Submitted: 11 March, 2026 Accepted: 14 April, 2026 Published: 03 September, 2026 DOI:10.22514/jocpd.2026.115

ORIGINAL RESEARCH

Journal of Clinical Pediatric Dentistry

Can AI see what clinicians see? A comparative study of GPT-5 and Gemini 2.5 in radiographic evaluation of regenerative endodontic treatments in immature permanent teeth

Enes Mustafa Aşar[1,*], Murat Selim Botsali[1]

[1]Department of Pediatric Dentistry, Selcuk University, 42130 Konya, Turkey

***Correspondence**
enesmustafa.asar@selcuk.edu.tr
(Enes Mustafa Aşar)

Abstract

Background: This study evaluates the diagnostic performance and agreement of GPT-5 and Gemini 2.5 in interpreting regenerative endodontic treatments (RET) on periapical radiographs compared with an expert reference standard. **Methods**: This retrospective diagnostic accuracy study included 51 paired maxillary anterior periapical radiographs (initial and follow-up) from 51 RET cases. Two experienced pediatric dentists established the reference standard for each fully visible tooth (n = 282) in terms of Fédération Dentaire Internationale (FDI) tooth number, RET presence/absence, apex status, periapical lesion presence/absence, and root development type (I–VI). GPT-5 and Gemini 2.5 were queried in separate reset sessions using the same standardized prompt and fixed output template. Diagnostic performance and agreement with the reference standard were assessed using standard classification metrics and agreement coefficients. **Results**: Both models classified FDI tooth numbers with a high degree of accuracy; Gemini 2.5 showed almost perfect agreement with the reference standard (Cohen's κ = 0.949, observed agreement 96.5%) and outperformed GPT-5 (κ = 0.733) in terms of FDI numbering. With regard to RET detection, Gemini 2.5 achieved substantial agreement (κ = 0.796, $F1$ = 0.898) compared with moderate agreement in the case of GPT-5 (κ = 0.537, $F1$ = 0.768). In contrast, performance with regard to apex status (Fleiss' κ = 0.270) and periapical lesion detection was limited for both models, and root development type classification was poor, with macro-averaged $F1$ scores of 0.130 (GPT-5) and 0.073 (Gemini 2.5) and agreement near chance. **Conclusions**: Off-the-shelf multimodal large language models (LLMs) demonstrated potential as assistive tools for structured, binary radiographic tasks in RET follow-up (FDI numbering and RET detection), but were not reliable in the case of complex or ordinal endpoints (apex status, lesions, root development type).

Keywords

Diagnostic accuracy; Immature permanent teeth; Large language models; Periapical radiography; Regenerative endodontic treatment

1. Introduction

Recent advances in multimodal large language models (LLMs) have enabled the integration of visual and textual information, allowing these models to analyze medical images while simultaneously processing contextual prompts [1]. In addition, models such as ChatGPT, Gemini, and Copilot have demonstrated remarkable performance in responding to educational, advisory, and clinical questions [2].

The evolution of LLMs toward visual content evaluation has accelerated in recent years with the transition to multimodal models. These models can detect and interpret relevant structures and generate text output not only with text input, but also with image input. These capabilities have been tested in various clinical dentistry applications, such as the detection of dental pathologies using panoramic radiographs, the evaluation of impacted canines, or the detection of dental anomalies using periapical radiographs [3–5]. Today, LLMs offer significant potential not only in natural language processing, but also for decision support in clinical disciplines such as medicine, radiology, and dentistry.

Regenerative Endodontic Treatments (RET) aim to biologically regenerate the pulp-dentin complex and promote continued root development in teeth with incomplete root development and necrotic pulp [6]. In traditional apexification methods, root development ceases, and the teeth may become

brittle. RET aims to overcome these limitations by continuing root development biologically [7]. The primary objective of RET is to restore the pulp-dentin complex biologically in necrotic, immature permanent teeth and to restart physiological root development. The tooth becoming clinically asymptomatic and responding positively to pulp tests are key success criteria. Additionally, radiographic analyses play a significant role in evaluating root development and determining long-term treatment outcomes [6, 7]. The main criteria used to determine the success and effectiveness of treatment include increases in root length, thickening of the dentin walls, closure of the root apex, and radiographic healing of the periapical lesion [8]. These radiographic parameters are of great importance in demonstrating the potential advantages offered by RET compared with traditional apexification treatments. The successful performance of this treatment has been shown by various methods that allow the quantitative measurement of changes in root length and dentin wall thickness [9]. Additionally, specific measurement techniques, such as Radiographic Root Area, have been proposed in some studies to monitor treatment effectiveness in a standardized, reproducible manner [10].

In this context, periapical radiographs are the most commonly used imaging modality in pre-treatment and follow-up evaluations of teeth treated with RET. However, radiographic evaluation is observer-dependent and may show inter-rater variability. Therefore, it remains to be investigated whether LLMs can perform this radiographic analysis task and accurately detect and classify variables such as tooth number, RET presence, apex closure, root development type, and periapical lesion status. If these models can reliably interpret images within a standardized prompt framework, they may offer significant potential for automated evaluation and comparative analysis.

The present study aimed to evaluate whether two commercially available multimodal LLMs (GPT-5 and Gemini 2.5) could interpret periapical radiographs of teeth treated with RET with acceptable diagnostic accuracy when compared to an expert reference standard, without any task-specific fine-tuning.

2. Materials and methods

This study is a retrospective, diagnostic accuracy-based study designed to evaluate the ability of two different LLMs (ChatGPT-5 and Gemini 2.5) to interpret periapical radiographs of teeth treated with RET. This research was conducted with the approval of the Selcuk University Faculty of Dentistry Non-Invasive Clinical Research Ethics Committee in Türkiye (Decision No: 2025/38). The Board approved the use of archived periapical radiographs with completely concealed identity information without the need for individual patient consent. Therefore, periapical radiographs obtained directly from the faculty archive were anonymized for use in this study.

This diagnostic accuracy study was prepared in accordance with STARD-AI (Standards for Reporting Diagnostic Accuracy Studies Involving Artificial Intelligence). The completed STARD-AI checklist is provided as **Supplementary Table 1**. Two different LLMs were used in the study.

2.1 Reference standard and radiographic evaluation

Two independent pediatric dentists (EMA and MSB) with at least 8 years of clinical experience in regenerative endodontics analyzed the data by reviewing anonymized periapical images on a calibrated screen and established a reference standard. For each fully visible tooth seen in the initial and final radiographs, the reference standard was recorded separately based on the following variables:

- FDI tooth classification: 11, 12, 21, 22.
- Regenerative endodontic treatment (RET): Present/Absent.
- Apex status: Open/Closed.
- Presence of periapical lesions: Present/Absent.
- Root development types: Type I, Type II, Type III, Type IV, Type V, Type VI.

To ensure objectivity and reduce the risk of observer-related bias, all measurements were recorded independently by the evaluators, without interaction or guidance. Agreement with regard to FDI tooth classification, regenerative endodontic treatment, apex status, and the presence of periapical lesions was evaluated using Cohen's kappa coefficient, with corresponding values of 0.976, 0.962, 0.848, and 0.882, respectively. In terms of root development type, an ordinal variable, inter-rater agreement was assessed using Krippendorff's alpha with an ordinal metric, which yielded a value of 0.997. These coefficients demonstrate a very high level of consistency between evaluators for all study parameters. The evaluators were utterly blind to the LLM outputs, as the reference standard was created before the LLM outputs were generated. In cases of discrepancy between the two evaluators, a consensus-based approach was adopted: the evaluators reviewed the case together and reached a final agreed label through discussion. No third-party arbitration was required, as all discrepancies were resolved through direct consensus.

2.2 Test groups and experimental design

Sample size was determined using Buderer's method for diagnostic accuracy studies [11]. Assuming a 95% confidence level ($Z = 1.96$), target sensitivity = 0.80, specificity = 0.90, prevalence = 0.25, and an allowable accepted margin of error of ± 0.10, the largest requirement was for sensitivity (N = 246), while specificity required N = 46. Accordingly, the target sample was set at N = 246 teeth and inflated by 10% for potential losses to N = 270. The final dataset comprised 282 teeth from 51 paired (initial–final) periapical radiographs, thereby exceeding the planned minimum.

The study evaluated 102 periapical radiographs (pre- and post-treatment) from 51 different cases treated with RET using two commercial LLM-based AI models (GPT-5 and Gemini 2.5) on a tooth-by-tooth basis. Pre-treatment diagnostic radiographs and follow-up radiographs taken at least 1 year after treatment were used for each case. The 2021 Clinical Considerations guide published by the American Association of Endodontists (AAE) for RET states that 6–12 months are required for apical radiolucency to resolve [12]. For these reasons, follow-up X-rays were selected from patients with follow-up scheduled for at least one year later. The follow-

up radiographs for all cases in the study ranged from 1 to 5 years.

Tooth-level evaluation was adopted because the AI output was generated per tooth, making it the natural unit of analysis for assessing classification performance. It is acknowledged, however, that multiple teeth from the same case share the same image context, which may introduce a degree of within-case clustering. As no clustering correction was applied, the independence of observations may not be fully established, and confidence intervals (CIs) and agreement coefficients should be interpreted with this limitation in mind.

The study analyzed 282 teeth from 102 periapical radiographs that were fully visible and evaluable by both models. In each new chat session, periapical radiographs taken before and after treatment were sent to the language model along with the prompt input. The responses obtained from the teeth detected by the model were recorded. The FDI classification, RET status detection, apex status detection, and root development types after RET were recorded for the teeth detected on both radiographs, listed from left to right. A total of 282 teeth were evaluated by AI in this manner and statistically analyzed according to the reference standard of specialist physician data. The study flow plan is shown in Fig. 1.

2.3 Standardization of radiographic data

All periapical radiographs used were completely anonymized, containing no patient information or embedded metadata. All periapical radiographs were from the maxillary anterior region to ensure standardization. The images were obtained from the institution's secure PACS (Picture Archiving and Communication System) archive and were used for research purposes only. The study scanned 2100 maxillary periapical radiographs taken from patients between January 2020 and January 2024. A total of 51 radiographs that met the inclusion criteria were used in the study. All maxillary periapical radiographs were obtained using a single intraoral x-ray unit (Siray Plus; Zhuhai Siger, Zhuhai, China) under a standardized pediatric protocol. Exposure parameters for the anterior region were optimized to achieve diagnostic image quality at the lowest reasonable dose (*e.g.*, 60 kVp, 4 mA, 20 ms). A phosphor storage plate (PSP) system (DIGORA Optime, Soredex, Tuusula, Finland) with size-2 plates (31 × 41 mm) was used for all acquisitions. Images were exported from the PACS using the default profile (JPEG, 470 × 620 px, 150 dpi; default brightness/contrast), with no additional window/level, brightness/contrast, or filtering adjustments applied. For image review, all radiographs were displayed on the same calibrated monitor under consistent ambient lighting, and no display adjustments were permitted.

Since the study included both pre- and post-treatment evaluations, the radiographs were standardized and aligned parallel to one another. To this end, the initial and final periapical radiographs for the same case were adjusted to ensure equal numbers of visible teeth and placed on a standard canvas at the same resolution. Image editing software (Web version, Canva Inc., Sydney, NSW, Australia) was used exclusively for anonymization, cropping, and the standardized alignment of paired radiographs. No diagnostic image features, such as brightness, contrast, or radiographic structures, were altered during this process (Fig. 2).

For spatial orientation, an angular measurement tool (Protractor) based on a browser plug-in was used, with the median palatal suture (MPS) serving as the anatomical reference; images were rotated so that the MPS was parallel to the vertical axis, ensuring that the occlusal plane was approximately parallel to the horizontal reference. After vertical/horizontal alignment was completed, the initial and final radiographs were scaled to preserve the approximate size and arrangement characteristics of the visible teeth, and then recorded in a standard format (Fig. 2).

All rotation and scaling operations were performed using the Canva default interpolation settings. No additional JPEG recompression was intentionally applied beyond the platform's default export profile (JPEG, 470 × 620 px, 150 dpi), which was used consistently across all images. Following each alignment procedure, all radiographs were visually verified by the same operator to confirm that the crown and apex of each tooth of interest remained fully within the field of view and that no diagnostically-relevant apical structures were cropped or obscured during the standardization process.

2.4 Model and prompt

LLM models were used in the period 01–12 October 2025. GPT-5 (https://chatgpt.com/) was run via OpenAI's commercial ChatGPT web interface, and Gemini 2.5 (https://gemini.google.com/app) was run via Google Gemini's commercial web interface. A single researcher manually entered all queries in separate chat sessions. User-level parameters (temperature, top-p, seed, max_tokens) were not configurable, and platform defaults were used. The primary objective of the study was to examine the performance of accessible AI models. In this context, diagnostic accuracy was analyzed by reading periapical radiographs with a standardized prompt, without retraining the models.

ChatGPT-5: Released on 07 August 2025.

Gemini 2.5 Flash: Released on 25 September 2025.

Prompt input: The prompt input was detailed and tailored to the desired outputs. The same prompt was used in both models to ensure standardization. The prompt included orientation rules (film left = patient's right), listed only fully visible teeth (crown + apex) from left to right using FDI, and noted partially visible teeth only as "Partially visible: [FDI]" and excluded them from the analysis; it used the criteria of no continuous filling in the coronal/cervical distinct Mineral Trioxide Aggregate (MTA) radiopacity and apical 2/3; mandatory reporting of RET presence/absence, apex open/closed, and lesion presence/absence for each fully visible tooth on both initial and final films; and reported root development type [13] only if RET was present and indicated on the final film.

All radiographs were evaluated using identical prompts for both models to ensure methodological consistency. The analyses were conducted within the same evaluation period to minimize variability related to potential model updates. All prompts were written in English to ensure optimal model interpretation, as these systems are primarily trained using English-language datasets.

To limit free text and facilitate statistical comparison, the

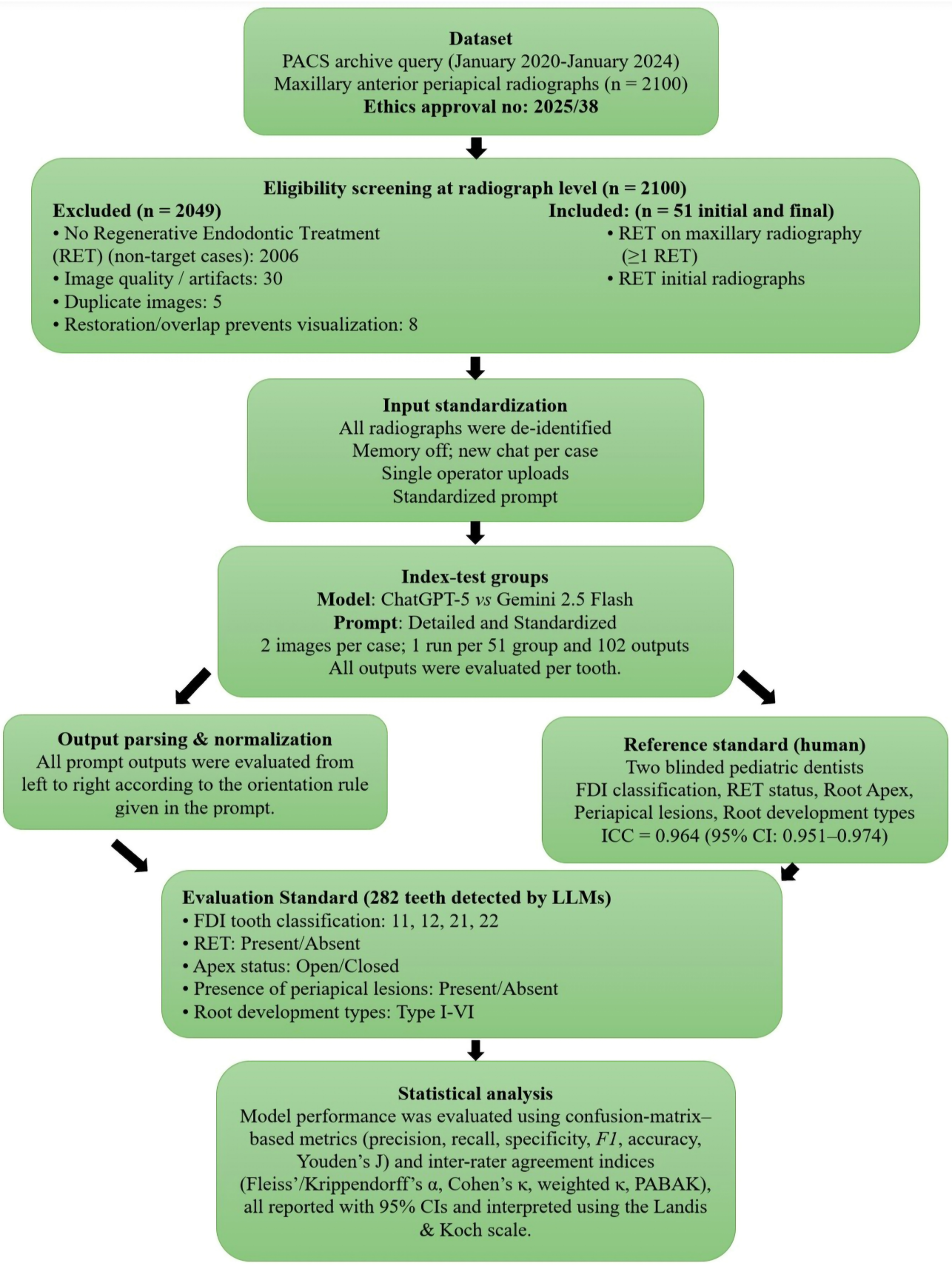


FIGURE 1. Flowchart of the study design and evaluation process (STARD-AI item 24). RET, regenerative endodontic treatment; PACS, Picture Archiving and Communication System; FDI, Fédération Dentaire Internationale; ICC, Intraclass Correlation Coefficient; CI, Confidence Interval; LLM, Large Language Model; PABAK, Prevalence-Adjusted Bias-Adjusted Kappa.

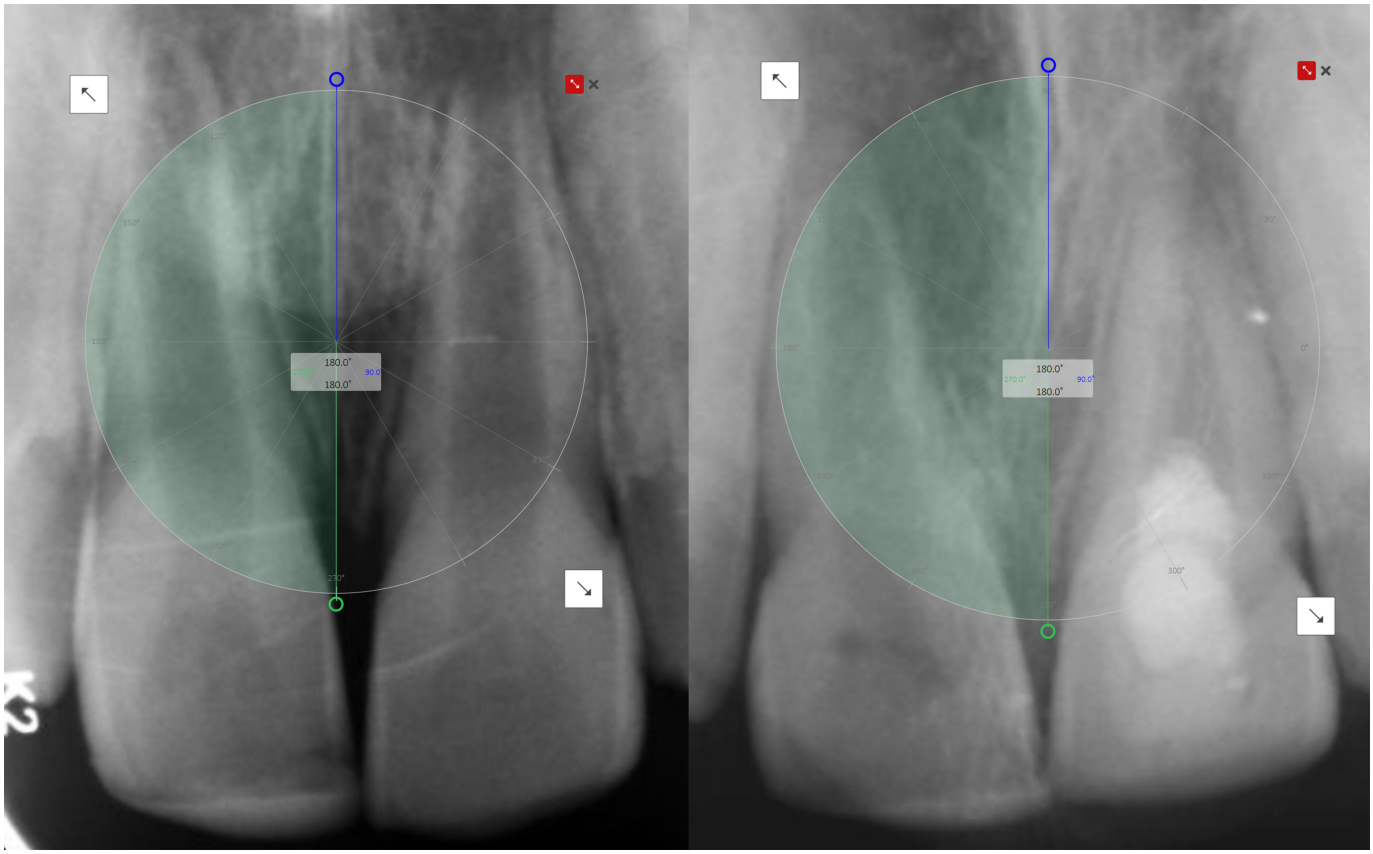


FIGURE 2. Standardization and alignment of paired periapical radiographs (initial *vs.* final radiograph).

output was linked to a fixed template: a line for each film stating "Visible teeth (count = N)… [list]" and a single-line summary for each tooth ("FDI: RET = …, Apex = …, Lesion = …, Root development type = …/No RET"). This structure ensured that both models produce standard, machine-parsable outputs that can be directly matched with the human reference standard, thereby reducing orientation and visibility errors. The complete standard prompt used in the study is provided Table 1.

2.5 Index test (ChatGPT-5 and Gemini 2.5)

Once the reference standard was fully established, data evaluation began via LLM. To ensure privacy and impartiality in all interactions with the GPT-5 and Gemini 2.5 models, the models' "memory" function was disabled. A new chat session was initiated for each radiograph in both language models. This ensured data privacy and prevented previously provided data from influencing the models. To standardize the process, all data entries were made by the same person (EMA).

In each chat session, the initial and final radiographs, along with the prompt, were sent to the model, which was then run. In total, the initial and final radiographs of 51 different cases were run in separate chat sessions using the same prompt inputs. All AI outputs were recorded in the order specified by the reference standard (FDI classification, RET status, apex status, presence of periapical lesion, root development type). All outputs were logged and evaluated left to right on the radiograph, exactly as prescribed by the prompt's orientation rule. Representative screenshots of the outputs from GPT-5 and Gemini 2.5 for the same case (Case 17) are shown in Figs. 3,4. Screenshots of the model outputs from different cases in which all four maxillary teeth were reported are also shown in Figs. 5,6.

2.6 Inclusion and exclusion criteria

Inclusion Criteria:

Cases with RET performed on at least one tooth in the maxillary anterior region.

Cases with a RET related to the maxillary anterior region and at least 1 year of follow-up radiographs.

Radiographs in which the crown and apex of the tooth that underwent RET are fully visible in the initial and final images.

Radiographs selected by two specialist dentists with consensus on diagnostic clarity and appropriateness.

Setting and dates: Selcuk University Faculty of Dentistry; January 2020–January 2024.

Exclusion Criteria:

Radiographs containing motion, exposure errors, or technical artifacts; radiographs with low resolution and difficult to interpret.

Radiographs containing overlapping anatomical structures or restorative material that prevented clear visualization of the tooth which had undergone RET.

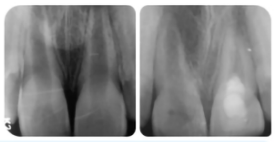
2.7 Statistical analysis

Model performance (GPT-5, Gemini 2.5) was summarized using confusion matrices and standard metrics (precision, recall/sensitivity, specificity, $F1$, accuracy; macro-averaged; Youden's J). Agreement with the human reference standard

TABLE 1. Standardized prompt structure used for GPT-5 and Gemini 2.5.

Parameter Category	Settings/Content
Orientation Rule	Film left = patient's right. Only fully visible teeth (crown + apex) listed left to right; partially visible teeth noted separately and excluded from evaluation.
RET Identification	Distinct MTA radiopacity in cervical/coronal portion; no continuous radiopaque canal filling in apical 2/3.
Reporting Rules	Apex evaluated in both films; all fully visible teeth reported; root development type recorded in final film only if RET present.
Root Development Types	Type I: wall thickening + root length increase + closed-apex. Type II: wall thickening, no length increase, blunt closed-apex. Type III: continued development, open foramen. Type IV: canal obliteration. Type V: no continued development. Type VI: segmental development.
Output Format	Per-film header + single-line summary per tooth: FDI/RET = Present/Absent/Apex = Open/Closed/Lesion = Present/Absent/Root development type.
Complete Prompt	Orientation rule: • Film left = patient's right. • Film right = patient's left. • First, determine the exact number of fully visible teeth (crown + apex within the field of view). • List only fully visible teeth from left to right on the film = patient's right to left (order: 12, 11, 21, 22 if all visible). • If partially visible teeth exist, note them separately as "Partially visible: [numbers]" but do not include them in RET/apex/root development evaluation. Regenerative Endodontic Treatment (RET) identification: • Distinct MTA radiopacity in cervical/coronal portion. • No continuous radiopaque canal filling in apical 2/3. Reporting rules: • Apex status must be evaluated in both Initial and Final films. • All fully visible teeth must be reported (even without RET). • Partially visible teeth are listed but not evaluated. • Root development type only in Final film if RET present. Root development types: • Type I: Increased thickening of the canal walls with a significant increase in root length, with the root apex becoming closed. • Type II: Increased thickening of the canal walls without a significant increase in root length, with the root apex becoming blunt and closed. • Type III: Continued root development with the apical foramen remaining open. • Type IV: Severe calcification (obliteration) of the canal space. • Type V: No continued root development. • Type VI: An uncommon type of segmental root development. Answer Format: Initial periapical radiograph Visible teeth (count = N, left → right on film = patient's right → left): [list teeth] • FDI: RET = Present/Absent; Apex = Open/Closed; Lesion = Present/Absent; Root development type: No RET Final periapical radiograph Visible teeth (count = N, left → right on film = patient's right → left): [list teeth] • FDI: RET = Present/Absent; Apex = Open/Closed; Lesion=Present/Absent; Root development type: [Type I–VI if RET present, otherwise "No RET"]

MTA, mineral trioxide aggregate; RET, regenerative endodontic treatment; FDI, Fédération Dentaire Internationale.



Orientation rule:
• Film left = patient's right.
• Film right = patient's left.
• First, determine the exact number of fully visible teeth (crown + apex within the field of view).
• List only fully visible teeth from left to right on the film = patient's right to left (order: 12, 11, 21, 22 if all visible).
• If partially visible teeth exist, note them separately as "Partially visible: [numbers]" but do not include them in RET/apex/root development evaluation.
Regenerative Endodontic Treatment (RET) identification:
• Distinct MTA radiopacity in cervical/coronal portion.
• No continuous radiopaque canal filling in apical 2/3.
Reporting rules:
• Apex status must be evaluated in both Initial and Final films.
• All fully visible teeth must be reported (even without RET).
• Partially visible teeth are listed but not evaluated.
• Root development type only in Final film if RET present.
Root development types:
• Type I: Increased thickening of the canal walls with a significant increase in root length, with the root apex becoming closed.
• Type II: Increased thickening of the canal walls without a significant increase in root length, with the root apex becoming blunt and closed.
• Type III: Continued root development with the apical foramen remaining open.
• Type IV: Severe calcification (obliteration) of the canal space.
• Type V: No continued root development.
• Type VI: An uncommon type of segmental root development.

Answer Format
Initial periapical radiograph
Visible teeth (count = N, left → right on film = patient's right → left): [list teeth]
• FDI: RET=Present/Absent; Apex=Open/Closed; Lesion=Present/Absent; Root development type: No RET
Final periapical radiograph
Visible teeth (count = N, left → right on film = patient's right → left): [list teeth]
• FDI: RET=Present/Absent; Apex=Open/Closed; Lesion=Present/Absent; Root development type: [Type I–VI if RET present, otherwise "No RET"]

Here's the structured report based on the two periapical radiographs you provided:

Initial periapical radiograph

Visible teeth (count = 2, left → right on film = patient's right → left): 11, 21

- **11**: RET = Absent; Apex = Open; Lesion = Absent; Root development type: No RET
- **21**: RET = Absent; Apex = Open; Lesion = Absent; Root development type: No RET

Partially visible: 12, 22

Final periapical radiograph

Visible teeth (count = 2, left → right on film = patient's right → left): 11, 21

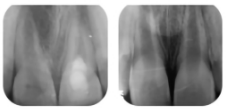
- **11**: RET = Absent; Apex = Closed; Lesion = Absent; Root development type: No RET
- **21**: RET = Present (distinct MTA radiopacity visible); Apex = Closed; Lesion = Absent; Root development type: **Type II** (thickening of canal walls, root blunt and closed without significant increase in root length)

Partially visible: 12, 22

Ask anything

ChatGPT can make mistakes. Check important info.

FIGURE 3. Representative GPT-5 output for a paired periapical radiograph set (Case 17). MTA, mineral trioxide aggregate; RET, regenerative endodontic treatment; FDI, Fédération Dentaire Internationale.



Orientation rule:
• Film left = patient's right.
• Film right = patient's left....

Analysis

Based on the provided periapical radiographs and the defined orientation and reporting rules, here is the evaluation:

Initial periapical radiograph (17ilk.jpg)

The image shows the maxillary anterior region, specifically the central and lateral incisors.

- **Fully visible teeth (count = 4, left → right on film = patient's right → left):** 12, 11, 21, 22
- **Partially visible:** None

FDI	RET	Apex	Lesion	Root development type
12	Absent	Open	Absent	No RET
11	Absent	Open	Absent	No RET
21	Absent	Open	Absent	No RET
22	Absent	Open	Absent	No RET

E-Tablolar'a aktar

Final periapical radiograph (17son.png)

The image is a follow-up of the same maxillary anterior region.

- **Fully visible teeth (count = 4, left → right on film = patient's right → left):** 12, 11, 21, 22
- **Partially visible:** None

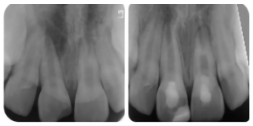
FDI	RET	Apex	Lesion	Root development type
12	Absent	Open	Absent	No RET
11	Absent	Open	Absent	No RET
21	Present	Closed	Absent	Type I
22	Absent	Open	Absent	No RET

Gemini'a sorun

Araçlar

Gemini hata yapabildiğinden verdiği yanıtları doğrulayın.

FIGURE 4. Representative Gemini 2.5 output for the same paired periapical radiograph set (Case 17). RET, regenerative endodontic treatment; FDI, Fédération Dentaire Internationale.



Orientation rule:
• Film left = patient's right.
• Film right = patient's left.
• First, determine the exact number of fully visible teeth (crown + apex within the field of view).
• List only fully visible teeth from left to right on the film = patient's right to left (order: 12, 11, 21, 22 if all visible).
• If partially visible teeth exist, note them separately as "Partially visible: [numbers]" but do not include them in RET/apex/root development evaluation.

Regenerative Endodontic Treatment (RET) identification:
• Distinct MTA radiopacity in cervical/coronal portion.
• No continuous radiopaque canal filling in apical 2/3.

Reporting rules:
• Apex status must be evaluated in both Initial and Final films.
• All fully visible teeth must be reported (even without RET).
• Partially visible teeth are listed but not evaluated.
• Root development type only in Final film if RET present.

Root development types:
• Type I: Increased thickening of the canal walls with a significant increase in root length, with the root apex becoming closed.
• Type II: Increased thickening of the canal walls without a significant increase in root length, with the root apex becoming blunt and closed.
• Type III: Continued root development with the apical foramen remaining open.
• Type IV: Severe calcification (obliteration) of the canal space.
• Type V: No continued root development.
• Type VI: An uncommon type of segmental root development.

Answer Format

Initial periapical radiograph
Visible teeth (count = N, left → right on film = patient's right → left): [list teeth]
• FDI: RET=Present/Absent; Apex=Open/Closed; Lesion=Present/Absent; Root development type: No RET

Final periapical radiograph
Visible teeth (count = N, left → right on film = patient's right → left): [list teeth]
• FDI: RET=Present/Absent; Apex=Open/Closed; Lesion=Present/Absent; Root development type: [Type I–VI if RET present, otherwise "No RET"]

Here's the structured radiographic interpretation following your specified format:

Initial periapical radiograph

Visible teeth (count = 4, left → right on film = patient's right → left): 12, 11, 21, 22

- **12**: RET = Absent; Apex = Open; Lesion = Absent; Root development type: No RET
- **11**: RET = Absent; Apex = Open; Lesion = Absent; Root development type: No RET
- **21**: RET = Absent; Apex = Open; Lesion = Absent; Root development type: No RET
- **22**: RET = Absent; Apex = Open; Lesion = Absent; Root development type: No RET

Final periapical radiograph

Visible teeth (count = 4, left → right on film = patient's right → left): 12, 11, 21, 22

- **12**: RET = Absent; Apex = Closed; Lesion = Absent; Root development type: No RET
- **11**: RET = Present; Apex = Closed; Lesion = Absent; Root development type: **Type I**
- **21**: RET = Present; Apex = Closed; Lesion = Abse… Root development type: **Type I**
- **22**: RET = Absent; Apex = Closed; Lesion = Absent; Root development type: No RET

Ask anything

ChatGPT can make mistakes. Check important info.

FIGURE 5. Examples of GPT-5 outputs from cases with all four maxillary anterior teeth fully visible (11, 12, 21, 22). MTA, mineral trioxide aggregate; RET, regenerative endodontic treatment; FDI, Fédération Dentaire Internationale.

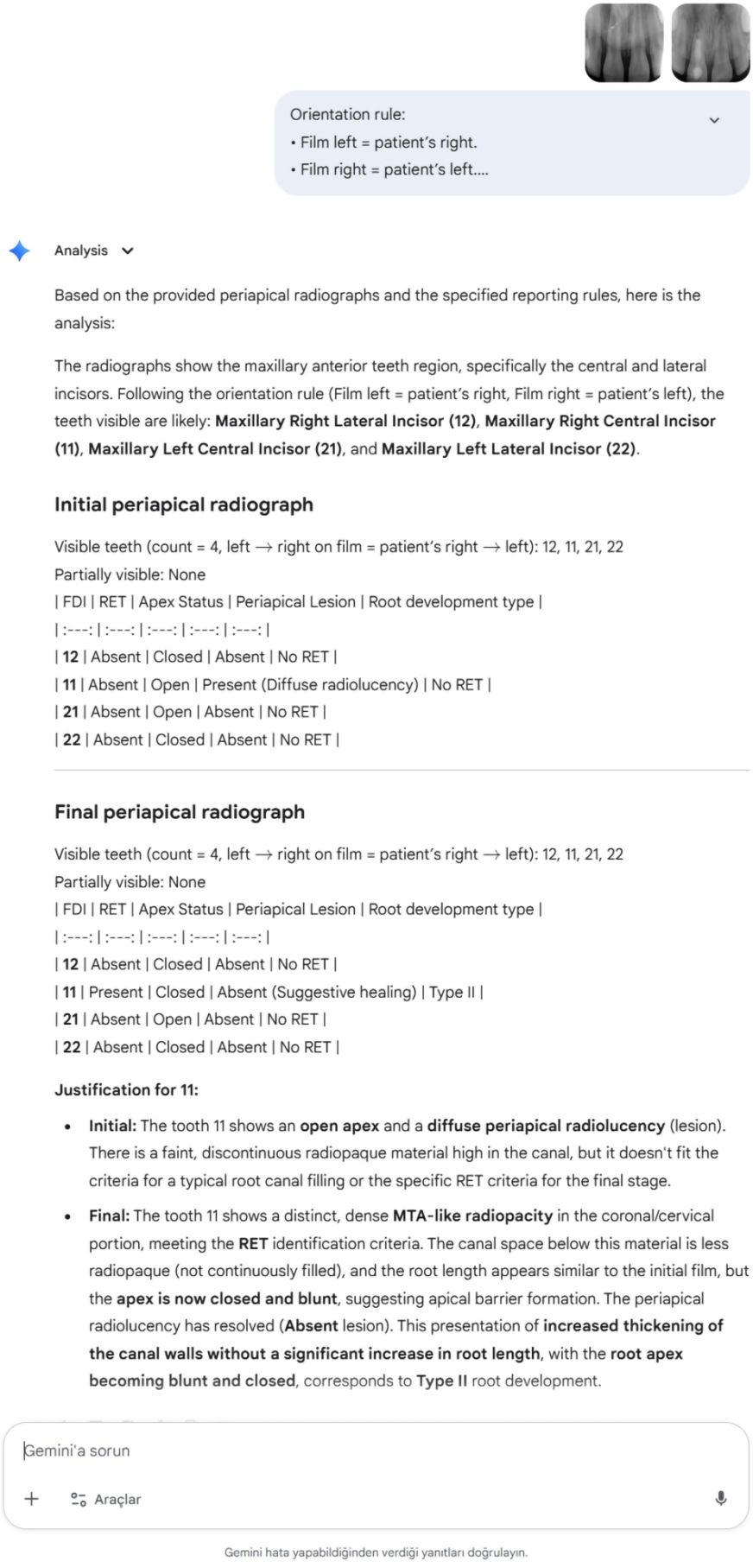


FIGURE 6. Examples of Gemini 2.5 outputs from cases with all four maxillary anterior teeth fully visible (11, 12, 21, 22). MTA, mineral trioxide aggregate; RET, regenerative endodontic treatment; FDI, Fédération Dentaire Internationale.

used Fleiss' kappa (nominal) and Krippendorff's alpha (ordinal); pairwise agreement used Cohen's kappa and quadratic-weighted kappa. Results are presented with 95% CIs and observed agreement; where prevalence/bias influenced κ ($|P_o - \kappa| > 0.25$), prevalence-adjusted bias-adjusted kappa (PABAK)/Brennan–Prediger were additionally reported. Interpretation of agreement coefficients followed the Landis and Koch scale: <0.00, poor; 0.00–0.20, slight; 0.21–0.40, fair; 0.41–0.60, moderate; 0.61–0.80, substantial; 0.81–1.00, almost perfect agreement.

3. Results

A total of 51 paired periapical radiographs were included in the study. Because multiple teeth were visible within many radiographic fields, each clearly identifiable tooth was evaluated individually. As a result, a total of 282 teeth were analyzed across the initial and follow-up radiographs. According to the reference standard, the prevalence rates at baseline were recorded as 21% (60 teeth) for RET, 74% (209 teeth) for open apex, and 10% (29 teeth) for lesion presence. Root development types among teeth with RET were recorded as type 1, 10% (6 teeth), type 2, 21.7% (13 teeth), type 3, 21.7% (13 teeth), type 4, 3.3% (2 teeth), type 5, 41.7% (25 teeth), and type 6, 1.6% (1 tooth).

3.1 FDI tooth classification (11/12/21/22)

The GPT-5 model classified FDI tooth numbers as follows: numbers 11, 12, 21, and 22 had correct classification rates of 74.5%, 100%, 74.5%, and 94.1%, respectively. Teeth numbers 11 and 21 were found to be the teeth that caused the most confusion for the model.

GPT-5 achieved its best per-class performance for FDI 22 ($P = 1.000$, $S = 1.000$, $F1 = 0.970$, $A = 0.993$). FDI 12 showed perfect sensitivity ($R = 1.000$), whereas the lowest accuracy occurred in the case of FDI 11 ($A = 0.816$). Accuracies for FDI 12 and 21 were similar (0.908 and 0.901). Misclassifications were concentrated mainly between 11 and 21.

In classifying the FDI tooth numbers of the Gemini 2.5 model, the correct classification rates for teeth numbers 11, 12, 21, and 22 are 96.1%, 100%, 96.1%, and 94.1%, respectively. The model misclassified 3.9% of teeth numbers 11 and 21, and 5.9% of tooth number 22. Gemini 2.5 achieved its strongest per-class performance for FDI 22 ($P = 1.000$, $S = 1.000$, $F1 = 0.970$, $A = 0.993$), with perfect sensitivity for FDI 12 ($R = 1.000$). The macro-average accuracy ($A = 0.982$) indicates very high overall classification performance. The performance metrics of the GPT-5 and Gemini 2.5 models are presented in Table 2.

At the macro-level averages, the Gemini 2.5 model achieved higher values than GPT-5 with regard to all performance metrics in the FDI classification. The differences in precision ($\Delta P = -0.127$), $F1$ score ($\Delta F1 = -0.131$), and Youden index ($\Delta J = -0.161$), are particularly noticeable. These findings indicate that Gemini 2.5 demonstrates a more balanced and robust classification performance in terms of both capturing true positives and preventing false positives.

There was substantial three-rater agreement for FDI classification across the human reference standard, GPT-5, and Gemini 2.5 (Fleiss' $\kappa = 0.807$; 95% CI: 0.762–0.853; observed agreement $P_o = 86.3\%$; $Z = 38.149$; $p < 0.001$). Pairwise agreement was almost perfect between the reference standard and Gemini 2.5 (Cohen's $\kappa = 0.949$; 95% CI: 0.919–0.980; $P_o = 96.5\%$), substantial between the reference standard and GPT-5 ($\kappa = 0.733$; 95% CI: 0.667–0.797; $P_o = 80.8\%$), and substantial between GPT-5 and Gemini 2.5 ($\kappa = 0.742$; 95% CI: 0.678–0.804; $P_o = 81.6\%$).

3.2 Regenerative endodontic treatment (RET)

The performance of GPT-5 was higher for RET absent than present ($P = 0.922$ *vs.* 0.579; $F1 = 0.888$ *vs.* 0.647). Sensitivity was greater for absent ($R = 0.856$) than present ($R = 0.733$), while specificity showed a complementary pattern ($S = 0.733$ *vs.* 0.856). Macro-averages: $P = 0.751$, $R = 0.795$, $S = 0.795$, $F1 = 0.768$ (Table 2).

The performance of Gemini 2.5 was high overall and stronger for absent than present ($P = 0.968$ *vs.* 0.803; $F1 = 0.954$ *vs.* 0.841); sensitivity remained high in both classes ($R = 0.941$ *vs.* 0.883), with complementary specificity ($S = 0.883$ *vs.* 0.941). Macro-averages: $P = 0.886$, $R = 0.912$, $S = 0.912$, $F1 = 0.898$ (Table 2).

At the macro average level, the Gemini 2.5 model achieved higher values than GPT-5 in predicting the presence/absence of RET across all performance metrics. In particular, the differences in the Youden index ($\Delta J = -0.234$), precision ($\Delta P = -0.213$), $F1$ score ($\Delta F1 = -0.196$), and specificity ($\Delta S = -0.192$), were significant. These findings show that Gemini 2.5 exhibits a more balanced and robust classification performance in terms of both capturing true positives and preventing false positives.

There was significant and substantial three-rater agreement in terms of the human reference standard, GPT-5, and Gemini 2.5 (Fleiss' $\kappa = 0.610$; 95% CI: 0.526–0.698; observed agreement $P_o = 85.8\%$; $Z = 17.737$; $p < 0.001$). The prevalence-insensitive Brennan–Prediger coefficient was 0.716, also indicating substantial agreement. Pairwise agreement was substantial between the human reference standard and Gemini 2.5 (Cohen's $\kappa = 0.796$; 95% CI: 0.707–0.883; $P_o = 92.9\%$), and moderate between the human reference standard and GPT-5 and between GPT-5 and Gemini 2.5 ($\kappa = 0.537$; 95% CI: 0.418–0.653; $P_o = 83.0\%$ and $\kappa = 0.511$; 95% CI: 0.399–0.626; $P_o = 81.6\%$, respectively). After adjusting for prevalence/bias, PABAK values were 0.660 (reference standard—GPT-5) and 0.632 (GPT-5—Gemini 2.5), both indicating substantial agreement.

3.3 Apex status (open *vs.* closed)

Sensitivity was lower for open-apex teeth compared with closed-apex teeth in the GPT-5 model ($R = 0.526$ *vs.* 0.808); precision was higher for open-apex teeth ($P = 0.887$ *vs.* 0.373). The $F1$ values were 0.661 for open-apex teeth and 0.511 for closed-apex teeth. The macro average values were found to be $P = 0.630$, $R = 0.667$, $S = 0.667$, $F1 = 0.586$; overall accuracy $A = 0.599$ (Table 2).

Sensitivity was found to be higher in open-apex teeth than

TABLE 2. Per-tooth diagnostic performance of GPT-5 and Gemini 2.5—FDI tooth-class identification, RET detection, apex status, and apical lesion.

Model and Category	True Positive (TP)	False Positive (FP)	False Negative (FN)	True Negative (TN)	Precision (*P*)	Sensitivity (*R*)	Specificity (*S*)	*F*1 Score	Accuracy (*A*)
				FDI					
GPT-5									
11	76	26	26	154	0.745	0.745	0.856	0.745	0.816
12	44	26	0	212	0.629	1.000	0.891	0.772	0.908
21	76	2	26	178	0.974	0.745	0.989	0.844	0.901
22	32	0	2	248	1.000	0.941	1.000	0.970	0.993
Macro-average					0.837	0.858	0.934	0.833	0.904
Gemini 2.5									
11	98	4	4	176	0.961	0.961	0.978	0.961	0.972
12	44	4	0	234	0.917	1.000	0.983	0.957	0.986
21	98	2	4	178	0.980	0.961	0.989	0.970	0.979
22	32	0	2	248	1.000	0.941	1.000	0.970	0.993
Macro-average					0.964	0.966	0.987	0.964	0.982
				RET (Absent/Present)					
GPT									
Absent	190	16	32	44	0.922	0.856	0.733	0.888	
Present	44	32	16	190	0.579	0.733	0.856	0.647	0.830
Macro-average					0.751	0.795	0.795	0.768	
Gemini									
Absent	209	7	13	53	0.968	0.941	0.883	0.954	
Present	53	13	7	209	0.803	0.883	0.941	0.841	0.929
Macro-average					0.886	0.912	0.912	0.898	
				Root Apex (Open/Closed)					
GPT									
Open	110	14	99	59	0.887	0.526	0.808	0.661	
Closed	59	99	14	110	0.373	0.808	0.526	0.511	0.599
Macro-average					0.630	0.667	0.667	0.586	
Gemini									
Open	139	30	70	43	0.822	0.665	0.589	0.735	
Closed	43	70	30	139	0.381	0.589	0.665	0.462	0.645
Macro-average					0.602	0.627	0.627	0.599	
				Apical Lesion (Absent/Present)					
GPT									
Absent	231	13	22	16	0.947	0.913	0.552	0.930	
Present	16	22	13	231	0.421	0.552	0.913	0.478	0.876
Macro-average					0.684	0.732	0.732	0.704	
Gemini									
Absent	246	26	7	3	0.904	0.972	0.103	0.937	
Present	3	7	26	246	0.300	0.103	0.972	0.154	0.883
Macro-average					0.602	0.538	0.538	0.545	

P = TP/(TP + FP); R = TP/(TP + FN); S = TN/(TN + FP); F1 = 2·(P·R)/(P + R); A = (TP + TN)/(TP + FP + FN + TN).
FDI, Fédération Dentaire Internationale; RET, regenerative endodontic treatment.

in closed-apex teeth (R = 0.665 *vs.* 0.589). Precision was significantly higher in favor of open-apex teeth (P = 0.822 *vs.* 0.381). Specificity, however, was higher in closed-apex teeth (S = 0.665 *vs.* 0.589). $F1$ values were 0.735 in open-apex teeth and 0.462 in closed-apex teeth. The macro averages were found to be P = 0.602, R = 0.627, S = 0.627, $F1$ = 0.599, and overall accuracy A = 0.645 (Table 2).

Overall, the GPT-5 model showed higher values than Gemini 2.5 in terms of precision (ΔP = 0.028), sensitivity (ΔR = 0.040), specificity (ΔS = 0.040), and Youden index (ΔJ = 0.080). However, the $F1$ score ($\Delta F1$ = −0.013) and accuracy (ΔA = −0.046) values showed a slight superiority in favor of Gemini 2.5. The difference in the Youden index indicates that GPT-5 has a higher diagnostic discrimination capacity than Gemini 2.5 in distinguishing between open and closed apex cases.

There was significant, yet fair, three-rater agreement for apex status among the human reference standard, GPT-5, and Gemini 2.5 (Fleiss' κ = 0.270; 95% CI: 0.196–0.344; observed agreement P_o = 64.8%; Z = 7.855; $p < 0.001$). Pairwise agreement was fair for human reference standard—Gemini 2.5 (κ = 0.216; 95% CI: 0.083–0.313; P_o = 64.5%) and human reference standard—GPT-5 (κ = 0.243; 95% CI: 0.135–0.351; P_o = 59.9%), and moderate for GPT-5–Gemini 2.5 (κ = 0.411; 95% CI: 0.306–0.516; P_o = 69.9%). Prevalence-adjusted coefficients (PABAK) were 0.290, 0.198, and 0.398, respectively, corroborating these levels of agreement.

3.4 Apical lesion (absent/present)

The performance of GPT-5 was significantly higher in those teeth without apical lesions compared with those with apical lesions (P = 0.947 *vs.* 0.421; R = 0.913 *vs.* 0.552; $F1$ = 0.930 *vs.* 0.478). Macro averages: P = 0.684, R = 0.732, S = 0.732, $F1$ = 0.704; overall accuracy A = 0.876 (Table 2).

The performance of Gemini 2.5 was significantly stronger in those teeth without apical lesions compared with those with apical lesions (P = 0.904 *vs.* 0.300, R = 0.972 *vs.* 0.103, $F1$ = 0.937 *vs.* 0.154). Specificity showed a complementary pattern (S = 0.972 *vs.* 0.103). Macro averages: P = 0.602, R = 0.538, S = 0.538, $F1$ = 0.545; overall accuracy A = 0.883 (Table 2).

GPT-5 exceeded Gemini 2.5 across key metrics: ΔP = 0.082, ΔR = 0.194, ΔS = 0.194. The Youden's J difference was also substantial (ΔJ = 0.388), indicating superior diagnostic discrimination by GPT-5, independent of prevalence.

In terms of lesion presence, three-rater agreement was significant, yet fair (Fleiss' κ = 0.271; 95% CI 0.161–0.380; P_o = 87.9%; Z = 7.892; $p < 0.001$). The prevalence/bias-adjusted coefficient was PABAK = 0.759, indicating substantial underlying agreement. Pairwise: human reference standard—Gemini 2.5 κ = 0.107 (P_o = 88.3%; PABAK = 0.766), human reference standard—GPT-5 κ = 0.409 (P_o = 87.6%; PABAK = 0.752), and GPT-5–Gemini 2.5 κ = 0.250 (P_o = 87.9%; PABAK = 0.758). These relatively high observed agreements ($P_o \approx 88\%$), coupled with fair κ values, reflect prevalence/bias effects; the prevalence-adjusted coefficients (PABAK) indicate substantial underlying agreement.

3.5 Root development types (type 1–6)

Classification at the root development type level remained limited for both models. Sensitivity and $F1$ were low in most root development types; specificity, however, was high due to the sparse distribution of root development types. GPT-5 achieved P = 0.258, R = 0.165, S = 0.866, $F1$ = 0.130, A = 0.771 in terms of macro averages, while Gemini 2.5 achieved P = 0.089, R = 0.152, S = 0.859, $F1$ = 0.073, A = 0.753; the differences were particularly in favor of GPT-5 for precision (ΔP = +0.169) and $F1$ ($\Delta F1$ = +0.057) (Youden J was low in both models: 0.031 *vs.* 0.011).

Overall agreement was poor/slight: Krippendorff's α = −0.134 (95% CI, −0.294 to 0.037), indicating no agreement beyond chance. Pairwise weighted Cohen's κ values were likewise near zero with CIs crossing zero—Human reference standard vs Gemini 2.5: κ = 0.051 (95% CI, −0.011 to 0.112; P_o = 13.3%), Human reference standard *vs.* GPT-5: κ = 0.054 (95% CI, −0.049 to 0.168; P_o = 20.0%), and GPT-5 *vs.* Gemini 2.5: κ = 0.015 (95% CI, −0.200 to 0.210; P_o = 21.7%). These findings are consistent with low prevalence and class imbalance in the type labels. Detailed confusion matrices and all metrics are presented in Table 3 and Fig. 7.

4. Discussion

This study evaluated the diagnostic performance of two commercially-available multimodal LLMs—GPT-5 and Gemini 2.5—in interpreting periapical radiographs of teeth treated with RET, without task-specific fine-tuning. Both models demonstrated task-dependent performance: results were clinically promising for structured binary tasks such as FDI tooth numbering and RET detection, whereas performance in terms of apex status, periapical lesion detection, and root development type classification remained insufficiently accurate for clinical use. These findings suggest that the current generation of general-purpose LLMs may serve as assistive tools for selected radiographic tasks, but are not yet suitable as standalone diagnostic instruments in RET follow-up.

The limited performance observed for complex tasks may be attributed to several factors. Dental periapical radiographs often present low contrast, overlapping anatomical structures, and subtle pathological changes that might challenge automated interpretation. Additionally, general-purpose multimodal LLMs are typically trained on large-scale natural image datasets, rather than specialized dental radiographic images, which may limit their ability to recognize fine diagnostic features specific to endodontic evaluation.

In the literature, AI-based models have been used in dentistry for topics such as dental traumatology [14], caries classification [15], oral cancer diagnosis [16], and tooth numbering [17]. In endodontic studies, the performance of deep learning models has been examined with regard to the detection of periapical lesions, the evaluation of root canal morphology, the estimation of working length, and the prediction of treatment outcomes in mature permanent teeth. In these studies, AI-based models have demonstrated promising performance [18–20]. Developing all these AI-based models requires extensive

TABLE 3. Comparison of macro-averaged performance metrics for root development types classification (GPT-5 *vs.* Gemini 2.5).

Metrics	GPT-5	Gemini 2.5	Difference (Δ = GPT-5–Gemini 2.5)
Precision (*P*)	0.258	0.089	0.169
Sensitivity (*R*)	0.165	0.152	0.013
Specificity (*S*)	0.866	0.859	0.007
*F*1 Score	0.130	0.073	0.057
Accuracy (*A*)	0.771	0.753	0.018
Youden Index (*J*)	0.031	0.011	0.020

All metrics are macro-averaged across classes. P = TP/(TP + FP); R = TP/(TP + FN); S = TN/(TN + FP); F1 = 2·(P·R)/(P + R); A = (TP + TN)/(TP + FP + FN + TN); J = R + S − 1.
TP, True Positive; FP, False Positive; FN, False Negative; TN, True Negative.

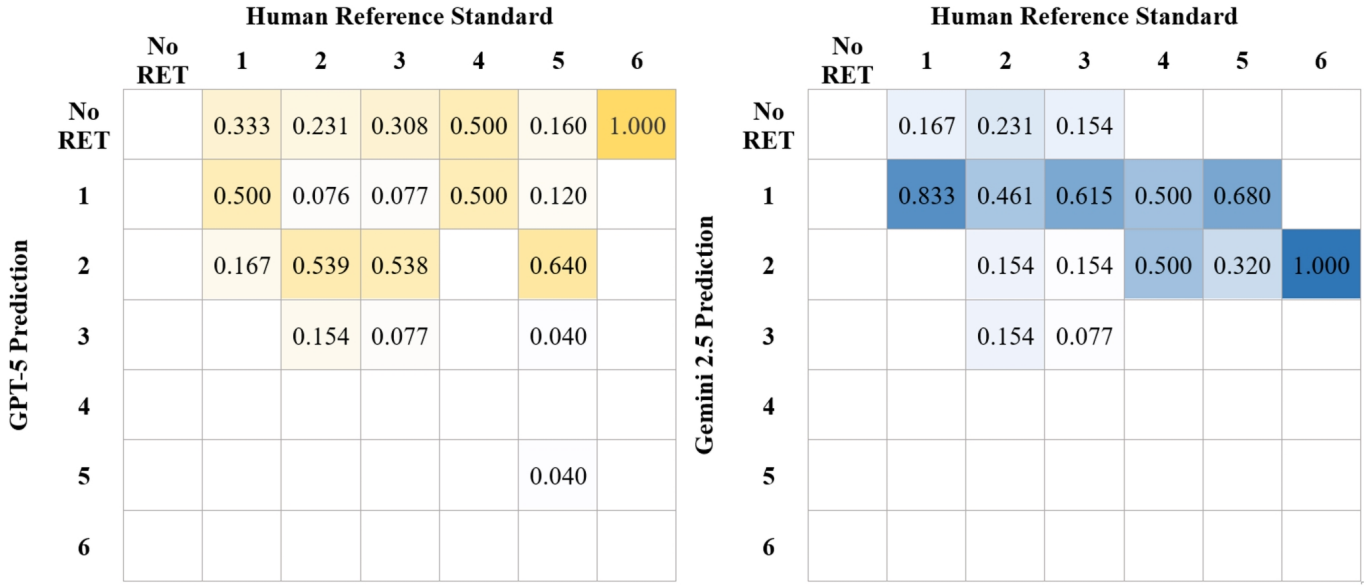


FIGURE 7. Row-normalized probability (confusion) matrices for root development type classification: GPT-5 (left) and Gemini 2.5 (right), referenced to the human standard. RET, regenerative endodontic treatment.

model training and data input. The fact that each clinical scenario requires new model training limits the generalizability of these studies to all areas of dentistry [21]. With the introduction of the use of LLMs in dentistry, increasingly practical diagnostic performance studies have emerged, indicating that they can operate without extensive training, unlike those that require detailed training processes with large datasets [22]. In these studies, model performance has been evaluated in areas such as the detection of common dental conditions [23], dental age determination [24], the localisation of impacted canines [4], the detection of supernumerary teeth [3], and the identification of complex endodontic anomalies [25].

AI applications in the field of RET remain limited. Lu *et al.* [26] demonstrated the potential of five trained machine learning models for predicting RET prognosis, while Baraka *et al.* [27] showed that a Three-dimensional U-shaped convolutional network (3D U-NET) deep learning model could accelerate pulp volume assessments following RET using Cone-beam computed tomography (CBCT) segmentation. Ekmekci *et al.* [28] reported that GPT-4o provided highly accurate responses to RET-related clinical questions. To the best of our knowledge, no prior study has directly evaluated periapical radiographs of RET-treated teeth using multimodal LLMs, positioning the present work as an initial contribution in this area.

Diagnostic studies using dental radiographs in multimodal LLMs are limited, and most have used panoramic radiographs. In these studies, LLMs have generally demonstrated limited diagnostic performance. Camlet *et al.* [29] evaluated tooth numbering and bone height estimation in panoramic radiographs. In this study, they reported that LLMs did not perform as well as convolutional neural networks. Liu *et al.* [23] reported acceptable performance using GPT-4o for radiopaque findings such as restorations and implants; however, this study also yielded quite limited results with regard to caries, periapical lesions, and fractures. Similarly, Mine *et al.* [30], Silva *et al.* [5], and Salmanpour *et al.* [4] reported limited clinical applicability for detecting various pathologies. Evidence derived from periapical radiographs is even more limited. Although task-specific fine-tuning has been shown to significantly improve

diagnostic accuracy in periapical imaging [3], LLMs have performed well below expert levels when faced with complex endodontic anomalies [25]. All of these studies demonstrate the increasing use of LLMs in the interpretation of dental radiographs. However, it has been reported that general-purpose LLMs without domain-specific adaptation have significant diagnostic limitations.

Both models achieved a high degree of accuracy in FDI tooth numbering, with Gemini 2.5 demonstrating almost perfect agreement with the human reference standard (Cohen's κ = 0.949). GPT-5 performed acceptably overall, but showed a tendency for misclassification between the maxillary central incisors (FDI 11 and 21) which are anatomically similar and share a symmetric position on the radiograph; error rates were considerably lower for the morphologically more distinct lateral incisors. These findings suggest that multimodal LLMs are nearing clinical application in structured tasks such as FDI classification. These models could be considered as auxiliary tools for automatic labeling or as support for second-read reviews. However, the distribution of central and lateral incisors in the dataset was imbalanced. This may have influenced the results. Validation with more balanced samples is required before moving to broader clinical applications.

In terms of RET presence/absence classification, GPT-5 achieved acceptable overall accuracy (≈83%, $F1 \approx 0.77$), but demonstrated a marked performance asymmetry between classes, performing considerably better in detecting RET absence than presence (precision 0.92 *vs.* 0.58). These results suggest that the model exhibits a bias towards the negative class. Therefore, the risk of false negatives increases in RET-positive cases. Consequently, this creates clinical limitations in terms of the model's testing capacity. In contrast, Gemini 2.5 can distinguish the presence and absence of RET with a higher degree of accuracy (≈93%), a more balanced $F1$ score (≈0.90), and a Youden index of 0.82. Thus, it shows a higher agreement with the human reference standard. Considering all these results, Gemini 2.5 has, on one hand, demonstrated superior performance as a clinical support tool for RET detection. GPT-5, on the other hand, provides limited clinical support due to its tendency to miss RET-positive cases. Therefore, it may be more appropriate to consider it as a secondary assessment tool for clinicians.

When evaluating the performance of both models, the results were insufficient for independent clinical use. Limited success was achieved in distinguishing between open- and closed-root apex. GPT-5 identified closed-apex with relatively higher accuracy. It also misclassified approximately half of the open-apex. Gemini 2.5 demonstrated slightly better overall accuracy and $F1$ score. GPT-5's slightly higher Youden index may indicate it has slightly greater diagnostic discrimination. While both models showed moderate agreement with each other (κ = 0.411), they demonstrated only fair agreement with the human reference standard (κ = 0.243 and κ = 0.216, respectively). In conclusion, both models, when compared with the human reference standard, exhibited similar diagnostic limitations in detecting apical morphology on periapical radiographs. Another possible reason for this limited performance could be the significant prevalence of open-apex cases.

In evaluating periapical lesion status, the GPT-5 model demonstrated relatively balanced performance despite an imbalanced data distribution. It performed significantly better in detecting the absence of lesions than in detecting their presence. In contrast, the Gemini 2.5 model evaluated cases without lesions almost flawlessly, but showed very low sensitivity with regard to teeth with lesions. As a result, it missed a significant portion of actual lesions. The fact that teeth with lesions were much less common in data distribution may have led to the results being inaccurately reflected. Based on these data, the GPT-5 demonstrated comparatively more balanced performance in terms of periapical lesion detection; however, neither model achieved a level considered adequate for independent clinical use in this domain.

Finally, both LLMs showed poor performance in terms of classifying root development types (Types I–VI). GPT-5 provided only marginally acceptable results in the early development stages (Types I–II), while sensitivity and $F1$ scores remained very low in the case of more advanced and rare types (Types III–VI); a similar, even more limited picture was observed for Gemini 2.5. The low average precision, sensitivity, and $F1$ scores, along with Youden indices close to zero, indicate that the model is insufficient at distinguishing the subtle morphological differences between root development stages. The lack of statistical significance in human–model and model–model agreement coefficients also reveal that LLMs have not yet captured human decision logic on this ordinal scale. These results also show that these models are not yet capable of evaluating apical development types in teeth that have undergone root canal treatment. However, data limitations and skewed distributions are also issues in this regard, and it may not be accurate to conclude solely from this data. This topic requires further investigation into more powerful, trainable LLMs.

A closer examination of the two models' error patterns reveals behaviorally distinct profiles. Gemini 2.5 provides more reliable responses on binary tasks such as tooth numbering and RET detection. GPT-5, while seemingly giving more balanced responses across classes, exhibited more inconsistent performance. Both models struggled with classified morphological features such as root development type and apical status. This may be attributed to the fact that the models were classified with large image sets without specific adaptation to dental radiographs. In terms of data regarding the models' performance, Gemini 2.5 does not give positive results unless certain and clear detection, while GPT-5, conversely, gives positive responses more easily. This hypothesis aligns with the high sensitivity values of GPT-5 in lesion detection. As the current dataset does not provide access to model decision probabilities or calibration data, these interpretations remain hypothetical and would warrant dedicated investigation in future work.

Behavioral differences between the models may affect how both models will function in clinical practice. The findings suggest that, on one hand, the Gemini 2.5 model can be a helpful annotation tool for clinicians in tasks such as FDI tooth numbering and RET detection, as it demonstrated high agreement with the human reference standard. GPT-5, on the other hand, can be considered a complementary tool to reduce the risk of overlooking subtle or difficult-to-interpret findings. In any case, the final diagnostic decision should

remain the responsibility of the clinician, and no model should be considered a substitute for or an independent determination of clinical judgment. The models performed poorly with regard to apex status and root development type classification. This indicates that LLMs have not demonstrated sufficient success for clinical practice in these areas. However, the development of multimodal LLMs may provide opportunities to improve performance with regard to these tasks in the future.

Prompt design has been shown to significantly affect the diagnostic performance of LLMs. According to various studies, preparing domain-specific commands generally increases accuracy [31, 32]. However, an overly-detailed or information-laden prompt can reduce accuracy [33]. The prompt used in this study was designed to provide sufficient detail without being overly long. This approach was effective in simple tasks such as orientation determination and tooth numbering. However, despite the detailed definition of root development types, the model performance was insufficient in this task. This result suggests that complex sequential categories exceed the capacity of existing general-purpose LLMs. In other words, the problem may not stem from the inadequacy of the prompt, but from the limitations of the model in making such distinctions.

The imbalance in class distribution should be taken into account when interpreting the diagnostic criteria in this study. The proportion of periapical lesion-positive cases was only 10%, while the proportion of Type VI root development was 1.6%. This imbalance may mask low sensitivity while inflating overall accuracy values.

A related interpretive consideration concerns the agreement statistics reported for tasks with highly imbalanced class distributions, particularly periapical lesion detection. In such cases, a phenomenon known as the prevalence paradox may arise, whereby high observed agreement rates coexist with paradoxically low kappa values. This occurs because Cohen's kappa corrects for chance agreement based on marginal frequencies; when one class heavily dominates, the expected chance agreement is already high, and even clinically-meaningful model performance yields a low kappa. For this reason, prevalence- and bias-adjusted kappa (PABAK) was additionally reported for tasks where the absolute difference between observed agreement and kappa exceeded 0.25, as this threshold may indicate a meaningful prevalence or bias effect. In interpreting the agreement statistics of the present study, PABAK may be considered a more informative indicator of underlying agreement for highly imbalanced outcomes, while kappa remains the appropriate primary metric for tasks with more balanced class distributions such as FDI numbering and RET detection.

This study has a number of limitations. First, the dataset is based on a limited number of RET cases treated at a single center; notably, class imbalance across outcome variables, particularly for periapical lesion presence and rare root development subtypes, may have constrained model performance and widened CIs in these categories. Secondly, task-specific fine-tuning was not performed on the GPT-5 and Gemini 2.5 models, as the study was designed to evaluate the general performance and reliability of accessible LLMs. Model-specific optimization could improve the reliability of the results. Third, the analyses are limited to standardized periapical radiographs of maxillary anterior teeth; it is uncertain whether similar results would be obtained in different tooth groups, other imaging protocols, or broader RET populations. Finally, only two commercial LLM versions, specific versions at a given point in time, were evaluated; given the rapid update cycles of these models, there is potential for both improvement and unintended deterioration in performance in future versions, necessitating validation of our findings in future independent studies. Additionally, as user-level parameters, such as temperature and sampling settings, were not configurable via the web interface, and platform defaults were used throughout, output variability cannot be fully excluded; the same input may yield different outputs across sessions, which represents an inherent reproducibility constraint of web-based LLM evaluation.

Furthermore, as multiple teeth from the same radiographic case were included in the analysis, the assumption of full statistical independence between observations may not hold; the potential clustering effect represents an additional interpretive constraint on the reported CIs and agreement statistics.

5. Conclusions

The present findings indicate that GPT-5 and Gemini 2.5 demonstrate task-dependent performance in the radiographic evaluation of RET-treated immature permanent teeth. Gemini 2.5 achieved almost perfect agreement with the human reference standard for FDI tooth numbering (κ = 0.949) and substantial agreement for RET detection (κ = 0.796), outperforming GPT-5 in terms of these tasks (κ = 0.733 and κ = 0.537, respectively), whereas performance for apex status, periapical lesion detection, and root development type classification remained insufficient for independent clinical use in both models. These results suggest that current general-purpose LLMs can be used as support tools for physicians with regard to certain simple tasks. However, at this point in time, these models should not be considered as independent diagnostic tools. The findings may pave the way for more focused model development studies in this field.

ABBREVIATIONS

RET, regenerative endodontic treatments; LLMs, large language models; AI, artificial intelligence; AAE, American Association of Endodontists; PACS, Picture Archiving and Communication System; MPS, median palatal suture; FDI, Fédération Dentaire Internationale; STARD, Standards for Reporting Diagnostic Accuracy; PSP, phosphor storage plate; MTA, mineral trioxide aggregate; CI, confidence interval; 3D, three-dimensional; U-NET, U-shaped convolutional network; CBCT, cone-beam computed tomography; PABAK, prevalence-adjusted bias-adjusted kappa.

AVAILABILITY OF DATA AND MATERIALS

The datasets generated and analysed during the current study are not publicly available due to patient privacy and institu-

tional data protection regulations; the periapical radiographic images contain sensitive clinical information and were used under ethics committee approval (Selcuk University Faculty of Dentistry Non-Invasive Clinical Research Ethics Committee, Decision No: 2025/38). However, the anonymised numerical data (AI model outputs, reference standard scores, and statistical outputs) that support the findings of this study are available from the corresponding author upon reasonable request.

AUTHOR CONTRIBUTIONS

EMA—conceptualized the manuscript; carried out data analysis, drafted and edited the manuscript. EMA and MSB—carried out methodology. Both authors subsequently revised the drafts. Both authors read and approved of the final manuscript.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

All procedures in this study were performed in accordance with the ethical standards of the Selcuk University Faculty of Dentistry Non-Invasive Clinical Research Ethics Committee in Türkiye (Approval Date: 30 June 2025, Decision No: 2025/38) and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. Due to the retrospective nature of the study, the Selcuk University Faculty of Dentistry Non-Invasive Clinical Research Ethics Committee waived the requirement for informed consent to participate and approved the use of institutional archival data without individual consent, provided that all data were anonymized prior to analysis.

ACKNOWLEDGMENT

The authors thank Hakan Benli for statistical consultation. During the preparation of this work, DeepL was used for AI-assisted English translation and Grammarly for grammar and language editing; all content was subsequently reviewed and edited by the authors, who take full responsibility for the final manuscript. The multimodal LLMs GPT-5 and Gemini 2.5 were used solely as index tests for radiographic evaluation as described in the Methods section and were not used to generate or translate any part of the manuscript text.

FUNDING

This research received no external funding.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

SUPPLEMENTARY MATERIAL

Supplementary material associated with this article can be found, in the online version, at https://oss.jocpd.com/files/article/2095390600662532096/attachment/Supplementary%20material.docx.

REFERENCES

[1] Umer F, Batool I, Naved N. Innovation and application of large language models (LLMs) in dentistry—a scoping review. BDJ Open. 2024; 10: 90.

[2] Çekiç EC, Tavşan O. Evaluating large language models using national endodontic specialty examination questions: are they ready for real-world dentistry? BMC Medical Education. 2025; 25: 1308.

[3] Aşar EM, İpek İ, Bi Lge K. Customized GPT-4V(ision) for radiographic diagnosis: can large language model detect supernumerary teeth? BMC Oral Health. 2025; 25: 756.

[4] Salmanpour F, Akpınar M. Performance of Chat Generative Pretrained Transformer-4.0 in determining labiolingual localization of maxillary impacted canine and presence of resorption in incisors through panoramic radiographs: a retrospective study. American Journal of Orthodontics and Dentofacial Orthopedics. 2025; 168: 220–231.

[5] Silva TP, Andrade-Bortoletto MFS, Ocampo TSC, Alencar-Palha C, Bornstein MM, Oliveira-Santos C, *et al*. Performance of a commercially available Generative Pre-trained Transformer (GPT) in describing radiolucent lesions in panoramic radiographs and establishing differential diagnoses. Clinical Oral Investigations. 2024; 28: 204.

[6] Wei X, Yang M, Yue L, Huang D, Zhou X, Wang X, *et al*. Expert consensus on regenerative endodontic procedures. International Journal of Oral Science. 2022; 14: 55.

[7] Meschi N, Palma PJ, Cabanillas-Balsera D. Effectiveness of revitalization in treating apical periodontitis: a systematic review and meta-analysis. International Endodontic Journal. 2023; 56: 510–532.

[8] Erdogan O, Casey SM, Bahammam A, Son M, Mora M, Park G, *et al*. Radiographic evaluation of regenerative endodontic procedures and apexification treatments with the assessment of external root resorption. Journal of Endodontics. 2024; 50: 1420–1428.e1.

[9] Alfahadi HR, Al-Nazhan S, Alkazman FH, Al-Maflehi N, Al-Nazhan N. Clinical and radiographic outcomes of regenerative endodontic treatment performed by endodontic postgraduate students: a retrospective study. Restorative Dentistry & Endodontics. 2022; 47: e24.

[10] Flake NM, Gibbs JL, Diogenes A, Hargreaves KM, Khan AA. A standardized novel method to measure radiographic root changes after endodontic therapy in immature teeth. Journal of Endodontics. 2014; 40: 46–50.

[11] Buderer NM. Statistical methodology: I. Incorporating the prevalence of disease into the sample size calculation for sensitivity and specificity. Academic Emergency Medicine. 1996; 3: 895–900.

[12] AAE clinical considerations for a regenerative procedure revised 5/18/2021. 2021. Available at: https://www.aae.org/specialty/wp-content/uploads/sites/2/2021/08/ClinicalConsiderationsApprovedByREC062921.pdf (Accessed: 01 October 2025).

[13] Jiang X, Liu H, Peng C. Continued root development of immature permanent teeth after regenerative endodontics with or without a collagen membrane: a randomized, controlled clinical trial. International Journal of Paediatric Dentistry. 2022; 32: 284–293.

[14] Bani-Hani T, Wedyan M, Al-Fodeh R, Shuqeir R, Al Jundi S, Tewari N. Artificial intelligence model for application in dental traumatology. European Archives of Paediatric Dentistry. 2025; 26: 1117–1124.

[15] Taleb A, Rohrer C, Bergner B, De Leon G, Rodrigues JA, Schwendicke F, *et al*. Self-supervised learning methods for label-efficient dental caries classification. Diagnostics. 2022; 12: 1237.

[16] Ilhan B, Guneri P, Wilder-Smith P. The contribution of artificial intelligence to reducing the diagnostic delay in oral cancer. Oral Oncology. 2021; 116: 105254.

[17] Bilgir E, Bayrakdar İŞ, Çelik Ö, Orhan K, Akkoca F, Sağlam H, *et al*. An artificial intelligence approach to automatic tooth detection and numbering in panoramic radiographs. BMC Medical Imaging. 2021; 21: 124.

[18] Albitar L, Zhao T, Huang C, Mahdian M. Artificial Intelligence (AI) for detection and localization of unobturated second mesial buccal (MB2) canals in cone-beam computed tomography (CBCT). Diagnostics. 2022; 12: 3214.

[19] Qu Y, Lin Z, Yang Z, Lin H, Huang X, Gu L. Machine learning models for prognosis prediction in endodontic microsurgery. Journal of Dentistry. 2022; 118: 103947.
[20] Sadr S, Mohammad-Rahimi H, Motamedian SR, Zahedrozegar S, Motie P, Vinayahalingam S, *et al*. Deep learning for detection of periapical radiolucent lesions: a systematic review and meta-analysis of diagnostic test accuracy. Journal of Endodontics. 2023; 49: 248–261.e3.
[21] Patil SR, Karobari MI. Exploring artificial intelligence for enhanced endodontic practice: applications, challenges, and future directions. Advances in Public Health. 2024; 2024: 8075515.
[22] Setzer FC, Li J, Khan AA. The use of artificial intelligence in endodontics. Journal of Dental Research. 2024; 103: 853–862.
[23] 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: 2315.
[24] Dursun D, Bilici Geçer R. Dental age estimation from panoramic radiographs: a comparison of orthodontist and ChatGPT-4 evaluations using the London Atlas, Nolla, and Haavikko Methods. Diagnostics. 2025; 15: 2389.
[25] Erkal D, Felek T, Butean OP, Er K. Dens invaginatus as a diagnostic challenge: evaluating large language models against expert endodontic reasoning. BMC Oral Health. 2025; 25: 1552.
[26] Lu J, Cai Q, Chen K, Kahler B, Yao J, Zhang Y, *et al*. Machine learning models for prognosis prediction in regenerative endodontic procedures. BMC Oral Health. 2025; 25: 234.
[27] Baraka M, El-Kateb N, Gamal M, Elwan AH, Sharaf P, Cevidanes L, *et al*. Assessment of volumetric changes after regenerative endodontic procedures using semiautomated and 3D U-NET automated CBCT segmentation: a retrospective cohort study. BMC Oral Health. 2025; 25: 1409.
[28] Ekmekci E, Durmazpinar PM. Evaluation of different artificial intelligence applications in responding to regenerative endodontic procedures. BMC Oral Health. 2025; 25: 53.
[29] Camlet A, Kusiak A, Ossowska A, Świetlik D. Advances in periodontal diagnostics: application of multimodal language models in visual interpretation of panoramic radiographs. Diagnostics. 2025; 15: 1851.
[30] Mine Y, Iwamoto Y, Okazaki S, Nishimura T, Tabata E, Takeda S, *et al*. Challenges and limitations of multimodal large language models in interpreting pediatric panoramic radiographs. International Journal of Paediatric Dentistry. 2026; 36: 74–80.
[31] AlFarabi Ali S, AlDehlawi H, Jazzar A, Ashi H, Esam Abuzinadah N, AlOtaibi M, *et al*. The diagnostic performance of large language models and oral medicine consultants for identifying oral lesions in text-based clinical scenarios: prospective comparative study. JMIR AI. 2025; 4: e70566.
[32] Freire Y, Santamaría Laorden A, Orejas Pérez J, Ortiz Collado I, Gómez Sánchez M, Thuissard Vasallo IJ, *et al*. Evaluating the influence of prompt formulation on the reliability and repeatability of ChatGPT in implant-supported prostheses. PLOS ONE. 2025; 20: e0323086.
[33] Rebitschek FG, Carella A, Kohlrausch-Pazin S, Zitzmann M, Steckelberg A, Wilhelm C. Evaluating evidence-based health information from generative AI using a cross-sectional study with laypeople seeking screening information. npj Digital Medicine. 2025; 8: 343.

How to cite this article: Enes Mustafa Aşar, Murat Selim Botsali. Can AI see what clinicians see? A comparative study of GPT-5 and Gemini 2.5 in radiographic evaluation of regenerative endodontic treatments in immature permanent teeth. Journal of Clinical Pediatric Dentistry. 2026; 50(5): 53-70. doi: 10.22514/jocpd.2026.115.