



Consistency and accuracy of different artificial intelligence models in evaluating longitudinal cracks and fractures of teeth

Zeynep Yağmur Özdemir Tekin¹ · Arzu Kaya Mumcu¹

Received: 10 February 2026 / Accepted: 15 September 2026

© The Author(s), under exclusive licence to The Society of The Nippon Dental University 2026

Abstract

Large language models such as ChatGPT and Google Gemini have attracted increasing attention in dentistry due to their potential to support clinical decision-making; however, their reliability in responding to endodontic questions remains unclear. This study compared the accuracy, consistency, and temporal stability of three artificial intelligence models (ChatGPT-3.5, ChatGPT-4o, and Google Gemini) when answering dichotomous clinical questions on vertical root fractures and tooth cracks using expert responses as the reference standard. Seventy-two true/false questions were developed based on the European Society of Endodontology position statement on longitudinal tooth fractures, and 60 validated items with an item-level content validity index of 1.00 were included and categorized into easy, moderate, and difficult levels. Each model was queried over 10 consecutive days at three daily time points, generating 5,400 responses in total. Overall accuracy differed significantly among models ($p < 0.001$), with ChatGPT-4o achieving the highest accuracy (86.1%), followed by ChatGPT-3.5 (85.6%) and Gemini (81.8%). No differences were observed between morning and afternoon sessions, whereas evening performance was significantly lower for Gemini ($p = 0.011$). Differences between models were evident only for moderate-difficulty questions ($p < 0.001$). Inter-model agreement was very low (Cohen's kappa (κ) = 0.019), and temporal stability differed significantly among models ($p < 0.001$), with ChatGPT-4o showing the highest stability. Although acceptable accuracy levels were observed, limited agreement and temporal variability indicate that these systems cannot replace clinician judgment and should currently be used only as supportive tools in endodontic decision-making.

Keywords Artificial intelligence · Chatbot · ChatGPT · Google Gemini · Endodontics · Large language models · Vertical root fracture

Introduction

Vertical root fractures and tooth cracks are pathologies that pose significant diagnostic and clinical challenges in endodontic practice [1]. In particular, the increasing incidence of vertical root fractures in endodontically treated teeth has led to greater complexity in the cases encountered by clinicians [2]. In many cases, symptoms are nonspecific, radiographic findings may be insufficient, and definitive differentiation of the fracture can be time-consuming. Moreover, it has been reported that if left untreated, tooth cracks may progress to more severe stages, leading to irreversible pulpal diseases

or tooth loss [3]. Therefore, the early and accurate detection of these conditions is of critical importance for determining an appropriate treatment approach and preserving the long-term prognosis of the tooth.

Diagnostic challenges in endodontics have increased interest in artificial intelligence-based systems, which have the potential to support clinical decision-making, particularly through their capabilities in big data analysis and pattern recognition. Artificial intelligence is defined as computer systems capable of performing tasks that require human intelligence and can generate predictions in complex clinical situations by processing large volumes of data [4]. Moreover, modern artificial intelligence approaches, particularly through machine learning and deep learning techniques, possess the ability to learn from data, recognize abstract relationships, and analyze increasingly complex features [5]. In recent years, large language models have attracted increasing attention in the healthcare field due to

✉ Zeynep Yağmur Özdemir Tekin
zeynepyagmur.ozdemir@ksbu.edu.tr

¹ Department of Endodontics, Faculty of Dentistry, Kutahya Health Sciences University, Kutahya 43100, Turkey

their capacity to understand natural language and generate contextually appropriate responses. Among these models, ChatGPT has drawn particular attention for its ability to produce human-like responses and to clarify verbal clinical questions [6]. Although studies in the medical field have shown that ChatGPT can achieve high accuracy rates in certain areas of basic knowledge, its performance has been reported to be variable in clinically specialized questions, with the potential to generate inaccurate or misleading information at times [7]. The complex details encountered in the imaging-based evaluation of endodontic pathologies render artificial intelligence applications an important complementary tool in this field. In particular, the ability of deep learning models to distinguish subtle anatomical structures and pathological changes on radiographs offers a notable advantage in endodontic diagnostic processes [8]. In this context, studies have demonstrated that artificial intelligence-based models can be effectively applied in the detection of periapical lesions, vertical root fractures, and root canal morphology, as well as in working length determination [9–11]. On the other hand, large language models with natural language processing capabilities have also attracted the interest of researchers for answering verbal and knowledge-based questions related to endodontics [12]. However, it has been reported that the performance of these models may vary depending on the level of expertise required by the question, and that limitations in terms of consistency and accuracy persist, particularly in more complex clinical scenarios [13].

In recent years, next-generation versions of large language models, namely ChatGPT-4o and Google Gemini Advanced, have attracted attention as tools that provide rapid access to information in the field of endodontics. Owing to their natural language processing capabilities, both models are able to instantly respond to frequently posed clinical questions, thereby facilitating access to endodontic knowledge [14]. The multimodal architecture of ChatGPT-4o enables the integrated processing of text, audio, and visual inputs, allowing it to provide more contextualized and consistent responses compared with its previous versions [15]. Similarly, Google's advanced model, Gemini Advanced, stands out for its capacity to process large datasets and its near-expert human-level, multidisciplinary performance [16]. Although both models have the potential to provide information in endodontics, it has been reported that they demonstrate different levels of performance in terms of clinical accuracy, and that the validity of their responses decreases markedly, particularly under strict evaluation criteria [14].

Therefore, comparing the accuracy, consistency, and temporal variability of the responses generated by artificial intelligence-based models on diagnostically critical issues such as vertical root fractures and tooth cracks is of great

importance for evaluating the feasibility of using these systems as reliable supportive tools in endodontic practice. In this context, to better understand the performance of contemporary large language models in endodontic knowledge generation, the present study aimed to compare the responses of ChatGPT-3.5, ChatGPT-4o, and Google Gemini models to questions related to vertical root fractures and tooth cracks in terms of accuracy, consistency, and temporal variability.

Materials and methods

Ethical approval

As this study did not involve human participants or animal subjects, approval from an ethics committee was not required.

Question design

The questions used in this study were developed based on the *Position Statement on Longitudinal Cracks and Fractures of Teeth* published by the European Society of Endodontology [3]. To evaluate the performance of three large language models (ChatGPT-3.5, ChatGPT-4o, and Google Gemini Advanced), a total of 72 dichotomous (yes/no) questions were prepared by two faculty members with at least 8 years of academic and clinical experience in the Department of Endodontics. To assess the content validity of the prepared questions, the question set was independently reviewed by two additional experienced endodontists. The experts rated each item using a four-point Likert scale according to the Content Validity Index (I-CVI) method proposed by Yusoff et al. [17].

Scores of “3=quite relevant” and “4=highly relevant” were considered acceptable; for a question to be deemed valid, both experts were required to assign a score of 3 or 4 to the item, achieving an I-CVI value of 1.00. Based on expert feedback, some items were simplified, while those with limited scientific contribution were removed from the item pool. Following the validation process, items that did not meet the criteria were excluded, and a final question pool consisting of a total of 60 validated questions was established for use in the study.

Categorization of questions

The 60 validated questions were classified into three difficulty levels—easy, moderate, and difficult—by two endodontists. Prior to classification, the evaluators jointly defined the criteria, including the complexity of the clinical

concept, the required level of knowledge, and the likelihood of incorrect responses by general dental practitioners.

In cases of disagreement between the evaluators, a third endodontist made the final decision. All questions were included in the study, with each difficulty level comprising 20 items. Representative examples from each difficulty category are provided in Table 1.

Response generation process in artificial intelligence models

Three large language models were evaluated:

1. ChatGPT-3.5
2. ChatGPT-4o
3. Google Gemini Advanced

Response generation was conducted over a 10-day period, three times daily (morning, afternoon, and evening). For each individual prompt, a new conversation (“new chat”) was initiated to provide a clean context window and to prevent the model from being influenced by previous prompts or its own earlier responses within the same conversation thread. Accordingly, each question was presented in an isolated conversational context. Each model answered all 60 questions in every session, resulting in 180 responses per day, 1,800 responses per model over 10 days, and a total of 5,400 responses across all models.

All responses were recorded in structured datasets and exported for statistical analysis.

Human expert answers

To evaluate the response accuracy of the artificial intelligence models, two experienced endodontists independently answered the 60 questions.

Statistical analysis

All artificial intelligence responses were transferred into an Excel-format dataset, and statistical analyses were performed using IBM SPSS Statistics for Windows, Version 25.0 (IBM Corp., Armonk, NY, USA). To compare the accuracy performance of the three artificial intelligence models, Pearson’s chi-square test was applied based on the

distribution of correct and incorrect responses. When pairwise comparisons between models were required, Bonferroni correction was used to reduce the risk of type I error resulting from multiple testing.

To evaluate intraday response variability, results obtained during the morning, afternoon, and evening sessions were analyzed as separate categories, and differences among these three time periods were examined using the chi-square test. As the questions were categorized into predefined difficulty levels (easy, moderate, and difficult), separate chi-square analyses were conducted for each category to assess whether the performance of the artificial intelligence models varied according to difficulty level.

Within-model temporal stability was evaluated by determining whether the accuracy status of a given question (correct or incorrect) remained unchanged across consecutive days within the corresponding time-of-day sessions. For each question, the switch rate was defined as the proportion of consecutive-day comparisons in which the accuracy status changed. Temporal stability was calculated as $1 - \text{switch rate}$, with higher values indicating greater stability of the binary outcome over time. Because only binary correct/incorrect outcomes were analyzed, semantic, lexical, or reasoning-level similarity between repeated responses was not assessed. For all tests, the level of statistical significance was set at $p < 0.05$.

Results

In this study, a total of 5,400 responses were generated by three large language models (ChatGPT-3.5, ChatGPT-4o, and Google Gemini) to 60 clinical endodontic questions over a 10-day period, with three daily sessions (morning, afternoon, and evening).

Overall accuracy rates differed significantly among the models (Pearson $\chi^2 = 15.448$, $p < 0.001$). The highest accuracy was observed for ChatGPT-4o (86.1%), followed by ChatGPT-3.5 (85.6%) and Gemini (81.8%) (Table 2).

Analysis according to time of day revealed no significant differences among models during the morning ($p = 0.151$) and afternoon ($p = 0.197$) sessions; however, model performance diverged significantly during the evening session ($p = 0.011$), with Gemini exhibiting the lowest performance.

Table 1 Representative examples of questions according to difficulty level

Difficulty level	Representative question
Easy	Is cold sensitivity one of the symptoms of cracked teeth? Only yes or no answer.
Moderate	Does Cracked Teeth incidence increase with age and become most prevalent in patients over 40 years of age? Only yes or no answer.
Difficult	Can CBCT reliably detect vertical root fractures (VRFs) within the root due to the image resolution being insufficient to detect fractures that are typically 50–100 μm in width? Only yes or no answer.

Table 2 Comparison of overall accuracy and temporal stability of the models among the models

Metric	ChatGPT-3.5	ChatGPT-4o	Gemini	Group difference (<i>p</i>)
Overall accuracy (%)	85.6	86.1	81.8	<0.001 †
95% confidence interval	84.0–87.2	84.5–87.7	80.0–83.6	—
SD (question-based)	0.83	0.82	0.91	—
Switch rate (%)	2.84	2.28	6.06	<0.001 ‡
95% confidence interval	2.03–3.65	1.56–3.01	4.89–7.22	—
Temporal stability (%)	97.16	97.72	93.94	<0.001 ‡
Inter-model agreement (Cohen's κ)	—	—	—	0.019

† Pearson chi-square test (correct/incorrect $\times$ model)

‡ Kruskal–Wallis test (question-based switch rates)

Table 3 Accuracy and temporal stability of the models according to question difficulty

Difficulty	Metric	ChatGPT-3.5	ChatGPT-4o	Gemini	Group difference (<i>p</i>)
Easy	Accuracy (%)	96.7	97.2	96.0	0.534 †
	Stability (%)	97.04	98.52	94.63	0.091 ‡
Moderate	Accuracy (%)	92.5	95.2	86.8	<0.001 †
	Stability (%)	97.96	96.11	93.70	0.003 ‡
Difficult	Accuracy (%)	67.7	66.0	62.4	0.151 †
	Stability (%)	96.48	98.52	93.49	0.078 ‡

† Pearson chi-square test (correct/incorrect $\times$ model)

‡ Kruskal–Wallis test (question-based switch rates)

When questions were categorized by difficulty level (20 easy, 20 moderate, and 20 difficult), no significant differences were found among models for easy ($p=0.534$) and difficult ($p=0.151$) questions, whereas a significant difference in accuracy was observed for moderate-level questions ($p<0.001$). In pairwise comparisons, Gemini demonstrated significantly lower accuracy than both ChatGPT-3.5 ($p=0.001$) and ChatGPT-4o ($p<0.001$), while no significant difference was observed between ChatGPT-3.5 and ChatGPT-4o ($p=0.055$).

Inter-model response agreement assessed using Cohen's kappa coefficient indicated a very low level of agreement ($\kappa=0.019$). This finding suggests that the patterns of correct and incorrect responses provided by the models to the same questions largely did not overlap.

As only binary correct/incorrect labels were analyzed, temporal stability represented outcome-based stability rather than similarity in response wording or reasoning. It was calculated as the probability of maintaining the same accuracy status across consecutive days ($1 - \text{switch rate}$). Thus, the switch rate represented temporal inconsistency, whereas temporal stability represented its complement. A significant difference in temporal stability was detected among the models (Kruskal–Wallis, $p<0.001$). The highest stability was observed for ChatGPT-4o (97.72%), followed by ChatGPT-3.5 (97.16%), whereas Gemini exhibited lower stability (93.94%) (Table 2).

Difficulty-based analyses indicated that moderate-level questions primarily drove the divergence among models. Significant differences among models were observed for moderate-difficulty questions in terms of both accuracy ($p<0.001$) and stability ($p=0.003$), whereas differences at easy and difficult levels were not significant (Table 3). These findings indicate that Google Gemini generated less

accurate and more temporally variable responses, particularly for questions requiring intermediate-level clinical reasoning.

Discussion

In recent years, growing interest in the application of large language models (LLMs) in medicine has been accompanied by a marked increase in studies investigating their potential clinical utility [12, 18, 19]. Although information obtained from internet-based sources is valuable due to its constant accessibility and ability to provide diverse perspectives, the application of such information in clinical settings requires a comprehensive evaluation of model performance and potential sources of error. Accordingly, the present study compared the responses of ChatGPT-3.5, ChatGPT-4o, and Google Gemini to dichotomous questions related to vertical root fractures and tooth cracks in terms of accuracy, consistency, and temporal variability.

The questions used in this study were developed in accordance with established clinical guidelines and were based on the *Position Statement on Longitudinal Cracks and Fractures of Teeth* published by the European Society of Endodontology in 2025. Question content was formulated to reflect clinical scenarios most frequently encountered by clinicians and patients. A similar guideline-based approach has been adopted in previous studies, both during question development and in the evaluation of responses, to ensure scientific validity and clinical relevance [12, 20].

When unable to perform a given task, LLMs may generate persuasive but incorrect content rather than explicitly indicating uncertainty, potentially producing fabricated references, incorrect terminology, or contextual errors [21]. This situation indicates that although ChatGPT possesses

the ability to recognize and organize data patterns, it is unable to fully comprehend the meaning and context of the information [4]. To reduce this risk, only dichotomous (“yes/no”) questions were used in the present study, consistent with the methodology employed by Suárez et al. [12].

Due to the probabilistic nature of LLMs, variability may be observed in their responses [22]. For this purpose, 60 questions at different difficulty levels were developed, and a total of 5,400 responses generated by the LLMs at different times of the day were analyzed. According to the findings of the present study, ChatGPT-4o demonstrated an accuracy of 86.1%, followed by ChatGPT-3.5 with 85.6%, whereas Gemini exhibited an accuracy of 81.8%. Umer and Habib [8], argue that an acceptable accuracy threshold for diagnostic tasks should exceed 90%. In the present study, although the accuracy levels of the ChatGPT-4o and ChatGPT-3.5 models approached 90%, none of the models exceeded this threshold, and Gemini exhibited the lowest accuracy level. Although ChatGPT-4o and ChatGPT-3.5 demonstrated promising results, they did not achieve the desired levels of accuracy and consistency. In a study conducted by Díaz-Flores García et al. [20] the accuracy of Gemini’s responses to endodontic diagnostic and treatment-related questions was evaluated, and only 37.11% of the responses were found to be correct. This low level of accuracy is considered insufficient for clinical application and highlights the need for caution when using LLMs, particularly in situations involving sensitive medical decision-making.

To more accurately represent model performance and to examine whether the responses exhibited variation or fluctuation at different times of the day, the same questions were posed over a 10-day period at different time points (morning, afternoon, and evening), using a new chat session each time. When accuracy levels across different times of the day were evaluated, a significant difference was observed only during the evening sessions, whereas no differences were found between the other time periods. During the evening sessions, Gemini demonstrated a lower accuracy level compared with ChatGPT-4o and ChatGPT-3.5. The lower evening performance observed for Gemini may indicate that LLM performance is not entirely stable across different times of use. Although external factors such as server load or peak usage hours may have contributed to this variability, these factors were not measured in the present study; therefore, no causal relationship can be established. From a practical perspective, clinicians should be aware that LLM performance may vary over time and should avoid relying on a single AI-generated response for clinically relevant decisions. To evaluate the temporal robustness of model performance, we examined whether each model maintained the same correct/incorrect status for a given question across consecutive days. Although temporal stability

was numerically high across all models (93.94–97.72%), it was not complete and differed significantly among the models. The corresponding switch rates ranged from 2.28% to 6.06%, with Gemini exhibiting the greatest temporal variability. These findings reflect only the temporal stability of binary accuracy outcomes and should not be interpreted as indicating consistency in the wording, semantic content, or reasoning of repeated responses.

Therefore, even when overall accuracy status remains relatively stable over time, occasional changes in correctness may still have clinical relevance when LLM-generated information is used to support decision-making.

Previous research has shown that LLMs may demonstrate limitations in tasks requiring reasoning and knowledge integration, despite satisfactory performance on more general or lower-order questions [23]. In the present study, easy questions yielded high and relatively similar accuracy across all models, which may reflect the more direct retrieval of established factual or clinically familiar information. Difficult questions, by contrast, resulted in lower absolute accuracy across all models but did not produce significant inter-model differences, suggesting that highly complex questions may have posed a comparable challenge across systems. In contrast, moderate-difficulty questions produced the greatest divergence among the models. One possible explanation is that these questions require an intermediate level of contextual interpretation, discrimination between closely related concepts, and integration of clinical information rather than simple factual recall. Such tasks may therefore be more sensitive to differences in how individual models process and prioritize clinically relevant information. This pattern may help explain why moderate questions differentiated model performance more clearly than either easy or difficult questions.

As LLMs continue to advance, they are increasingly used by both patients and clinicians in the search for health-related information. Patients may consult these models to obtain information about health conditions that may affect them, while clinicians may use LLMs to question recommended diagnostic tests or treatment plans for a specific patient and to obtain supportive information. Nevertheless, it should be borne in mind that the outputs of such models may contain inaccurate or unverified information, that algorithms may carry inherent biases, or that they may be influenced by data containing confounding effects [24]. Therefore, further research is required to ensure that these models can be used reliably and ethically in healthcare, particularly to prevent potential harm to vulnerable populations [24]. Indeed, the findings of the present study indicate that LLMs such as ChatGPT-4o, ChatGPT-3.5, and Google Gemini have reached a certain level of competence in answering questions related to vertical root fractures and

cracked teeth in endodontics; however, they also reveal that substantial limitations persist in complex cognitive abilities, including knowledge transfer, multi-step reasoning, logical inference, and interdisciplinary information integration. In this context, regular and comprehensive evaluation of performance differences among models will contribute to the more reliable and effective use of LLMs across various medical fields.

Conclusion

The findings of this study demonstrate that the three artificial intelligence models evaluated exhibited differing levels of accuracy and consistency in their responses to dichotomous clinical questions related to vertical root fractures and cracked teeth in endodontics. Although ChatGPT-4o generally achieved the highest accuracy, all models showed notable limitations, particularly in questions requiring moderate levels of clinical reasoning. Moreover, the very low inter-model agreement ($\kappa=0.019$) indicates that these models should not currently be regarded as a gold standard or a reliable standalone second opinion for endodontic decision-making. Furthermore, the variability observed in responses across different times of the day indicates that the decision stability of current artificial intelligence models remains limited.

These findings suggest that existing artificial intelligence models have not yet reached a level of maturity that would allow their use as independent clinical decision-making tools in endodontic practice. Nevertheless, ongoing advancements in model development indicate that artificial intelligence may be more effectively integrated as a supportive adjunct to clinician expertise in the future. To ensure the safe, accurate, and reliable clinical application of artificial intelligence in endodontics, regular longitudinal evaluation of different models and versions, supported by more comprehensive and clinically oriented studies, is essential.

Acknowledgements This study was presented as an oral presentation at the 8th International Health Sciences and Life Congress, held in Burdur, Türkiye, from April 16 to 19, 2025.

Author contributions Conceptualization: Z.Y.Ö.T., A.K.M.; Methodology: Z.Y.Ö.T., A.K.M.; Data curation: Z.Y.Ö.T., A.K.M.; Formal analysis: Z.Y.Ö.T., A.K.M.; Investigation: Z.Y.Ö.T., A.K.M.; Validation: Z.Y.Ö.T., A.K.M.; Visualization: Z.Y.Ö.T., A.K.M.; Supervision: A.K.M.; Writing – original draft: Z.Y.Ö.T., A.K.M.; Writing – review and editing: Z.Y.Ö.T., A.K.M.

Funding This research received no external funding.

Data availability The datasets generated and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Declarations

Conflict of interest The authors declare that they have no competing interests.

Ethical approval This study did not involve human participants or animals. Therefore, ethical approval was not required.

References

1. Kishen A. Mechanisms and risk factors for fracture predilection in endodontically treated teeth. *Endodontic Top.* 2006;13(1):57–83. <https://doi.org/10.1111/j.1601-1546.2006.00201.x>.
2. Pan X, Tang R, Gao A, Hao Y, Lin Z. Cross-sectional study of posterior tooth root fractures in 2015 and 2019 in a Chinese population. *Clin Oral Invest.* 2022;26(10):6151–7. <https://doi.org/10.1007/s00784-022-04564-0>.
3. Patel S, Teng PH, Liao WC, Davis MC, Fidler A, Haupt F, et al. Position statement on longitudinal cracks and fractures of teeth. *Int Endod J.* 2025;58(3):379–90. <https://doi.org/10.1111/iej.14186>.
4. Sinha RK, Roy AD, Kumar N, Mondal H, Sinha R Applicability of ChatGPT in assisting to solve higher order problems in pathology. *Cureus.* 2023. <https://doi.org/10.7759/cureus.35237>.
5. Xu L, Sanders L, Li K, Chow JC. Chatbot for health care and oncology applications using artificial intelligence and machine learning: systematic review. *JMIR cancer.* 2021;7(4):e27850. <https://doi.org/10.2196/27850>.
6. Abd-Alrazaq A, AlSaad R, Alhuwail D, Ahmed A, Healy PM, Latifi S, et al. Large language models in medical education: opportunities, challenges, and future directions. *JMIR Med Educ.* 2023;9(1):e48291. <https://doi.org/10.2196/48291>.
7. Korteling JE, van de Boer-Visschedijk GC, Blankendaal RA, Boonekamp RC, Eikelboom AR. Human-versus artificial intelligence. *Front Artif Intell.* 2021;4:622364. <https://doi.org/10.3389/frai.2021.622364>.
8. Umer F, Habib S. Critical analysis of artificial intelligence in endodontics: a scoping review. *J Endod.* 2022;48(2):152–60. <https://doi.org/10.1016/j.joen.2021.11.007>.
9. Yang P, Guo X, Mu C, Qi S, Li G. Detection of vertical root fractures by cone-beam computed tomography based on deep learning. *Dentomaxillofac Radiol.* 2023;52(3):20220345. <https://doi.org/10.1259/dmfr.20220345>.
10. Hiraiwa T, Ariji Y, Fukuda M, Kise Y, Nakata K, Katsumata A, et al. A deep-learning artificial intelligence system for assessment of root morphology of the mandibular first molar on panoramic radiography. *Dentomaxillofac Radiol.* 2019;48(3):20180218. <https://doi.org/10.1259/dmfr.20180218>.
11. Wu W, Chen S, Chen P, Chen M, Yang Y, Gao Y, et al. Identification of root canal morphology in fused-rooted mandibular second molars from X-ray images based on deep learning. *J Endod.* 2024;50(9):1289–97. <https://doi.org/10.1016/j.joen.2024.05.014>.
12. Suárez A, Díaz-Flores García V, Algar J, Gómez Sánchez M, Llorente de Pedro M, Freire Y. Unveiling the ChatGPT phenomenon: evaluating the consistency and accuracy of endodontic question answers. *Int Endod J.* 2024;57(1):108–13. <https://doi.org/10.1111/iej.13985>.
13. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLoS Digit health.* 2023;2(2):e0000198. <https://doi.org/10.1371/journal.pdig.0000198>.

14. Dufey-Portilla N, Frisman AB, Robles MG, Peña-Bengoa F, Ávila CC, Nagendrababu V, et al. Assessing the validity of ChatGPT-4o and Google Gemini Advanced when responding to frequently asked questions in endodontics. *J Appl Oral Sci.* 2025;33:e20250321. <https://doi.org/10.1590/1678-7757-2025-0321>.
15. OpenAI, Hello. GPT-4o 2024 [Available from: <https://openai.com/index/hello-gpt-4o/>]
16. Team G, Anil R, Borgeaud S, Alayrac J-B, Yu J, Soricut R, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv.* 2023;2312.11805. <https://doi.org/10.48550/arXiv.2312.11805>.
17. Yusoff MSB. ABC of content validation and content validity index calculation. *Educ Med J.* 2019;11(2):49–54. <https://doi.org/10.21315/eimj2019.11.2.6>.
18. Ozden I, Gokyar M, Ozden ME, Sazak Ovecoglu H. Assessment of artificial intelligence applications in responding to dental trauma. *Dent Traumatol.* 2024;40(6):722–9. <https://doi.org/10.1111/edt.12965>.
19. Snigdha NT, Batul R, Karobari MI, Adil AH, Dawasaz AA, Hameed MS, 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. <https://doi.org/10.1155/2024/1119816>.
20. Díaz-Flores García V, Freire Y, Tortosa M, Tejedor B, Estevez R, Suárez A. Google Gemini's performance in endodontics: a study on answer precision and reliability. *Appl Sci.* 2024;14(15):6390. <https://doi.org/10.3390/app14156390>.
21. Seth I, Kenney PS, Bulloch G, Hunter-Smith DJ, Thomsen JB, Rozen WM. Artificial or augmented authorship? A conversation with a chatbot on base of thumb arthritis. *Plast Reconstr Surgery–Global Open.* 2023;11(5):e4999. <https://doi.org/10.1097/GOX.0000000000004999>.
22. Antaki F, Touma S, Milad D, El-Khoury J, Duval R. Evaluating the performance of ChatGPT in ophthalmology: an analysis of its successes and shortcomings. *Ophthalmol Sci.* 2023;3(4):100324. <https://doi.org/10.1016/j.xops.2023.100324>.
23. Shojaei-Mend H, Mohebbati R, Amiri M, Atarodi A. Evaluating the strengths and weaknesses of large language models in answering neurophysiology questions. *Sci Rep.* 2024;14(1):10785. <https://doi.org/10.1038/s41598-024-60405-y>.
24. Duong D, Solomon BD. Analysis of large-language model versus human performance for genetics questions. *Eur J Hum Genet.* 2024;32(4):466–8. <https://doi.org/10.1038/s41431-023-01396-8>.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.