Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

The Performance of Artificial Intelligence Applications in the Diagnosis, Radiographic Localization, and Treatment Planning of Periapical Lesions Under Zero-Shot Conditions

Elif Yıldız

elif.yldz.99@gmail.com

Zonguldak Bülent Ecevit University

Emre Bodrumlu

Zonguldak Bülent Ecevit University

Research Article

Keywords: Apical periodontitis, Artificial intelligence, Diagnostic accuracy, Panoramic radiography, Treatment planing

Posted Date: November 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11666448/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

The aim of this study is to comparatively evaluate the performance of three different artificial intelligence applications (Claude, ChatGPT and Kimi Chat) in the diagnosis of periapical lesions, radiographic localization and treatment planning in panoramic radiographs, tested under blind conditions, using the assessment of two endodontic specialists as the reference standard.

Methods

The study included 100 panoramic radiographs. Each radiograph was presented to the artificial intelligence applications alongside the relevant clinical scenario and using a standard prompt. The model responses were classified as correct or incorrect by comparing them with the reference standard in terms of diagnosis, radiographic marking and treatment planning. Accuracy rates were calculated as n (%) and 95% confidence intervals were determined using the Wilson method. The Cochran's Q test was used for comparisons between models, whilst the Bonferroni-corrected McNemar exact test was used for pairwise comparisons. Agreement between models was assessed using Fleiss' Kappa, whilst inter-rater and intra-rater agreement were assessed using Cohen's Kappa coefficient.

Results

The correct diagnosis rates were 36% for ChatGPT, 20% for Kimi, and 15% for Claude ($p = 0.002$), with ChatGPT performing significantly better than Claude ($p < 0.001$). In radiographic localization, Claude achieved the highest accuracy (90%), followed by Kimi (70%) and ChatGPT (54%), with significant differences in all pairwise comparisons ($p < 0.0167$). Accuracy in treatment planning ranged from 95% to 100%, with no significant difference between the models ($p = 0.082$). Inter- and intra-rater agreement was perfect ($\kappa = 1.00$).

Conclusion

The findings indicate that off-the-shelf (general-purpose) AI applications are not suitable for use as a primary, independent diagnostic tool in the interpretation of endodontic radiographs. However, their success in tasks such as radiographic annotation and treatment planning suggests that these models could be used as decision-support tools under specialist supervision or as auxiliary tools in dental education.

Background

Chronic periapical periodontitis is a clinical condition typically evident on radiographs as a periapical radiolucent area, and which frequently arises following pulp necrosis. This condition reflects a localised inflammatory reaction developing in the periapical tissues in response to microbial invasion originating from the root canal system (1, 2). Lesions forming in the periapical region of teeth may progress for a long time, even for years, without causing any symptoms and are often discovered incidentally on radiographs taken during routine dental examinations. The condition may progress from a chronic, asymptomatic state to an acute, symptomatic flare-up. This change is typically triggered by a disruption in the host-microbe relationship and results in a rapid increase in inflammatory mediators, tissue oedema and the formation of purulent exudate, culminating in a clinical presentation characterised by an acute apical abscess. The accompanying pain is usually severe, constant and throbbing in nature; it is exacerbated by percussion and palpation due to increased pressure within the rigid boundaries of the alveolar bone (3). Treatment of such acute exacerbations involves the removal of the causative agent (pulp debris and microorganisms), ensuring drainage and maintaining infection control. Whilst the use of antibiotics and analgesics is indicated in the presence of systemic symptoms (e.g. fever, lymphadenopathy) or in cases of spreading infection, treatment is primarily achieved through cavity preparation and chemomechanical debridement of the root canal system (4–6).

Panoramic radiography is one of the most commonly used imaging methods in dentistry for the detection of lesions in the periapical region of the teeth. It provides an image of both jaws using a low radiation dose (7, 8). These images are requested during patients' initial examinations. For this reason, periapical lesions are often first detected on panoramic images. However, when using panoramic radiography to assess periapical regions, certain limitations arise, such as its relatively low resolution, as well as distortion and ghosting that may occur in the image. Furthermore, as periapical radiolucency only becomes visible in two-dimensional images once significant mineral loss has occurred, small or early-stage lesions may not be detected (9, 10). These limitations pose challenges even for experienced clinicians in accurately identifying periapical lesions and formulating an appropriate treatment plan.

In recent years, there has been a growing focus on utilising technological advancements to support such clinical decision-making processes and improve treatment outcomes; in this context, artificial intelligence-based approaches have begun to attract attention. Artificial Intelligence (AI) encompasses computer systems capable of performing tasks traditionally associated with human intelligence, such as learning, problem-solving and decision-making. AI enables the identification of data that is difficult to detect using traditional methods. The development of AI systems has become widespread across many areas of medicine and the healthcare sector. In dentistry, AI can be utilised through question-and-answer applications that assist with diagnosis and suggest treatment methods (11).

AI is being extensively researched to support radiographic diagnosis in endodontics. AI applications trained for specific tasks are showing promising performance in the detection of apical lesions on panoramic radiographs. The accuracy, sensitivity and specificity values reported in studies are comparable to those of specialist dentists (10, 12, 13). However, these systems are designed for a single, specifically defined task. Furthermore, these systems mostly merely report the presence of a

lesion and do not translate this finding into a clinical treatment recommendation (14, 15). AI applications developed in recent years have become accessible to everyone via simple chat interfaces. Systems such as Claude Opus 5.5 (Anthropic, San Francisco, CA, USA), ChatGPT (GPT-6, OpenAI, San Francisco, CA, USA) and Kimi (K3.1 Pro, Moonshot AI, Beijing, China) are capable of processing an uploaded image alongside a question posed in natural language and generating responses, including diagnoses and treatment recommendations (16). Unlike AI applications trained for a specific task, these models are used without domain-specific training, that is, under zero-shot conditions (17, 18).

In the current literature, AI systems generally evaluate a single task on periapical radiographs. It has been observed that there has been insufficient research into studies that address lesion detection, localisation at the tooth level and treatment planning recommendations from panoramic radiographs within the same clinical scenario. To this end, the aim of this study is to comparatively evaluate the performance of three artificial intelligence applications (Claude, ChatGPT, Kimi) in the diagnosis of periapical lesions, radiographic localization, and treatment planning on panoramic radiographs, using the assessment of two endodontic specialists as the reference standard.

Methods

Sample selection

Panoramic radiographs were retrieved from an open-access radiography website (19) featuring well-documented clinical cases. Panoramic radiographs containing at least one periapical lesion were included in the study. Blurred or manipulated panoramic radiographs, as well as images containing markings such as arrows, circles, lines or text, were excluded from the study. No human or animal subjects were used in this study, and no personally identifiable data was processed. Consequently, there was no need for ethical committee approval or informed consent.

Sample size calculation

The a priori sample size calculation was performed using the G*Power 3.1 software (Heinrich Heine University, Düsseldorf, Germany), based on the McNemar exact test. The expected accuracy rates were taken from the study by Adıgüzel et al. (18), which evaluated three artificial intelligence applications under zero-shot conditions and reported diagnostic accuracies of 54.2 per cent and 71.7 per cent for the lowest and highest values, respectively. As all models assessed the same radiographs, assuming a moderate correlation between paired responses, the discordance rate was calculated as 0.3412 and the odds ratio as 3.1054. It was determined that at least 94 matched observations (radiographs) were required with a two-sided $\alpha = 0.05$ and 80 per cent power. Consequently, 100 panoramic radiographs were included in the study, and the estimated power was calculated as 0.83.

Clinical scenarios

For each radiograph, a standard clinical scenario consistent with the radiographic findings was prepared by an endodontist. The scenarios were developed based on previous studies that evaluated the performance of artificial intelligence applications in endodontic diagnosis and treatment planning using structured clinical case data (20–22). Each scenario includes the patient's presenting complaint, medical and dental history, and clinical findings as defined in the American Association of Endodontists (AAE) diagnostic terminology (spontaneous pain, duration of symptoms, swelling, presence of a fistula, response to percussion and palpation, pulp sensitivity tests) (6, 23), and was written using a pre-determined template to ensure consistency across cases.

Reference standard

Assessments carried out by endodontists were used as a reference standard to evaluate the performance of language models (24). All assessments were conducted by two endodontists with over ten years' clinical experience under standardised digital imaging conditions. The assessors examined the panoramic radiographs and associated clinical scenarios independently of one another and without access to the responses generated by the language models, thereby determining the diagnosis, the radiographic localisation of the lesion and the treatment plan for each case. In cases where there was a discrepancy between the assessors, the images were re-examined jointly by the two specialists and the final classification was determined by consensus (18). Inter-assessor agreement was calculated using Cohen's Kappa coefficient based on the initial independent assessments. It was planned that, in the event of disagreement between assessors, the images would be re-examined jointly by two specialists and the final classification determined by consensus (18); however, no disagreement arose in any case. Inter-assessor agreement was calculated using Cohen's Kappa coefficient based on the initial independent assessments. To determine intra-rater reproducibility, all images were re-evaluated by the same assessors after a two-week interval (24). Agreement was determined using Cohen's Kappa coefficient for the categories of diagnosis, radiographic localisation and treatment plan for each radiograph, and was interpreted according to the Landis and Koch classification (25). Inter-rater agreement for diagnosis, radiographic localisation and treatment planning ($\kappa = 1.00$), as well as intra-rater agreement between the first and second assessors ($\kappa = 1.00$ for both), demonstrated perfect agreement; these values correspond to a level of agreement that is nearly perfect according to the Landis and Koch classification.

The artificial intelligence models evaluated and the evaluation process

Three publicly available artificial intelligence applications were evaluated: Claude Opus 5.5 (Anthropic, San Francisco, CA, USA), ChatGPT (GPT-6, OpenAI, San Francisco, CA, USA) and Kimi (K3.1 Pro, Moonshot AI, Beijing, China). All models were accessed via their official web interfaces on 29 October 2026 using paid-pro accounts, with no changes made to the default settings (26).

The models were tested under zero-shot conditions, without any example cases or pre-training (27, 28). To prevent cross-contamination between cases, each radiograph and clinical scenario was loaded into a new and independent chat session for each model. The order of the cases was determined by a computer-generated random sequence, and the same order was applied to all models (24). The following standard prompt was used for all cases: “This is an anonymised panoramic radiograph presented alongside the patient’s clinical information. Please assess the case and answer the following questions: 1- What is your diagnosis? 2- Which tooth or region is affected? Please mark it on the radiograph. 3- What is your recommended treatment protocol?”. Each case was assessed once by each model, and the first complete response was recorded for analysis (26).

The responses of each AI application were compared against the reference standard and evaluated across three distinct parameters:

1. **Diagnosis**: Whether the application identified the case solely as an acute apical abscess, or correctly identified it as an acute flare-up of pre-existing chronic apical periodontitis (29). Responses were classified as correct or incorrect based on the reference diagnosis (18, 27).
2. **Radiographic localization**: Whether the application indicated the relevant finding in the correct region of the panoramic radiograph (by directly marking it on the uploaded panoramic image using a circle, box or arrow). Responses were classified as correct marking (if the marker was placed on the periapical region of the affected tooth as defined in the reference standard) or incorrect marking (if the marker was placed on a different tooth or region) (24, 30).
3. **Treatment planning**: Whether the treatment recommended by the application is consistent with the treatment plan in the reference standard. Responses were defined as correct or incorrect (31).

Statistical Analysis

The data were analysed using the Python 3.13 (Python Software Foundation, Wilmington, DE, USA) programming language, with the statsmodels 0.15.0 and SciPy 1.18.1 libraries. The performance of the applications was evaluated in terms of descriptive ratios by comparing them with a reference standard established by endodontic specialists; the number and percentage [n (%)] of correct and incorrect responses were calculated for each of diagnosis, radiographic localization, and treatment planning, and 95% confidence intervals for the ratios were determined using the Wilson method. In radiographic marking, the rates for the ‘correct marking’, ‘incorrect marking’ and ‘no marking’ categories were reported separately; for inter-model comparison, responses were binary categorised as ‘correct marking’ and ‘incorrect marking’.

As the same radiographs were assessed by three YZ applications, the correct response rates between the applications were compared using Cochran’s Q test. Where a significant difference was detected, pairwise comparisons were performed using the McNemar exact test, and a Bonferroni correction was applied for multiple comparisons (corrected significance level $p < 0.0167$) (2). The agreement between the three models for each criterion was assessed using Fleiss’s Kappa coefficient. Inter-rater and intra-

rater agreement were calculated using Cohen's Kappa coefficient. All Kappa values were interpreted according to the classification by Landis and Koch. A p-value of < 0.05 was considered statistically significant.

Results

A total of 100 panoramic radiographs were evaluated by three artificial intelligence models (ChatGPT, Claude and Kimi). The number of correct responses, percentages and 95% confidence intervals for each model in terms of diagnosis, radiographic localization and treatment planning are presented in Table 1 and Fig. 1.

The highest correct diagnosis rate was observed in ChatGPT [36.0% (%95 CI: 27.3–45.8)]. This was followed by Kimi [20.0% (95% CI: 13.3–28.9)] and Claude [15.0% (95% CI: 9.3–23.3)]. There is a statistically significant difference between the models (Cochran's Q = 12.89; p = 0.002). In pairwise comparisons, ChatGPT's accuracy rate was found to be significantly higher than that of Claude (p < 0.001). The difference between ChatGPT and Kimi was not found to be significant after Bonferroni correction (p = 0.029). No significant difference was found between Claude and Kimi either (p = 0.458) (Table 2).

A statistically significant difference was found between the models in radiographic localization (Cochran's Q = 31.48, p < 0.001). Claude correctly marked the lesion in 90.0% of the radiographs (95% CI: 82.6–94.5), followed by Kimi (70.0%; 95% CI: 60.4–78.1) and ChatGPT (54.0%; 95% CI: 44.3–63.4). All pairwise comparisons were significant after Bonferroni correction (p < 0.0167): Claude outperformed both ChatGPT (p < 0.001) and Kimi (p = 0.001), and Kimi outperformed ChatGPT (p = 0.014) (Table 2, Fig. 1).

Claude recommended the correct treatment plan in all cases [100.0% (95% CI: 96.3–100.0)]. The correct response rate for ChatGPT and Kimi was 95.0% (95% CI: 88.8–97.8). The difference between the models was not statistically significant (Cochran's Q = 5.00; p = 0.082).

The agreement between the three AI applications was found to be poor across all criteria, and no statistically significant differences were observed. For diagnosis, $\kappa = -0.033$ (p = 0.564); for radiographic marking, $\kappa = -0.011$ (p = 0.854); and for treatment planning, $\kappa = -0.034$ (p = 0.550). The proportion of cases where the three models gave the same response for the same radiograph was 44.0% for diagnosis, 38.0% for radiographic marking and 90.0% for treatment planning (Table 3).

The two endodontic specialists showed perfect agreement with one another across all criteria (100%). The first and second measurements by both specialists yielded similar results across all assessment groups. Consequently, both inter-rater and intra-rater agreement were excellent.

Table 1
Correct response rates of the models, n (%) [95% CI]

Criterion	Claude	ChatGPT	Kimi
Diagnosis	15 (15.0) [9.3–23.3]	36 (36.0) [27.3–45.8]	20 (20.0) [13.3–28.9]
Radiographic localization	90 (90.0) [82.6–94.5]	54 (54.0) [44.3–63.4]	70 (70.0) [60.4–78.1]
Treatment planning	100 (100.0) [96.3–100.0]	95 (95.0) [88.8–97.8]	95 (95.0) [88.8–97.8]

CI, confidence interval (Wilson score method). n = 100 radiographs.

Table 2
Comparison of correct response rates between the models

Criterion	Cochran's Q (p)	Claude-ChatGPT	Claude-Kimi	ChatGPT-Kimi
Diagnosis	12.89 (0.002)	**< 0.001**	0.458	0.029
Radiographic localization	31.48 (< 0.001)	**< 0.001**	**0.001**	**0.014**
Treatment planning	5.00 (0.082)	–	–	–

Pairwise comparisons were performed using the exact McNemar test; values are p values. Values in bold are significant after Bonferroni correction (p < 0.0167). Pairwise comparisons were not performed for treatment planning, as Cochran's Q test revealed no significant difference.

Table 3
Agreement among the three models (Fleiss' kappa)

Criterion	κ	p	Complete agreement, %	Interpretation*
Diagnosis	−0.033	0.564	44.0	Poor
Radiographic localization	−0.011	0.854	38.0	Poor
Treatment planning	−0.034	0.550	90.0	Poor

Complete agreement: proportion of radiographs for which all three models gave the same response.
**According to the Landis and Koch classification.*

*Error bars represent 95% confidence intervals calculated using the Wilson score method. Exact McNemar test with Bonferroni correction: *p < 0.0167; **p < 0.01; ***p < 0.001.*

Figure 1. Correct response rates of the models for diagnosis, radiographic localization and treatment planning.

In our study, all three AI applications achieved high accuracy in treatment planning (95–100%), whereas correct diagnosis rates remained low (15–36%), making diagnosis the most challenging task for the models. In radiographic localization, by contrast, considerable and statistically significant differences were observed between the models (54–90%).

Discussion

The use of AI applications in dentistry for diagnosis, treatment planning and prognosis assessment is on the rise. These models are also emerging as clinical decision-support tools (14).

It has been shown that performance differences between models are related not only to theoretical knowledge, but also to data diversity, contextual processing capacity and the ability to generate reasoning (32). ChatGPT, which had the highest accuracy rate, was only able to reach the correct diagnosis in one-third of cases. The findings of our study are consistent with recent studies examining the diagnostic performance of three different AI applications on dental radiographs. One study reported that, under zero-shot conditions, none of the Claude 4.5 Sonnet, GPT-5.2 Thinking or Gemini 3.0 Pro models achieved sufficient performance for independent clinical use in the detection of apical periodontitis (18). Bubna et al. (27) also reported that ChatGPT and Gemini demonstrated high sensitivity but low specificity in the detection of periapical lesions. Similarly, Hatipoğlu et al. (30) demonstrated that the accuracy of dental condition diagnosis across six different models ranged between 45.5% and 60.0%, with no significant difference found between the models. Although some of these studies were conducted using periapical radiographs, the fact that similar results were obtained suggests that the limitation stems more from the models themselves than from the imaging method. In panoramic radiography studies using the same imaging method as our study, Liu et al.(33) reported that GPT-4o was inadequate at detecting findings such as caries and periapical lesions, whilst Silva et al. (34) reported that ChatGPT-3.5 had limited success in making a differential diagnosis of radiolucent lesions. The reason for this low diagnostic success is that AI models have not been trained using specialised codes specific to dental radiographs. Furthermore, in panoramic radiographs, the superimposition of anatomical structures and ghost images, along with regional distortions, makes it difficult to distinguish small and early-stage lesions in particular (35). This finding is consistent with the report by Ourang et al. (36), who noted that AI performs well with clearly defined, high-contrast radiographic lesions but struggles with low-contrast, artefact-prone or anatomically complex structures, Ariji et al. (37), Akyüz et al. (31) and Fontenele and Jacobs (38).

The diagnosis of periapical lesions requires the evaluation of radiographic findings in conjunction with clinical information(39). In our study, although clinical history was provided alongside radiographs for AI applications, diagnostic accuracy remained low and, in most cases, it was not possible to distinguish between an acute flare-up of a chronic lesion. This finding suggests that the models struggle to integrate

visual and textual information into a single diagnostic outcome. Similarly, Buckley et al. (40) reported that whilst an AI model could combine visual and textual information in challenging medical cases, its performance declined when an image was added to an already sufficiently informative text. Atakır et al. also found that adding images to the clinical context and diagnostic options in ChatGPT-4o did not significantly improve performance (41). The ability of AI applications to correctly identify the differential diagnosis depends on the clear and distinct presentation of radiographic or anamnestic clues indicating a chronic lesion in the case (e.g. the borders of the lesion, duration of symptoms, previous asymptomatic period, presence of a sinus tract) (3). It is thought that the models exhibit a tendency to favour the diagnosis of 'acute apical abscess', which is more frequently and prototypically represented in the literature. Furthermore, AI models tend to favour more common diagnostic categories when faced with clinical uncertainty; this tendency may have led the models to produce erroneous results by favouring common diagnoses, particularly in cases where differentiation is difficult (42, 43). Experts, however, have reached the correct diagnosis in all cases by evaluating radiographic and clinical findings together, drawing on their clinical experience.

It has been reported in the literature that the performance of AI models in detecting lesions on radiographs and determining their location may vary (30). Bubna et al. (27) reported that in 73–78 per cent of responses that correctly identified the lesion, the relevant tooth was incorrectly identified; whilst de Araujo et al. (24) demonstrated that there were marked differences between AI models in identifying the affected tooth and tooth group. When the findings from radiographic annotation were evaluated, they showed results similar to those of previous studies. It was found that Claude, which showed the lowest success rate in diagnosis, was the most successful model in marking, whilst ChatGPT, which was the most successful in diagnosis, had the lowest accuracy in marking. This discrepancy demonstrates that a model's ability to correctly identify the relevant region on a radiograph does not necessarily imply that it can correctly interpret the findings in that region. Visual localisation and diagnostic interpretation appear to be distinct, independently functioning capabilities in these models.

Treatment planning is a critical clinical decision-making process that requires a comprehensive assessment of clinical and radiographic findings and directly influences the outcome of endodontic treatment. All three artificial intelligence applications demonstrated high accuracy in treatment planning, and no significant differences were observed between the models. Similarly, Almeida et al. (44), in their study evaluating eight different chatbots, reported high accuracy rates of 97–100% in treatment planning, similar to those in our study. As endodontic treatment options are more limited and generalisable than diagnostic options, it is conceivable that the models may have recommended the correct treatment based on the symptoms and findings in the clinical history presented to them, rather than on the interpretation of the radiograph (45). The current literature suggests that reliable AI-assisted treatment planning will require the integration of multimodal data, combining radiographic, clinical, symptomatic and patient-reported information (46–48).

Conclusions

The findings indicate that off-the-shelf (general-purpose) AI applications are not suitable for use as a primary, independent diagnostic tool in the interpretation of endodontic radiographs. However, their success in tasks such as radiographic annotation and treatment planning suggests that these models could be used as decision-support tools under specialist supervision or as supplementary tools in dental education.

Abbreviations

AI Artificial Intelligence

CI Confidence Intervals

Declarations

>

Ethics approval and consent to participate

Not applicable.

Competing interests

The authors declare no competing interests.

Funding

The study received no funding.

Author Contribution

EB coordinated the research project. EB and EY participated in conceptualization of the study, interpretation of the results and drafting, writing, revising, and editing of the manuscript. All the authors have read and approved the final version of the manuscript.

Data Availability

The data will be made available upon reasonable request to the corresponding author.

References

1. Alyousef H, Almohaimeed AA, Alyousef HM. Healing of a large periapical lesion using non-surgical root canal retreatment: A case report. Cureus. 2024;16(9).
2. Tibúrcio-Machado CS, Michelon C, Zanatta F, Gomes MS, Marin JA, Bier CA. The global prevalence of apical periodontitis: a systematic review and meta-analysis. Int Endod J. 2021;54(5):712–35.
3. Sekar R, Maheswaran T, Chittrarasu M, Vijayparthiban S, Suresh R. CHITTRARASU M. Acute Exacerbation of Chronic Periapical Pathology in a Dentist: An Autobiographical Case Report. Cureus. 2026;18(1).
4. Lockhart PB, Tampi MP, Abt E, Aminoshariae A, Durkin MJ, Fouad AF, et al. Evidence-based clinical practice guideline on antibiotic use for the urgent management of pulpal-and periapical-related dental pain and intraoral swelling: A report from the American Dental Association. J Am Dent Association. 2019;150(11):906–21. e12.
5. Cope AL, Francis N, Wood F, Chestnutt IG. Systemic antibiotics for symptomatic apical periodontitis and acute apical abscess in adults. Cochrane database Syst reviews. 2018(9).
6. Duncan HF, Kirkevang LL, Peters OA, El-Karim I, Krastl G, Del Fabbro M, et al. Treatment of pulpal and apical disease: the European Society of Endodontology (ESE) S3-level clinical practice guideline. Int Endod J. 2023;56:238–95.
7. Różyło-Kalinowska I. Panoramic radiography in dentistry. Clin dentistry reviewed. 2021;5(1):26.
8. Antony DP, Thomas T, Nivedhitha M. Two-dimensional periapical, panoramic radiography versus three-dimensional cone-beam computed tomography in the detection of periapical lesion after endodontic treatment: A systematic review. Cureus. 2020;12(4).
9. Gliga A, Imre M, Grandini S, Marruganti C, Gaeta C, Bodnar D, et al. The limitations of periapical X-ray assessment in endodontic diagnosis—a systematic review. J Clin Med. 2023;12(14):4647.
10. Szabó V, Orhan K, Dobó-Nagy C, Veres DS, Manulis D, Ezhov M, et al. Deep learning-based periapical lesion detection on panoramic radiographs. Diagnostics. 2025;15(4):510.
11. Mladenovic R, Milosavljevic M, Stanisic D, Vasovic M. Importance of artificial intelligence in the analysis of children's CBCT imaging by dental students. J Dent Educ. 2023;87:914–6.
12. Ekert T, Krois J, Meinhold L, Elhennawy K, Emara R, Golla T, Schwendicke F. Deep learning for the radiographic detection of apical lesions. J Endod. 2019;45(7):917–22. e5.
13. Ba-Hattab R, Barhom N, Osman SAA, Naceur I, Odeh A, Asad A, et al. Detection of periapical lesions on panoramic radiographs using deep learning. Appl Sci. 2023;13(3):1516.
14. Schwendicke Fa, Samek W, Krois J. Artificial intelligence in dentistry: chances and challenges. J Dent Res. 2020;99(7):769–74.
15. Ong Ly C, Unnikrishnan B, Tadic T, Patel T, Duhamel J, Kandel S, et al. Shortcut learning in medical AI hinders generalization: method for estimating AI model generalization without external data. NPJ Digit Med. 2024;7(1):124.
16. Liu Z, Nalley A, Hao J, Ai QYH, Yeung AWK, Tanaka R, Hung KF. The performance of large language models in dentomaxillofacial radiology: a systematic review. Dentomaxillofacial Radiol.

2025;54(8):613–31.
17. Ferber D, Wölflein G, Wiest IC, Ligero M, Sainath S, Ghaffari Laleh N, et al. In-context learning enables multimodal large language models to classify cancer pathology images. Nat Commun. 2024;15(1):10104.
18. Adıgüzel M, Özeken F, Özeken T. Zero-shot detection of apical periodontitis on periapical radiographs using multimodal large language models: diagnostic accuracy and error patterns. BMC Oral Health. 2026;26(1):1525.
19. [Available from: https://radiopaedia.org/search?q=Orthopantomography.
20. Seo D, Cheong J, Choi Y, Shin Y, Park W. Large Language Models for Endodontic Symptom Assessment and Treatment Planning Using Image-Free Clinical Records: Comparative Evaluation Study. JMIR Med Inf. 2026;14(1):e86145.
21. Karaca B, Çakmak YE, Erkal D. Clinical relevance of large language models in endodontics: diagnostic appropriateness based on 50 simulated case scenarios. Australian Endodontic J. 2026;52(1):130–8.
22. Durmazpinar PM, Ekmekci E. Comparing diagnostic skills in endodontic cases: dental students versus ChatGPT-4o. BMC Oral Health. 2025;25(1):457.
23. Essam O, Umerji S, Blundell K. Endodontic assessment, complexity, diagnosis and treatment planning: Endodontic assessment, complexity, diagnosis and treatment planning. Br Dent J. 2025;238(7):441–7.
24. de Araujo BMM, Cavassin ME, Abuabara A, Giacon ML, Fariniuk LF, Kirschneck C et al. Diagnostic performance of multimodal large language models in detecting pulp stones on panoramic radiographs. BMC Oral Health. 2026.
25. Landis JR, Koch GG. The measurement of observer agreement for categorical data. biometrics. 1977:159 – 74.
26. Abuabara A, Baratto-Filho F, Gallego GR, Luiz LBP, Meger MN, Scariot R et al. Comparative evaluation of ChatGPT and Gemini in detecting external apical root resorption on panoramic radiographs of orthodontic patients: Artificial intelligence performance in detecting root resorption. J Orofac Orthopedics/Fortschritte der Kieferorthopädie. 2025:1–9.
27. Bubna DP, Mattos NHR, Luiz LBP, Baratto-Filho F, Mattos-Calil MTd, Silva-Sousa YTC, et al. Can Large Language Models Detect Periapical Lesions in Anterior Teeth? A Comparative Study. Braz Dent J. 2025;36:e25–6861.
28. Lisboa RM, Braido A, de-Jesus-Soares A, Tewari N, Soares CJ, Paranhos LR, Vieira WA. Performance of five free large language models in dental trauma: a 30-day longitudinal benchmark study. Front Oral Health. 2025;6:1737114.
29. Glickman G. AAE consensus conference recommended diagnostic terminology. J Endod. 2009;35(12):1634.
30. Hatipoğlu Ö, Baytar M, Pertek Hatipoğlu F. Diagnostic and localization performance of multimodal large language models in the interpretation of dental periapical radiographs. BMC Oral Health. 2026.

31. Akyüz İE, Kivircik BE, Aslan T. Artificial Intelligence-Assisted Periapical Radiographic Assessment: Lesion Detection, Endodontic Complication Analysis, and Review of Clinical Treatment Recommendations. J Endod. 2026.
32. DÜNDAR M, PINAR A, ÖZ H. Diş Hekimliği Uzmanlık Sınavı Endodonti Sorularının Üç Farklı Yapay Zekâ Modeli Tarafından Yanıtlanma Performansının Karşılaştırılması: 2012–2025 Retrospektif Analizi. Turkiye Klinikleri J Dent Sci. 2026;32(4):1087–96.
33. Liu Z, Ai QYH, Yeung AWK, Tanaka R, Nalley A, Hung KF. Performance of a vision-language model in detecting common dental conditions on panoramic radiographs using different tooth numbering systems. Diagnostics. 2025;15(18):2315.
34. Silva TP, Andrade-Bortoletto MFS, Ocampo TSC, Alencar-Palha C, Bornstein MM, Oliveira-Santos C, Oliveira ML. Performance of a commercially available Generative Pre-trained Transformer (GPT) in describing radiolucent lesions in panoramic radiographs and establishing differential diagnoses. Clin Oral Invest. 2024;28(3):204.
35. Estrela C, Bueno MR, Leles CR, Azevedo B, Azevedo JR. Accuracy of cone beam computed tomography and panoramic and periapical radiography for detection of apical periodontitis. J Endod. 2008;34(3):273–9.
36. Ourang SA, Sohrabniya F, Mohammad-Rahimi H, Dianat O, Aminoshariae A, Nagendrababu V, et al. Artificial intelligence in endodontics: fundamental principles, workflow, and tasks. Int Endod J. 2024;57(11):1546–65.
37. Ariji Y, Yanashita Y, Kutsuna S, Muramatsu C, Fukuda M, Kise Y et al. Automatic detection and classification of radiolucent lesions in the mandible on panoramic radiographs using a deep learning object detection technique. Oral surgery, oral medicine, oral pathology and oral radiology. 2019;128(4):424–30.
38. Fontenele RC, Jacobs R. Unveiling the power of artificial intelligence for image-based diagnosis and treatment in endodontics: An ally or adversary? Int Endod J. 2025;58(2):155–70.
39. Asgary S, Aminoshariae A, Wesselink PR. Apical periodontitis in vital and nonvital teeth: clinical and radiographic features. Iran Endodontic J. 2024;19(3):148.
40. Buckley TA, Diao JA, Srivastava CN, Brodeur PG, Rajpurkar P, Rodman A, Manrai AK. Multimodal foundation models exploit text to make medical image predictions. Nat Commun. 2026.
41. Atakır K, Işın K, Taş A, Önder H. Diagnostic accuracy and consistency of ChatGPT-4o in radiology: influence of image, clinical data, and answer options on performance. Diagn Interventional Radiol. 2026;32(5):555.
42. Mahajan A, Obermeyer Z, Daneshjou R, Lester J, Powell D. Cognitive bias in clinical large language models. npj Digit Med. 2025;8(1):428.
43. Schmidgall S, Harris C, Essien I, Olshvang D, Rahman T, Kim JW, et al. Evaluation and mitigation of cognitive biases in medical language models. NPJ Digit Med. 2024;7(1):295.
44. de Almeida PRZ, Barbosa IA, Alves MSA, de Menezes SAF, de Almeida Rodrigues P, Elgrably IS, et al. Evaluating large language models performance in endodontics: A clinical experimental study. Res

Soc Dev. 2026;15(1):e2515150523–e.
45. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023;29(8):1930–40.
46. Asgary S. Artificial intelligence in endodontics: a scoping review. Iran endodontic J. 2024;19(2):85.
47. Karkehabadi H, Khoshbin E, Ghasemi N, Mahavi A, Mohammad-Rahimi H, Sadr S. Deep learning for determining the difficulty of endodontic treatment: a pilot study. BMC Oral Health. 2024;24(1):574.
48. Khanagar SB, Alfadley A, Alfouzan K, Awawdeh M, Alaqla A, Jamleh A. Developments and performance of artificial intelligence models designed for application in endodontics: a systematic review. Diagnostics. 2023;13(3):414.

Figures

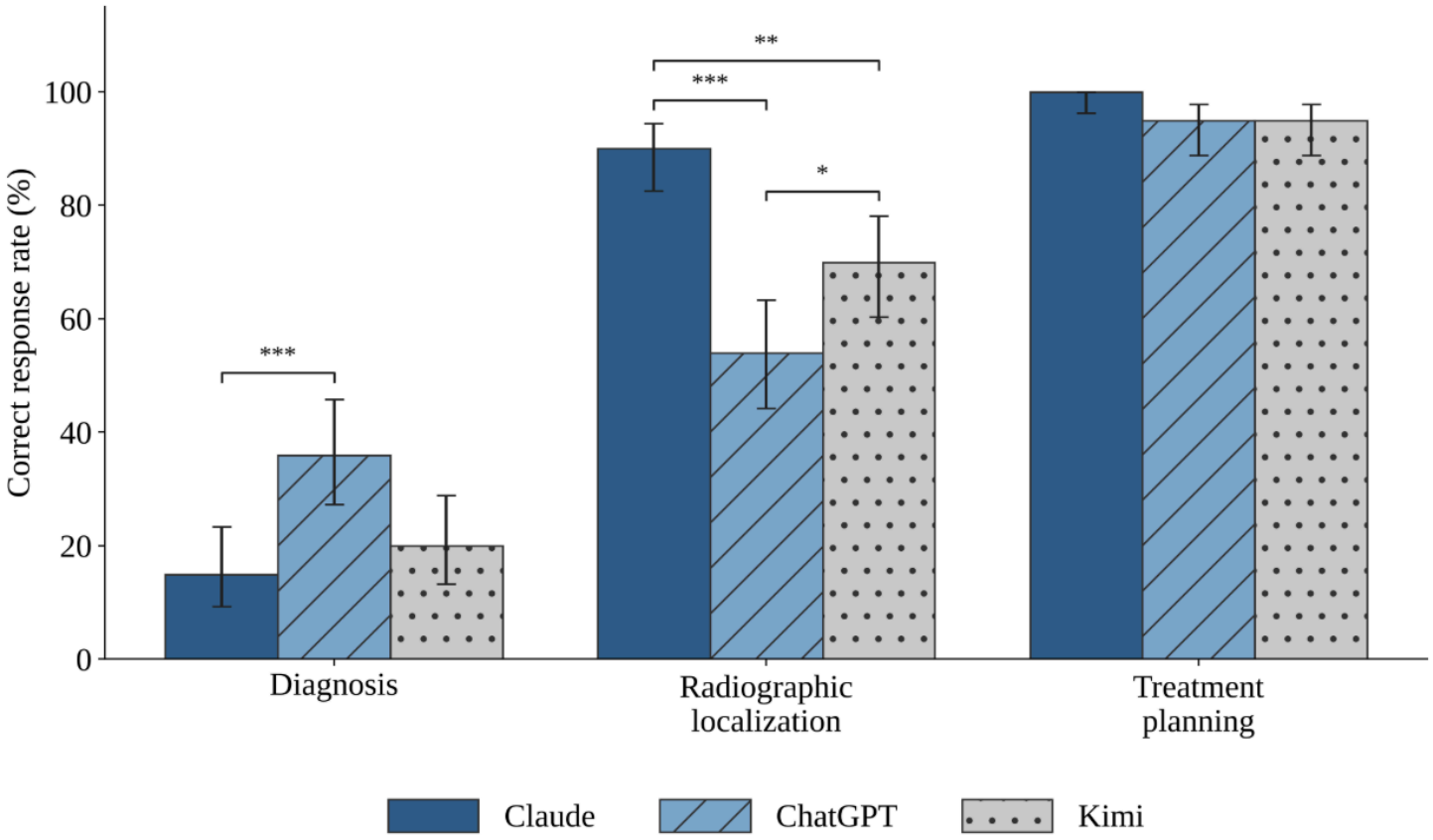

Figure 1

*Error bars represent 95% confidence intervals calculated using the Wilson score method. Exact McNemar test with Bonferroni correction: *p < 0.0167; **p < 0.01; ***p < 0.001.*

Correct response rates of the models for diagnosis, radiographic localization and treatment planning.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Table1.docx
- Table2.docx
- Table3.docx