ARTICLE

Brazilian Dental Journal. Volume 36 (2025), e25-6861
http://dx.doi.org/10.1590/0103-644020256861

ISSN 0103-6440

Can Large Language Models Detect Periapical Lesions in Anterior Teeth? A Comparative Study

Diulia Pereira Bubna [1], Natanael Henrique Ribeiro Mattos [1], Luana Beatriz das Portas Luiz [1], Flares Baratto-Filho [1,2], May Tostes de Mattos-Calil [3], Yara Teresinha Corrêa Silva-Sousa [3], Erika Calvano Küchler [4], Ângela Graciela Deliga Schroder [1], Cristiano Miranda de Araujo [1], Bianca Marques de Mattos de Araujo [1,2]

[1] Tuiuti University of Paraná, Curitiba, PR, Brazil.

[2] Department of Dentistry, University of Joinville (Univille), Joinville, SC, Brazil.

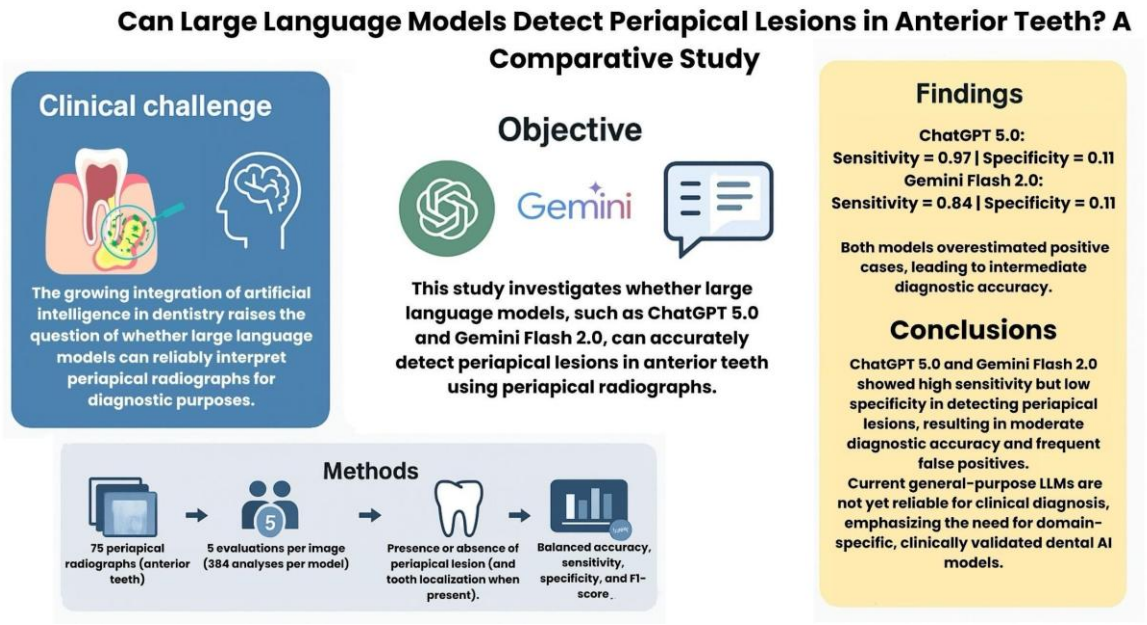
[3] School of Dentistry, University of Ribeirão Preto, Ribeirão Preto, SP, Brazil.

[4] Department of Orthodontics, University Hospital Bonn, Medical Faculty, Welschnonnenstr. 17, 53111, Bonn, Germany.

Correspondence: Flares Baratto Filho. Tuiuti University of Paraná. R. Padre Ladislau Kula, 395 – Santo Inácio. Curitiba, Brazil. Email: fbaratto1@gmail.com

This study evaluated the diagnostic performance of large language models (LLMs)—ChatGPT 5.0 (OpenAI) and Gemini Flash 2.0 (Google Inc.)—in detecting periapical lesions on periapical radiographs using a standardized multimodal prompt. Seventy-five anonymized periapical radiographs of anterior teeth from the maxilla and mandible were analyzed, evenly distributed between cases with and without lesions. A calibrated endodontic specialist provided the reference diagnosis. Each image was independently assessed five times by both LLMs using the prompt: "Does this image show a periapical lesion? Answer 'Yes' or 'No'. If 'Yes', which tooth?". Balanced accuracy, sensitivity, specificity, and F1-score were calculated with 95% confidence intervals (CIs) obtained via bootstrap resampling. Performance was also stratified by diagnostic difficulty, and the models were compared using the exact McNemar test ($\alpha = 0.05$). ChatGPT 5.0 showed higher overall performance than Gemini Flash 2.0, with sensitivity of 0.97 (95% CI: 0.95–0.99), specificity of 0.11 (95% CI: 0.07–0.16), and balanced accuracy of 0.54 (95% CI: 0.52–0.57). Gemini Flash 2.0 achieved sensitivity of 0.84 (95% CI: 0.79–0.89), specificity of 0.11 (95% CI: 0.07–0.16), and balanced accuracy of 0.48 (95% CI: 0.44–0.51). Both models showed high false-positive rates and frequent errors in tooth localization. The McNemar test confirmed a significant difference between models ($p < 0.05$), favoring ChatGPT 5.0. Both LLMs demonstrated high sensitivity but poor specificity, resulting in intermediate diagnostic performance and a bias toward positive classifications. General-purpose LLMs are therefore not yet suitable for radiographic diagnostic use.

Key Words: Artificial intelligence; Large language models; Periapical lesion; Diagnostic accuracy; Endodontics.



Introduction

Periapical lesions are chronic inflammatory alterations localized in the tissues surrounding the root apex, usually resulting from infections originating within the root canal system. They are characterized by resorption and destruction of the adjacent alveolar bone, representing one of the most clinically relevant pathologies in Endodontics (1, 2). Their prevalence in the population ranges from 23% to 31% and they are most often diagnosed through radiographic examinations (1-3). Radiographically, apical periodontitis may appear as a subtle widening of the periodontal ligament space or as more extensive radiolucent lesions associated with the tooth apex (2, 3). However, image assessment can be challenging, as the overlap of anatomical structures in two-dimensional radiographs hinders the detection of early lesions, potentially leading to delayed or inaccurate diagnoses (1, 3).

Among the available imaging methods, cone-beam computed tomography (CBCT) offers greater sensitivity for detecting periapical lesions. Nevertheless, its high cost and increased radiation dose limit its routine use (4). Consequently, periapical radiographs (PRs) remain the most widely used examination in endodontic practice (1, 3). They provide detailed information about the tooth, periodontal ligament, and alveolar bone, while being low-cost and widely accessible (1, 2). Despite these advantages, their interpretation heavily depends on the clinician's experience, which can introduce subjective bias and visual fatigue during extended working periods (2, 3). As a result, diagnostic accuracy may vary among professionals, highlighting the need for complementary tools that support standardization and reduce errors during radiographic image analysis (2, 3).

In this context, the incorporation of artificial intelligence (AI) into dentistry has expanded the possibilities for diagnostic enhancement and clinical support. Artificial intelligence has evolved from basic algorithmic systems developed in the 1950s to contemporary deep learning models capable of emulating human cognitive processes (5, 6). Deep learning, particularly through convolutional neural networks, has demonstrated remarkable performance in medical tasks such as tumor detection and radiographic image analysis, achieving results comparable to or even surpassing those of human specialists in certain contexts (2, 3). More recently, the development of large language models (LLMs), such as ChatGPT (OpenAI) and Gemini (Google), has broadened AI applications by enabling high-level processing, generation, and interpretation of natural language (5, 6). Beyond answering clinical questions and synthesizing scientific information, their multimodal versions have also gained the ability to interpret images, generating interest in their potential use for radiographic analysis. In dentistry, research on the use of LLMs as educational and diagnostic support tools is steadily increasing (6, 7). However, the literature still highlights important limitations, as these models remain susceptible to producing inaccurate information and lack robust validation for complex clinical tasks (5, 7).

Given the high prevalence of periapical lesions and the growing integration of artificial intelligence in dentistry, it is relevant to assess whether LLMs—specifically ChatGPT 5.0 and Gemini Flash 2.0—show potential to assist in detecting such alterations in periapical radiographs of anterior teeth. Therefore, the present study aimed to evaluate the diagnostic performance of LLMs in identifying periapical lesions on dental periapical radiographs of anterior teeth.

Materials and methods

Study Design

The present study employed a comparative and exploratory design aimed at evaluating the performance of LLMs—GPT-5o (OpenAI) and Gemini 2.5 Flash (Google Inc.)—in detecting periapical lesions on periapical radiographs of anterior teeth. The research was approved by the Research Ethics Committee under protocol no. 6.202.403 and conducted in accordance with the ethical principles of the Declaration of Helsinki. Administrative authorization was obtained to access the anonymized data, and informed consent was waived since no patient identification was involved. All data were anonymized prior to access, which took place between July and September 2025.

Data Source and Sample Selection

Digital periapical radiographs were retrieved from the database of a dental imaging service located in southern Brazil. Images were randomly selected, prioritizing anterior teeth.

Periapical radiographs of adult patients of both sexes with fully developed roots were included. Radiographs with poor image quality, structural superimposition, distortion, or anomalies that could compromise evaluation of the periapical region were excluded.

Image Classification

The periapical radiographs included in the study were categorized according to three main criteria:

1. Binary detection: presence or absence of periapical lesion.
2. Localization: identification of the affected tooth when a lesion was present.
3. Diagnostic difficulty: level of difficulty assigned by the specialist to determine the diagnosis, classified as 1 = easy, 2 = intermediate, and 3 = difficult.

This categorization was performed by an endodontic specialist with over ten years of clinical experience, who individually evaluated each image according to the predefined criteria and recorded the corresponding tooth whenever an alteration was observed. To ensure reliability, all images were re-evaluated after a two-week interval to determine intra-examiner agreement using Cohen's Kappa coefficient, which yielded a value greater than 0.90, indicating excellent reproducibility.

Sample Size

To determine the required number of tests per language model, a sample size calculation was performed under the assumption of an infinite population, a 5% margin of error, and an estimated accuracy rate of 50%. This conservative estimate was adopted to produce the largest possible sample size, resulting in 384 tests per model. Consequently, 75 periapical radiographs of anterior teeth from either the maxilla or the mandible were selected according to the eligibility criteria, evenly distributed between cases with and without periapical lesions. Each image was submitted five times to each model in independent sessions to assess response consistency, totaling 384 analyses per model and 768 analyses overall.

Prompt and LLM Design

Two publicly accessible multimodal LLMs capable of processing both text and images were tested: GPT-5o (OpenAI) and Gemini 2.5 Flash (Google Inc.). Both models were accessed in their most recent versions and evaluated independently under identical testing conditions.

Prompts were written in clear, objective English to ensure consistent instructions across models. Each test consisted of presenting a periapical radiograph accompanied by the following standardized instruction:

"Does this image show a periapical lesion? Answer 'Yes' or 'No'. If 'Yes', which tooth?"

This formulation was designed to assess the models' visual interpretation capability exclusively, without providing additional clinical context or prior examples, thereby ensuring the neutrality and reproducibility of the evaluation.

The models were tested without any parameter adjustments or contextual input, maintaining their baseline (zero-shot) behavior to simulate real-world clinical use by a general user. This strategy enabled the assessment of each LLM's intrinsic diagnostic performance without the influence of fine-tuning, advanced prompt-engineering, or supervised training, ensuring direct comparability between their outputs.

Evaluation Procedure

For each LLM, the following steps were carried out:

1. Access to the model platform: GPT-5o (https://chat.openai.com/) and Gemini 2.5 Flash (https://gemini.google.com/).

2. Individual submission of periapical radiographs, one at a time, along with the standardized prompt.
3. Recording of the responses generated, indicating whether a periapical lesion was identified (yes/no) and specifying the affected tooth when applicable.
4. Repetition of the procedure for each image in five independent sessions.
5. Storage and coding of all responses.

Model Performance Analysis

The results produced by each model were compared with the reference classification established by the endodontic specialist, enabling the calculation of the following diagnostic performance metrics:

• Balanced Accuracy — mean proportion of correctly identified positive and negative cases, calculated as the average between sensitivity and specificity.

• Sensitivity (Recall) — ability of the model to correctly identify positive cases, i.e., radiographs with periapical lesions.

• Specificity — ability of the model to correctly recognize negative cases, i.e., radiographs without periapical lesions.

• F1-score — harmonic mean of precision and sensitivity, providing a balanced measure of overall model performance.

In addition to individual metrics, confusion matrices were generated to visualize the distribution of true and false positives and negatives, allowing for a detailed examination of prediction patterns. Matrices were produced both globally (all cases) and stratified by diagnostic difficulty level (1 = easy; 2 = intermediate; 3 = difficult).

To illustrate the models' behavior across difficulty levels, metric profiles were constructed based on the bootstrapped mean values of Accuracy, Sensitivity, and Specificity, accompanied by semi-transparent bands representing the 95% confidence interval (95% CI).

For the direct comparison between ChatGPT and Gemini, the exact McNemar's test was applied, adopting a significance level of $\alpha = 0.05$. All analyses were conducted in Python, using the libraries pandas, numpy, scikit-learn, and statsmodels. All performance indicators were reported with their 95% confidence intervals, estimated by bootstrapping with 2000 resamples.

Results

Global Performance of the Models

In the overall analysis, ChatGPT 5.0 showed high sensitivity (0.97; 95% CI: 0.95–0.99), indicating a strong ability to identify radiographs with periapical lesions, but low specificity (0.11; 95% CI: 0.07–0.16), revealing a tendency to overestimate positive cases. This pattern resulted in intermediate accuracy (0.52; 95% CI: 0.48–0.57). Gemini Flash 2.0 exhibited a similar behavior, with lower sensitivity (0.84; 95% CI: 0.79–0.89), equally low specificity (0.11; 95% CI: 0.07–0.16), and overall accuracy of 0.46 (95% CI: 0.42–0.52) (Table 1 and Figure 1).

Among the radiographs correctly classified as containing a lesion, a tooth localization error was observed in 73.4% of ChatGPT's responses and 78.1% of Gemini's responses, indicating an additional limitation in anatomical correspondence.

Table 1. Global performance metrics of the LLMs.

Model	ChatGPT (n = 384)	Gemini (n = 384)
Accuracy (95% CI)	0.54 (0.52-0.57)	0.48 (0.44-0.51)
Sensitivity (95% CI)	0.97 (0.95-0.99)	0.84 (0.79-0.89)
Specificity (95% CI)	0.11 (0.07-0.16)	0.11 (0.07-0.16)
Precision (95% CI)	0.50 (0.45-0.55)	0.47 (0.42-0.52)
F1-Score (95% CI)	0.66 (0.62-0.71)	0.60 (0.55-0.65)

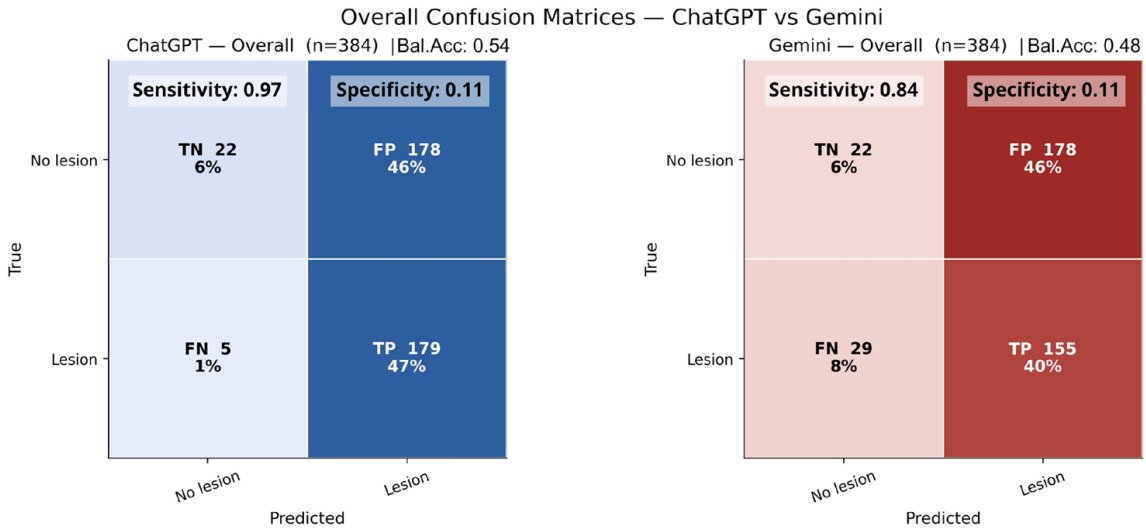


Figure 1. Overall confusion matrices for ChatGPT and Gemini in periapical lesion detection.

Performance by level of difficulty

The pattern of high sensitivity and low specificity persisted across all diagnostic difficulty levels, resulting in consistently low balanced accuracy. At level 1 (easy cases), ChatGPT 5.0 achieved a balanced accuracy of 0.54 (95% CI: 0.52–0.57), with sensitivity of 1.00 (95% CI: 1.00–1.00) and specificity of 0.09 (95% CI: 0.05–0.14), whereas Gemini Flash 2.0 showed a balanced accuracy of 0.48 (95% CI: 0.43–0.54), sensitivity of 0.85 (95% CI: 0.75–0.93), and specificity of 0.12 (95% CI: 0.07–0.17) (Table 2).

At the intermediate level (level 2), a slight increase in balanced accuracy was observed, although still within a range indicative of limited performance. ChatGPT 5.0 presented a balanced accuracy of 0.59 (95% CI: 0.54–0.65), sensitivity of 0.99 (95% CI: 0.97–1.00), and specificity of 0.20 (95% CI: 0.09–0.32) (Figure 2). Gemini Flash 2.0 reached a balanced accuracy of 0.47 (95% CI: 0.42–0.53), sensitivity of 0.85 (95% CI: 0.78–0.92), and specificity of 0.09 (95% CI: 0.02–0.18).

In the most complex cases (level 3), a further reduction in balanced accuracy was observed, with maintenance of high sensitivity and null specificity. ChatGPT 5.0 showed a balanced accuracy of 0.43 (95% CI: 0.37–0.48), sensitivity of 0.87 (95% CI: 0.73–0.97), and specificity of 0.00 (95% CI: 0.00–0.00), while Gemini Flash 2.0 achieved a balanced accuracy of 0.40 (95% CI: 0.33–0.47), sensitivity of 0.80 (95% CI: 0.66–0.93), and specificity of 0.00 (95% CI: 0.00–0.00).

These findings indicate that, regardless of diagnostic difficulty, both models displayed a systematic bias toward positive responses, which compromised their discriminative ability between radiographs with and without periapical lesions (Figure 3).

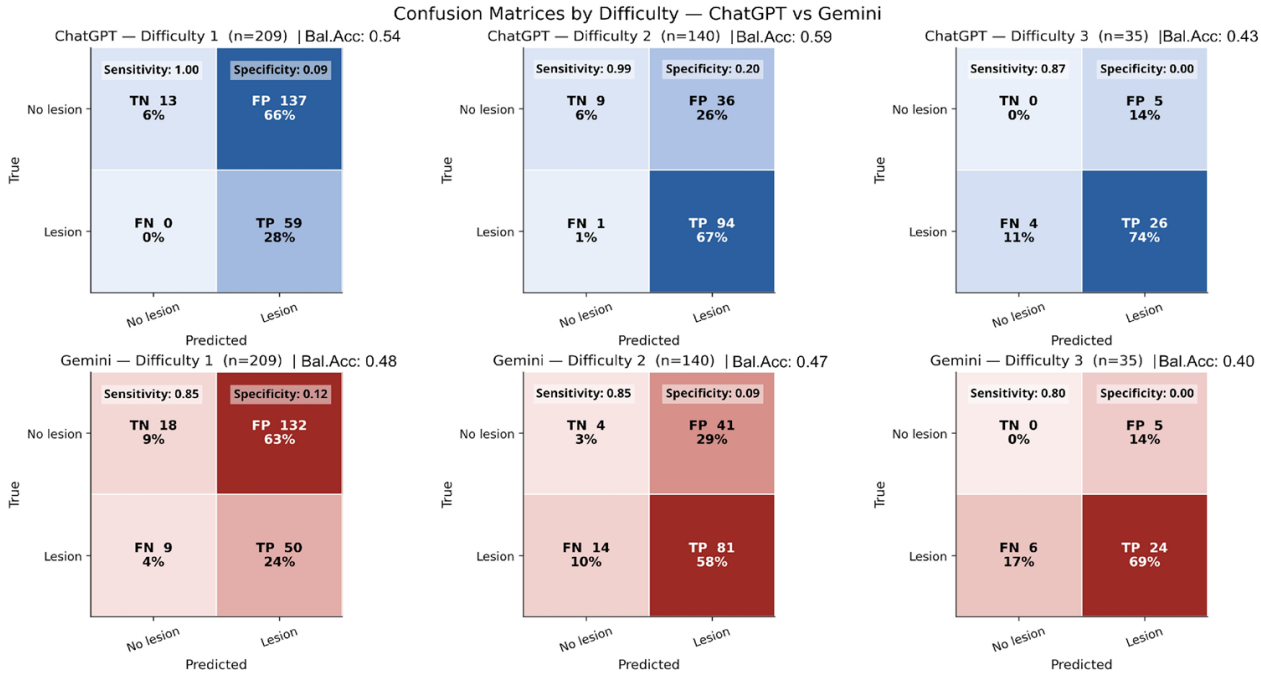


Figure 2. Confusion matrices by diagnostic difficulty for ChatGPT and Gemini in periapical lesion detection.

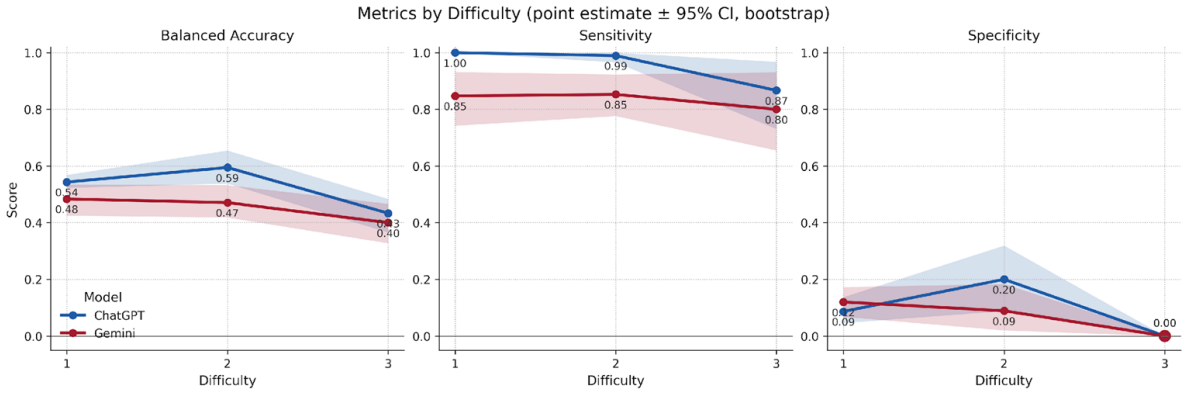


Figure 3. Diagnostic performance metrics by difficulty level for ChatGPT and Gemini.

Table 2. Performance metrics of the large language models (LLMs) by diagnostic difficulty level.

Model	Difficulty	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	Precision (95% CI)	F1-Score (95% CI)
ChatGPT	Easy (n = 210)	0.54 (0.52-0.57)	1.00 (1.00-1.00)	0.09 (0.05-0.14)	0.30 (0.24-0.37)	0.47 (0.39-0.54)
Gemini		0.48 (0.43-0.54)	0.85 (0.75-0.93)	0.12 (0.07-0.17)	0.28 (0.21-0.35)	0.42 (0.34-0.50)
ChatGPT	Intermediate (n = 140)	0.59 (0.54-0.65)	0.99 (0.97-1.00)	0.20 (0.09-0.32)	0.72 (0.64-0.80)	0.83 (0.78-0.88)
Gemini		0.47 (0.42-0.53)	0.85 (0.78-0.92)	0.09 (0.02-0.18)	0.67 (0.58-0.75)	0.75 (0.68-0.81)
ChatGPT	Difficult (n = 35)	0.43 (0.37-0.48)	0.87 (0.73-0.97)	0.00 (0.00-0.00)	0.84 (0.70-0.97)	0.85 (0.75-0.94)
Gemini		0.40 (0.33-0.47)	0.80 (0.66-0.93)	0.00 (0.00-0.00)	0.83 (0.68-0.96)	0.81 (0.70-0.91)

Comparison Between Models

In the direct comparison between models, the exact McNemar test revealed a statistically significant difference ($p < 0.05$), confirming the overall superior performance of ChatGPT 5.0 compared to Gemini Flash 2.0. This difference reflects ChatGPT's greater stability in detecting positive cases and a slight advantage in scenarios of intermediate complexity, although both models exhibited similar limitations regarding specificity.

Discussion

The growing use of LLMs for self-diagnosis reflects a significant shift in user behavior at the interface between technology and health. Expectations of performance and a positive perception of the risk–benefit balance foster user confidence in these tools, even in the absence of medical validation. This trend reinforces the acceptance of artificial intelligence as a legitimate source of clinical information while raising ethical and safety concerns in healthcare (8). In this context, the present study aimed to assess the diagnostic performance of large language models in identifying periapical lesions on dental radiographs. The results showed that although ChatGPT 5.0 and Gemini Flash 2.0 achieved high sensitivity in detecting lesions, both exhibited low specificity, with difficulty recognizing normal images. This pattern indicates a tendency toward overinterpretation of positive findings, limiting the autonomous clinical applicability of these tools.

In general, language models are prone to producing inaccurate or misleading responses. This limitation is associated with the absence of embedded medical specialization and the lack of evaluation frameworks designed for clinical use, compounded by the quality issues of training data (7). In a comparative evaluation including GPT-4, Qwen2.5, DeepSeek R1, and EndoQ, the domain-specific EndoQ model outperformed the others in clinical accuracy, contextual relevance, knowledge coverage, decision-making professionalism, and linguistic fluency, achieving high precision and recall in the information retrieval stage (7). These findings suggest that domain specialization and rigorous data curation are key determinants of clinical performance, explaining the inability of generalist models to fully meet demanding diagnostic objectives.

Adjustments in prompt formulation and optimized variants may improve performance, but do not address structural limitations. In an educational setting involving text-based and image-based questions, optimized versions and precise instructions improved accuracy depending on the task type and topic, without consistent gains across all contents (9). Similarly, providing specific contextual information increased ChatGPT-4's accuracy in topics such as restorations and obturation, but did not yield consistent improvements in other areas (10). These findings indicate that LLM performance depends more on task nature than on prompt refinement, which partly explains the high sensitivity but low specificity observed in this study.

Direct comparisons between LLMs have shown a recurrent advantage of newer ChatGPT versions over Gemini, particularly in applied tasks. Nevertheless, the overall performance of both

remains below that of human specialists. In assessments combining theoretical and clinical components, such as the Turkish dental specialization examination, ChatGPT-4 outperformed Gemini Advanced in several areas but showed lower performance specifically in Endodontics and Orthodontics (6). In a standardized radiological evaluation, ChatGPT-4o also achieved higher overall accuracy than Gemini Advanced, performing better in image-based tasks, though inconsistently across subfields(11). In the diagnosis of potentially malignant oral lesions, ChatGPT-4o demonstrated only moderate to substantial agreement with specialists (12). Overall, these findings indicate that despite technical progress and greater stability in recent ChatGPT versions, the models still exhibit important diagnostic limitations, particularly in tasks that require reliable differentiation between normal and pathological patterns.

Image interpretation in dentistry poses intrinsic technical challenges, especially in two-dimensional radiographs, where overlapping anatomical structures and the absence of depth hinder the detection of early lesions. To overcome these limitations, specific computer-vision architectures have been developed, including two-stage object detectors such as Faster R-CNN and single-stage detectors like YOLOv4, as well as deep convolutional neural networks designed for image classification and segmentation. These systems can identify patterns and contours with greater precision, even under noise and contrast variation. Studies have shown that such models can achieve performance comparable to that of specialists in detecting periodontal bone loss and identifying periapical lesions, in both cone-beam computed tomography and periapical radiographs, provided they are trained with properly annotated datasets (1, 13). In diagnostic support designs, incorporating models such as ConvNeXt and ResNet34 increased accuracy and reduced the evaluation time of novice dentists, with a statistically significant advantage of ConvNeXt (2). These results contrast with the profile of generalist language models, which, although effective in text comprehension and generation, are not designed for the clinical visual perception required in radiographic interpretation.

The high frequency of false-positive results observed in both LLMs has relevant implications for clinical practice. When used by dentists as decision-support tools, such overestimation of periapical lesions could lead to misinterpretation of radiographs, prompting unnecessary follow-up examinations or even overtreatment. This behavior reflects a lack of contextual anatomical understanding, as the models tend to classify any apical radiolucency as a lesion. Beyond diagnostic inaccuracies, the integration of unvalidated AI systems into professional workflows poses additional risks, including the propagation of systematic biases, reduced critical engagement by clinicians, and potential ethical or legal consequences if decisions are based on unverified outputs. Therefore, although their high sensitivity may help flag potentially altered cases for closer inspection, their elevated false-positive rate and associated risks emphasize that current general-purpose LLMs should only be used under professional supervision and not as autonomous diagnostic tools in dentistry.

Some limitations of the present study should be noted. CBCT imaging provides greater accuracy in detecting apical periodontitis; however, its use is recommended only as a supplementary diagnostic method—and not routinely—when a diagnosis cannot be reliably established through conventional means such as periapical radiography, due to its higher radiation dose (4). Nevertheless, many endodontists still routinely rely on periapical radiographs in daily clinical practice because of their accessibility, low cost, and diagnostic value in most cases. Although periapical lesions can affect any tooth, they occur more frequently in maxillary anterior teeth, mainly due to greater susceptibility to trauma, caries, and anatomical variations that complicate endodontic treatment (14). The results can therefore be extrapolated only to periapical lesions located in the anterior region of the maxilla and mandible, as the sample included exclusively incisor and canine teeth. The choice of anterior teeth was based on their more predictable anatomy and reduced structural overlap, which facilitated standardized analysis and minimized interpretive variability across models. Another limitation concerns the evaluation of only two LLM platforms, which restricts the generalizability of the findings. Despite these constraints, this study provides evidence that the evaluated LLMs still lack adequate diagnostic capability to accurately differentiate normal from pathological radiographs, highlighting the need for domain-specific models trained for dental radiographic interpretation.

Conclusion

ChatGPT 5.0 and Gemini Flash 2.0 models demonstrated high sensitivity in detecting periapical lesions on periapical radiographs but low specificity across all levels of difficulty, resulting in intermediate overall performance and a tendency toward false positives. These findings reinforce that general-purpose language models are not yet suitable for diagnostic applications, showing limited usefulness in directly supporting radiographic interpretation. Safe clinical use of artificial intelligence in this context requires the development of domain-specific models trained on large, curated dental datasets and validated under real clinical conditions to achieve a balance between sensitivity and specificity.

Conflict of Interest

The authors declare that they have no conflicts of interest related to this study.

Generative AI statement in scientific writing

During the preparation of this work, the authors used ChatGPT (OpenAI) for grammatical correction and language refinement. After using this tool, the authors reviewed and edited the content as needed and took full responsibility for the final content of the published article.

Funding

No funding was received or utilized to conduct this research.

Data availability

The research data are available within the article.

Resumo

Este estudo avaliou o desempenho diagnóstico de modelos de linguagem de grande porte (LLMs) — ChatGPT 5.0 (OpenAI) e Gemini Flash 2.0 (Google Inc.) — na detecção de lesões periapicais em radiografias periapicais, utilizando um prompt multimodal padronizado. Foram analisadas 75 radiografias periapicais anonimizadas de dentes anteriores da maxila e da mandíbula, distribuídas igualmente entre casos com e sem lesões. Um especialista em Endodontia, devidamente calibrado, forneceu o diagnóstico de referência. Cada imagem foi avaliada de forma independente cinco vezes por ambos os LLMs, utilizando o prompt: “Esta imagem apresenta uma lesão periapical? Em caso afirmativo, em qual dente? Responda com: A) Sim B) Não.” A acurácia balanceada, sensibilidade, especificidade e F1-score foram calculados com intervalos de confiança de 95% (IC95%) obtidos por reamostragem bootstrap. O desempenho também foi estratificado conforme o nível de dificuldade diagnóstica, e os modelos foram comparados pelo teste exato de McNemar ($\alpha = 0,05$). O ChatGPT 5.0 apresentou desempenho geral superior ao Gemini Flash 2.0, com sensibilidade de 0,97 (IC95%: 0,95–0,99), especificidade de 0,11 (IC95%: 0,07–0,16) e acurácia balanceada de 0,54 (IC95%: 0,52–0,57). O Gemini Flash 2.0 obteve sensibilidade de 0,84 (IC95%: 0,79–0,89), especificidade de 0,11 (IC95%: 0,07–0,16) e acurácia balanceada de 0,48 (IC95%: 0,44–0,51). Ambos os modelos apresentaram altas taxas de falso-positivos e erros frequentes na localização do dente. O teste de McNemar confirmou diferença estatisticamente significativa entre os modelos ($p < 0,05$), favorecendo o ChatGPT 5.0. Ambos os LLMs demonstraram alta sensibilidade, porém baixa especificidade, resultando em desempenho diagnóstico intermediário e viés para classificações positivas. Assim, modelos de linguagem de uso geral ainda não são adequados para aplicação diagnóstica radiográfica.

References

1. Viet DH, Son LH, Tuyen DN, Tuan TM, Thang NP, Ngoc VTN. Comparing the accuracy of two machine learning models in detection and classification of periapical lesions using periapical radiographs. Oral Radiol. 2024;40(4):493-500.

2. Liu J, Jin C, Wang X, Pan K, Li Z, Yi X, et al. A comparative analysis of deep learning models for assisting in the diagnosis of periapical lesions in periapical radiographs. BMC Oral Health. 2025;25(1):801.
3. Basso Á, Salas F, Hernández M, Fernández A, Sierra A, Jiménez C. Machine learning and deep learning models for the diagnosis of apical periodontitis: a scoping review. Clin Oral Investig. 2024;28(11):600.
4. Hilmi A, Patel S, Mirza K, Galicia JC. Efficacy of imaging techniques for the diagnosis of apical periodontitis: A systematic review. International endodontic journal. 2023;56 Suppl 3:326-39.
5. Özbay Y, Erdoğan D, Dinçer GA. Evaluation of the performance of large language models in clinical decision-making in endodontics. BMC Oral Health. 2025;25(1):648.
6. 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.
7. Xu X, Liu S, Zhu L, Long Y, Zeng Y, Lu X, et al. Development and evaluation of a retrieval-augmented large language model framework for enhancing endodontic education. Int J Med Inform. 2025;203:106006.
8. Shahsavar Y, Choudhury A. User Intentions to Use ChatGPT for Self-Diagnosis and Health-Related Purposes: Cross-sectional Survey Study. JMIR Hum Factors. 2023;10:e47564.
9. Wu YH, Tso KY, Chiang CP. Performance of ChatGPT in answering the oral pathology questions of various types or subjects from Taiwan National Dental Licensing Examinations. J Dent Sci. 2025;20(3):1709-15.
10. Lafourcade C, Kérourédan O, Ballester B, Richert R. Accuracy, consistency, and contextual understanding of large language models in restorative dentistry and endodontics. J Dent. 2025;157:105764.
11. Huang KA, Choudhary HK, Hardin WM, Prakash N. Comparative Analysis of ChatGPT-4o and Gemini Advanced Performance on Diagnostic Radiology In-Training Exams. Cureus. 2025;17(3):e80874.
12. 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-e31.
13. Karobari MI, Adil AH, Basheer SN, Murugesan S, Savadamoorthi KS, Mustafa M, et al. Evaluation of the Diagnostic and Prognostic Accuracy of Artificial Intelligence in Endodontic Dentistry: A Comprehensive Review of Literature. Comput Math Methods Med. 2023;2023:7049360.
14. Vera J, Thepris-Charaf J, Hernández-Ramírez A, García JG, Romero M, Vazquez-Carcaño M, et al. Prevalence of pulp canal obliteration and periapical pathology in human anterior teeth: A three-dimensional analysis based on CBCT scans. Australian Endodontic Journal. 2023;49(2):351-7.

Received: 09/10/2025
Accepted: 24/10/2025