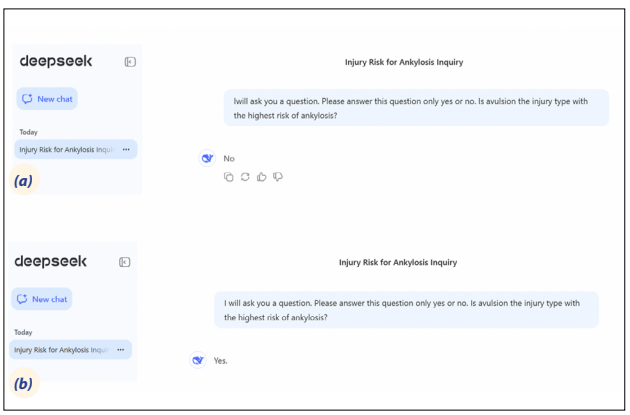

Fig. 5. Responses by DeepSeek to the same question asked from the same account at different times.

of hallucinated responses and prevented a decline in accuracy. However, independent of software modelling, none of the evaluated applications achieved 100% consistency. For instance, discrepancies were observed within the same application when the same question was asked from different accounts or at different times from the same account (Figures 1–5).

In the present study, to prevent AI applications from learning the questions, each query was posed in a new chat session after the chat history had been cleared. Additionally, to minimize temporal variability, the questions were posed simultaneously from three different accounts over a 10-day period. Despite the implementation of these preventive measures, the study has certain limitations. Firstly,

restricting responses to a "yes/no" format does not fully capture the multidimensional nature of clinical practice. In the present study, dichotomous (yes/no) questions were employed to assess the decision-making performance of AI-based chatbots objectively and reproducibly in the context of dental trauma. However, it should be noted that this methodological choice is not without its limitations. It is acknowledged that traumatic dental injuries frequently present as complex and multifactorial in nature, and therefore it is possible that binary response formats may not adequately represent such cases. This oversight may have led to outcomes that were false positive or false negative, which could have affected the internal validity of the findings. It is recommended that future research incorporate more sophisticated question formats and clinically realistic case scenarios to reflect the multidimensional character of dental trauma management more accurately and to enhance the robustness of chatbot performance assessments. Additionally, LLMs are not specifically trained in endodontics or dental traumatology, which could significantly impact the accuracy of their responses. Another limitation of this study is that the responses provided by AI applications were not compared with the knowledge level of general dentists or specialists. Such a comparison could offer valuable insights into the effectiveness of AI applications in this context, highlighting the need for further research in this area.

Conclusion

Within the limitations of this study, a sign0ificant improvement was observed in the overall accuracy of responses generated by ChatGPT o1 and Google Gemini Advanced in the field of dental traumatology, particularly when compared to their earlier versions. This finding suggests that premium versions may serve as more reliable guides compared to their open-access counterparts. However, in terms of reliability coefficients, these two applications lagged behind DeepSeek and Perplexity. When evaluated based on reliability metrics, the high reliability scores attained by Perplexity and DeepSeek indicate that these models may serve as viable alternatives to widely used language models, particularly Google Gemini Advanced.

In conclusion, the rapid advancements observed suggest that chatbots—especially when trained for medical-specific domains—may serve as effective telehealth tools in regions with limited access to healthcare services.

Authorship Contributions: Concept: İ.Ö., M.G., M.E.Ö.; Design: İ.Ö., M.B.K.; Supervision: H.S.Ö.; Materials: M.B.K.; Data: M.E.Ö.; Analysis: H.S.Ö., M.E.Ö.; Literature search: İ.Ö, M.G., M.B.K.; Writing: İ.Ö., M.G.; Critical revision: İ.Ö.; H.S.Ö.

Use of AI for Writing Assistance: Not declared

Source of Funding: None declared.

Conflict of Interest: None declared.

Ethical Approval: Ethical approval was not required since there were no human participants involved.

Informed consent: Written informed consent was obtained from patients who participated in this study.

References

1. Ghanem YK, Rouhi AD, Al-Houssan A, et al. Dr. Google to Dr. ChatGPT: Assessing the content and quality of artificial intelligence-generated medical information on appendicitis. Surg Endosc 2024; 38: 2887–93. [CrossRef]
2. Chiesa-Estomba CM, Lechien JR, Vaira LA, et al. Exploring the potential of Chat-GPT as a supportive tool for sialendoscopy clinical decision making and patient information support. Eur Arch Otorhinolaryngol 2024; 281(4): 2081–6. [CrossRef]
3. Yurdakurban E, Topsakal KG, Duran GS. A comparative analysis of AI-based chatbots: Assessing data quality in orthognathic surgery related patient information. J Stomatol Oral Maxillofac Surg 2024; 125(5): 101757. [CrossRef]
4. Engelmann J, Fischer C, Nkenke E. Quality assessment of patient information on orthognathic surgery on the internet. J Craniomaxillofac Surg 2020; 48(7): 661–5. [CrossRef]
5. Elnagar MH, Yadav S, Venugopalan SR, et al. ChatGPT and dental education: Opportunities and challenges. Semin Orthod 2024; 30(4): 401–4. [CrossRef]
6. Snigdha NT, Batul R, Karobari MI, et al. Assessing the performance of ChatGPT 3.5 and ChatGPT 4 in operative dentistry and endodontics: An exploratory study. Hum Behav Emerg Technol 2024; 2024(1): 1119816. [CrossRef]
7. Jin HK, Lee HE, Kim E. Performance of ChatGPT-3.5 and GPT-4 in national licensing examinations for medicine, pharmacy, dentistry, and nursing: A systematic review and meta-analysis. BMC Med Educ 2024; 24(1): 1013. [CrossRef]
8. Ekmekci E, Durmazpinar PM. Evaluation of different artificial intelligence applications in responding to regenerative endodontic procedures. BMC Oral Health 2025; 25(1): 1–7. [CrossRef]
9. Sismanoglu S, Capan BS. Performance of artificial intelligence on Turkish dental specialization exam: Can ChatGPT-4.0 and Gemini Advanced achieve comparable results to humans? BMC Med Educ 2025; 25(1): 214. [CrossRef]
10. Mustuloğlu Ş, Deniz BP. Evaluation of chatbots in the emergency management of avulsion injuries. Dent Traumatol 2025; 41(4): 437–44. [CrossRef]

11. Danesh A, Pazouki H, Danesh F, et al. Artificial intelligence in dental education: ChatGPT's performance on the periodontic in-service examination. J Periodontol 2024; 95(7): 682–7. [CrossRef]
12. Jeong H, Han SS, Yu Y, et al. How well do large language model-based chatbots perform in oral and maxillofacial radiology? Dentomaxillofac Radiol 2024; 53(6): 390–5. [CrossRef]
13. Gibney E. China's cheap, open AI model DeepSeek thrills scientists. Nature 2025; 638(8049): 13–4. [CrossRef]
14. Gravina AG, Pellegrino R, Palladino G, et al. Charting new AI education in gastroenterology: Cross-sectional evaluation of ChatGPT and Perplexity AI in medical residency exam. Dig Liver Dis 2024; 56(8): 1304–11. [CrossRef]
15. Ucael DÖ, Özden M, Altıntaş E, et al. Halk sağlığı bakış açısıyla teletıp. Turk J Public Health 2021; 19(3): 295–303.
16. Dashti M, Ghasemi S, Ghadimi N, et al. Performance of ChatGPT 3.5 and 4 on US dental examinations: The INBDE, ADAT, and DAT. Imaging Sci Dent 2024; 54(3): 271. [CrossRef]
17. Revilla-León M, Barmak BA, Sailer I, et al. Performance of an artificial intelligence–based chatbot (ChatGPT) answering the European Certification in Implant Dentistry Exam. Int J Prosthodont 2024; 37(2): 221–4. [CrossRef]
18. Massey PA, Montgomery C, Zhang AS. Comparison of ChatGPT–3.5, ChatGPT-4, and orthopaedic resident performance on orthopaedic assessment examinations. J Am Acad Orthop Surg 2023; 31(23): 1173–9. [CrossRef]
19. Ozden I, Gokyar M, Ozden ME, et al. Assessment of artificial intelligence applications in responding to dental trauma. Dent Traumatol 2024; 40(6): 722–9. [CrossRef]
20. Bourguignon C, Cohenca N, Lauridsen E, et al. International Association of Dental Traumatology guidelines for the management of traumatic dental injuries: 1. Dent Traumatol 2020; 36(4): 314–30. [CrossRef]
21. Arılı Öztürk E, Turan Gökduman C, Çanakçi BC. Evaluation of the performance of ChatGPT-4 and ChatGPT-4o as a learning tool in endodontics. Int Endod J 2025; 2025: 14217. [CrossRef]
22. Danesh A, Danesh A, Danesh F. Innovating dental diagnostics: ChatGPT's accuracy on diagnostic challenges. Oral Dis 2024; 2024: 14217. [CrossRef]
23. Pradhan P. Accuracy of ChatGPT 3.5, 4.0, 4o and Gemini in diagnosing oral potentially malignant lesions based on clinical case reports and image recognition. Med Oral Patol Oral Cir Bucal 2025; 30(2): e224–31. [CrossRef]
24. Frosolini A, Franz L, Benedetti S, et al. Assessing the accuracy of ChatGPT references in head and neck and ENT disciplines. Eur Arch Otorhinolaryngol 2023; 280(11): 5129–33. [CrossRef]
25. Mondillo G, Colosimo S, Perrotta A, et al. Comparative evaluation of advanced AI reasoning models in pediatric clinical decision support: ChatGPT O1 vs DeepSeek-R1. medRxiv 2025; 2025: 25321169. [CrossRef]
26. Normile D. Chinese firm's large language model makes a splash. Science 2025; 387(6731): 238. [CrossRef]
27. Temsah A, Alhasan K, Altamimi I, et al. DeepSeek in healthcare: Revealing opportunities and steering challenges of a new open-source artificial intelligence frontier. Cureus 2025; 17(2): e79221. [CrossRef]
28. Pandey S, Sharma S. A comparative study of retrieval-based and generative-based chatbots using deep learning and machine learning. Healthc Anal 2023; 3: 100198. [CrossRef]
29. Walker HL, Ghani S, Kuemmerli C, et al. Reliability of medical information provided by ChatGPT: Assessment against clinical guidelines and patient information quality instrument. J Med Internet Res 2023; 25: e47479. [CrossRef]