BMC Oral Health

RESEARCH

Open Access

Diagnostic and localization performance of multimodal large language models in the interpretation of dental periapical radiographs

Ömer Hatipoğlu[1*], Muhammed Baytar[1] and Fatma Pertek Hatipoğlu[2]

Abstract

Objective This study aimed to comparatively evaluate the performance of multiple multimodal large language models (MLLMs) in interpreting dental periapical radiographs using a structured framework that incorporates diagnostic accuracy, tooth-level anatomical precision, localization, and composite correctness across various attributes.

Materials and methods This paired, cross-sectional study assessed six contemporary MLLMs (Microsoft Copilot, ChatGPT-o4-mini-high, ChatGPT-5.2 Thinking, Gemini 2.5 Flash, Gemini 3 Pro, and Grok-3) using 55 de-identified dental periapical radiographs selected from a clinical archive according to predefined inclusion and exclusion criteria. Radiographs were selected to represent mutually exclusive tooth-status categories, including dental caries, composite restorations, amalgam restorations, and sound teeth. The reference labels for tooth status and anatomical identifiers were established before model evaluation through independent assessment and consensus by experienced clinicians. Each image was evaluated by all models under standardized prompting conditions. Performance was assessed across tasks of increasing complexity, including tooth-status diagnosis, jaw identification, tooth region, side determination, FDI tooth numbering, localization accuracy, and overall correctness across all attributes. Paired statistical analyses were performed using Cochran's Q test and generalized estimating equations with Holm's adjustment for multiple comparisons.

Results Tooth status diagnosis accuracy was moderate and comparable across models (45.5–60.0%), with no significant between-model differences. In contrast, significant heterogeneity was observed for localization-related tasks, including side determination, FDI tooth numbering, composite localization, and complete accuracy ($p < 0.05$). Adjusted analyses demonstrated that Gemini 3 Pro showed the highest adjusted accuracy for localization-dependent outcomes and overall correctness, with higher odds than several other models in pairwise comparisons. None of the models showed a consistent advantage for tooth status diagnosis alone.

Conclusion The performance of MLLMs in dental periapical radiograph interpretation is task-dependent and increasingly model-specific as anatomical precision requirements increase. While MLLMs may support education and structured discussions of radiographic findings, their reliability for autonomous tooth-level diagnosis and localization remains limited.

*Correspondence:
Ömer Hatipoğlu
omerhtp@gmail.com

Full list of author information is available at the end of the article

© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

Clinical Significance MLLMs may support dental education and patient communication by assisting with the explanation of basic radiographic features. However, their reliability decreases for tooth-level localization and comprehensive anatomical accuracy, indicating that MLLMs should not be used as stand-alone tools for clinical decision-making based on periapical radiographs.

Keywords Artificial Intelligence, Large Language Models, Periapical Radiography, Diagnostic Accuracy

Introduction

Artificial intelligence (AI) has become increasingly integrated into healthcare, with applications ranging from diagnostic support to clinical education and patient communication. In dentistry, earlier AI research has predominantly focused on task-specific image analysis systems, particularly convolutional neural network–based models designed for narrowly defined radiographic tasks. More recently, image-capable multimodal large language models (MLLMs), also referred to as vision-language models, have emerged as accessible tools that can process visual inputs together with natural-language prompts and generate structured textual responses. Unlike conventional text-only large language models (LLMs), MLLMs integrate image interpretation with language-based reasoning, making them conceptually distinct when evaluating radiographic interpretation performance rather than language generation alone [1–3].

Periapical radiographs are among the most frequently used imaging modalities in routine dental practice and education because they provide high-resolution tooth-level information with relatively low radiation exposure. These images are essential for detecting dental caries, evaluating restorative materials, and assessing periapical and root-related anatomy [4–6]. However, manual interpretation of periapical radiographs can be challenging in daily clinical practice because relevant findings may be subtle, overlap with adjacent anatomical structures, or be affected by projection geometry, image quality, and restorative materials. In busy clinical or educational settings, systematic tooth-level assessment may also be time-consuming and susceptible to observer variability, particularly when multiple teeth, restorations, or borderline radiographic changes are present. Accurate interpretation of periapical radiographs requires more than identifying the presence or absence of a pathological or restorative finding. It also requires precise anatomical localization, including the correct determination of the jaw, tooth region, laterality, and tooth number. Errors in localization may lead to inappropriate clinical decisions or miscommunication, even when the underlying diagnostic finding is correctly recognized [7, 8]. Therefore, a clinically meaningful evaluation of AI-based radiographic interpretation should incorporate not only diagnostic accuracy but also increasing levels of anatomical specificity that reflect real-world clinical and educational use [5, 9].

Despite the growing availability of MLLM-based tools in dentistry, their reliability for dental radiographic interpretation remains insufficiently validated. Most prior AI studies in dental radiology have examined task-specific image classification models, usually focusing on isolated outcomes such as caries, periapical pathology, or the presence of restorations [4–6]. In contrast, MLLMs generate free-text responses rather than fixed categorical outputs, which introduces additional challenges such as ambiguity, incomplete localization, and inconsistencies in tooth-level anatomical details [10, 11]. Moreover, relatively few studies have systematically assessed whether these models can maintain accuracy across a hierarchy of tasks, from basic image-level attributes to comprehensive tooth-level diagnosis and localization [12, 13].

Accordingly, the aim of the present study was to conduct a structured, paired evaluation of multiple contemporary MLLMs in the interpretation of dental periapical radiographs. Model performance was assessed across tasks of increasing complexity, including tooth-status diagnosis, jaw identification, tooth region, side determination, FDI tooth numbering, composite localization, and complete correctness across all attributes. By applying predefined scoring criteria, stratified analyses by reference category and tooth region, and model-based statistical comparisons, this study aimed to provide a clinically relevant benchmark of the capabilities and limitations of MLLMs in tooth-level dental radiographic interpretation.

Materials and methods

Study design

This study was designed as a paired, cross-sectional comparative performance evaluation of image-capable MLLM-generated interpretations of dental periapical radiographs. The evaluated models included Microsoft Copilot, ChatGPT-o4-mini-high (OpenAI), ChatGPT-5.2 Thinking (OpenAI), Gemini 2.5 Flash (Google), Gemini 3 Pro (Google), and Grok 3 (xAI). These models were selected according to predefined practical and methodological criteria. Eligible systems were required to support direct image input, generate natural-language responses to standardized prompts, be publicly accessible or widely available at the time of evaluation, and represent different major AI providers or model families that dental students, clinicians, or patients may realistically encounter in educational or informational contexts.

Conventional text-only LLMs, task-specific dental imaging algorithms, proprietary radiology platforms without general user access, and models that did not allow standardized image-based prompting were not included.

A total of 55 de-identified dental images obtained from the Recep Tayyip Erdoğan University Clinic Archive between 2024 and 2026 were included. All radiographs were independently assessed by each MLLM under standardized input and prompting conditions, enabling within-case comparisons between models. The reference standard for tooth status and anatomical identifiers was established prior to model evaluation through expert consensus by clinicians with at least 10 years of experience in restorative dentistry (Ö.H.). and endodontics (F.P.H.). The study protocol was approved by the Recep Tayyip Erdoğan University Non-Interventional Ethics Committee (2025/498). All methods were performed in accordance with the relevant institutional guidelines and regulations and in compliance with the Declaration of Helsinki. The requirement for informed consent was waived by the ethics committee because the study was retrospective in nature and involved fully anonymized radiographic data with no direct patient involvement.

Inclusion and exclusion criteria of dental radiographs

Dental periapical radiographs were screened according to predefined inclusion and exclusion criteria.

Inclusion criteria were as follows:

- fully de-identified periapical radiographs before analysis;
- sufficient diagnostic quality to allow reliable assessment of tooth status and anatomical identifiers, including jaw, tooth region, side, and FDI tooth number;
- adequate anatomical coverage and image contrast, with clearly visible key structures such as proximal contacts and coronal anatomy;
- representation of only one target tooth status category, with mutually exclusive categories defined as dental caries, composite restoration, amalgam restoration, or sound tooth;
- for carious, composite-restored, and amalgam-restored categories, the image contained at most one tooth exhibiting the target finding, with no additional teeth showing caries or restorations;
- for the sound control category, the radiograph showed no caries or restorations.

Exclusion criteria were as follows:

- presence of additional markings, annotations, arrows, text overlays, or pre-existing labels that could influence interpretation;
- pronounced motion blur, severe artifacts, or significant exposure errors compromising diagnostic interpretation;
- incomplete or excessively cropped field of view preventing accurate localization;
- residual patient-identifying information;
- more than one eligible radiograph from the same clinical encounter, in which case only one image was selected to reduce within-patient clustering.

A priori sample size calculation

The primary endpoint was tooth status correctness, coded as a binary outcome as correct or incorrect. Because each MLLM evaluated the same set of radiographs, the sample size was calculated for paired proportions using a McNemar framework. A two-sided α level of 0.05 and 80% power were used. An absolute accuracy difference of 20% points was specified a priori as a practically meaningful benchmark effect for this comparative MLLM evaluation. This threshold was not intended to represent a previously established minimal clinically important difference, because direct prior evidence on paired comparisons of MLLMs in periapical radiograph interpretation is limited. Instead, it was selected pragmatically to represent a moderate-to-large performance difference that would be relevant for distinguishing between models in educational, informational, or preliminary clinical-support contexts. Smaller differences may be statistically detectable in larger datasets but were considered less likely to alter the practical interpretation of model performance in this setting.

The discordant-pair proportion was conservatively assumed to be 25%, reflecting uncertainty regarding the expected proportion of radiographs for which two models would yield different correctness outcomes. Under these assumptions, the required sample size was 49 radiographs. Considering approximately 10% potentially non-evaluable cases, the target sample size was set to 55 radiographs.

Radiographic dataset and case distribution

The final dataset consisted of 55 dental periapical radiographs selected to represent a range of tooth types and restorative or carious conditions. The dataset included teeth with dental caries ($n = 15$), composite restorations ($n = 15$), amalgam restorations ($n = 10$), and sound teeth without caries or restorations ($n = 15$). For carious, composite-restored, and sound teeth, the cases were evenly distributed across the anterior, premolar, and molar regions (five teeth per category). Amalgam restorations were represented by premolar ($n = 5$) and molar ($n = 5$) teeth only, reflecting their clinical distribution in contemporary practices. Example dental radiographs from the study dataset are presented in Fig. 1.



Fig. 1 Example dental radiographs from the study dataset by tooth region and tooth status. (A–C) caries (anterior, premolar, molar), (D–F) composite restoration (anterior, premolar, molar), (G) sound tooth, (H–I) amalgam restoration (premolar, molar). Images were de-identified

No additional image preprocessing was performed by the investigators before model evaluation, apart from de-identification. Specifically, no brightness adjustment, contrast enhancement, cropping, resizing, or resolution normalization was applied. Each radiograph was uploaded to all evaluated MLLMs in the same de-identified image format as exported from the clinical archive. Before inclusion, all images were visually screened to ensure sufficient diagnostic quality and anatomical coverage. Images with pronounced motion blur, severe artifacts, significant exposure errors, incomplete field of view, or residual identifying information were excluded.

Reference standard establishment

The reference standard was established before model evaluation and independently of the MLLM-generated outputs. Two clinicians (Ö.H. and F.P.H.) with at least 10 years of experience in restorative dentistry and endodontics independently reviewed all included radiographs using a standardized reference form. For each radiograph, the reviewers recorded the target tooth status category, jaw, tooth region, side, visible FDI tooth number(s), and target tooth identity. After the independent assessments, disagreements were resolved through a consensus meeting in which both reviewers re-evaluated the radiograph together and agreed on a final label. The final consensus decision was used as the reference standard for scoring MLLM responses.

Interobserver agreement before consensus was calculated using Cohen's kappa for categorical variables, including tooth status, jaw identification, tooth region, side determination, and target tooth identification. Percentage agreement was also calculated for each reference-standard component (Supplementary Table 1). The agreement values were interpreted according to conventional benchmarks, and the final consensus labels were used for all subsequent model-performance analyses.

Prompting procedure and model execution

All model evaluations were performed using a standardized prompting procedure. Each de-identified periapical radiograph was uploaded individually to the official public user-facing web interface of each evaluated MLLM. No clinical history, patient demographic information, previous model response, reference label, or additional explanatory context was provided. The same image file and the same prompt wording were used for all models.

The evaluated MLLMs were accessed through public web interfaces rather than application programming interfaces or clinically configured software environments.

Therefore, the evaluations reflected the performance of these systems as encountered by typical users, such as students, clinicians, or patients, in publicly available settings. However, public interfaces may include non-visible preprocessing, safety filters, system-level instructions, model routing, or platform-specific image handling procedures; these factors could not be directly controlled or inspected.

Model evaluations were conducted according to model availability in two evaluation periods. Microsoft Copilot, Grok 3, Gemini 2.5 Flash, and ChatGPT-o4-mini-high were evaluated between March and April 2025. Gemini 3 Pro and ChatGPT-5.2 Thinking were evaluated in January 2026, after these models became available in the corresponding user-facing platforms. For each system, the model name displayed in the official public interface at the time of evaluation was recorded. Because the models were accessed through public web-based interfaces rather than fixed API endpoints, exact backend model snapshots, build numbers, or version hashes were not consistently available across platforms. Therefore, the findings should be interpreted as reflecting the user-facing performance of the displayed model versions during the specified evaluation periods.

For each radiograph, the following standardized prompt was used:

"Please evaluate this de-identified dental periapical radiograph and answer the following questions in a structured format:

1. Which jaw does this radiograph belong to: maxilla or mandible?
2. Does this radiograph correspond to the anterior or posterior region?
3. Does this radiograph belong to the right or left side?
4. Which FDI tooth numbers are visible in this radiograph?
5. Is there dental caries, a composite restoration, or an amalgam restoration present? If yes, please specify the affected tooth or teeth."

To minimize contextual carryover, each radiograph–model combination was evaluated in a new independent session. No follow-up prompts, clarifying questions, regenerated answers, or second attempts were allowed. If a model provided an uncertain or ambiguous answer, the first complete response generated after the standardized prompt was recorded exactly as produced and used for scoring. The order of radiographs was randomized using a computer-generated random sequence, and the same randomized order was applied across all models.

Because the models were evaluated through user-facing interfaces rather than application programming interfaces, parameters such as temperature, top-p, seed, and other decoding settings were not manually configurable across all platforms. Therefore, all systems were evaluated using their default platform settings.

Coding of model responses (correct/incorrect)

Responses to the first four questions were coded as binary outcomes (correct/incorrect) according to predefined criteria. For jaw identification, answers were coded as correct only if the model correctly identified the maxilla versus the mandible; otherwise, they were coded as incorrect. For the tooth region, answers were coded as correct only if the model correctly classified the image as anterior or posterior; incorrect classification was coded as incorrect. For side determination, responses were coded as correct only if the right/left side was correctly identified. For FDI tooth numbering, responses were coded as correct only if the model provided a complete and accurate list of all FDI tooth numbers visible in the radiograph; responses that were incomplete, partially correct, or contained additional incorrect tooth numbers were coded as incorrect.

For the tooth status outcome, correctness required both (i) accurate identification of the target condition (caries, composite restoration, or amalgam restoration) and (ii) correct specification of the affected tooth with sufficient anatomical precision. Specifically, a tooth status response was considered correct only if the model correctly reported the diagnosis and identified the tooth involved using tooth-type specificity (e.g., central incisor, canine, first premolar, second premolar, first molar, or second molar). Importantly, for this tooth-status endpoint, correct identification of the jaw (maxilla/mandible) and side (right/left) was not required for a response to be considered correct; these attributes were evaluated separately as image-level outcomes. Responses were coded as incorrect if the diagnosis was wrong, the involved tooth was not specified, the tooth identification was vague or mismatched (e.g., correct diagnosis but incorrect tooth type), or the model provided ambiguous multiple alternatives without a single clearly defined final answer. The radiograph selection workflow (inclusion/exclusion) and prespecified scoring rules used to define correct/incorrect responses across all endpoints are summarized in Fig. 2.

Primary and secondary outcomes

The study outcomes were defined in advance and evaluated as binary variables (correct or incorrect) for each radiograph–model combination. Five outcomes were obtained directly from the model responses: tooth status (diagnosis), jaw identification, tooth region (anterior/premolar/molar), side (right/left), and FDI tooth-numbering. Based on these variables, two additional composite outcomes were constructed. Localization accuracy was

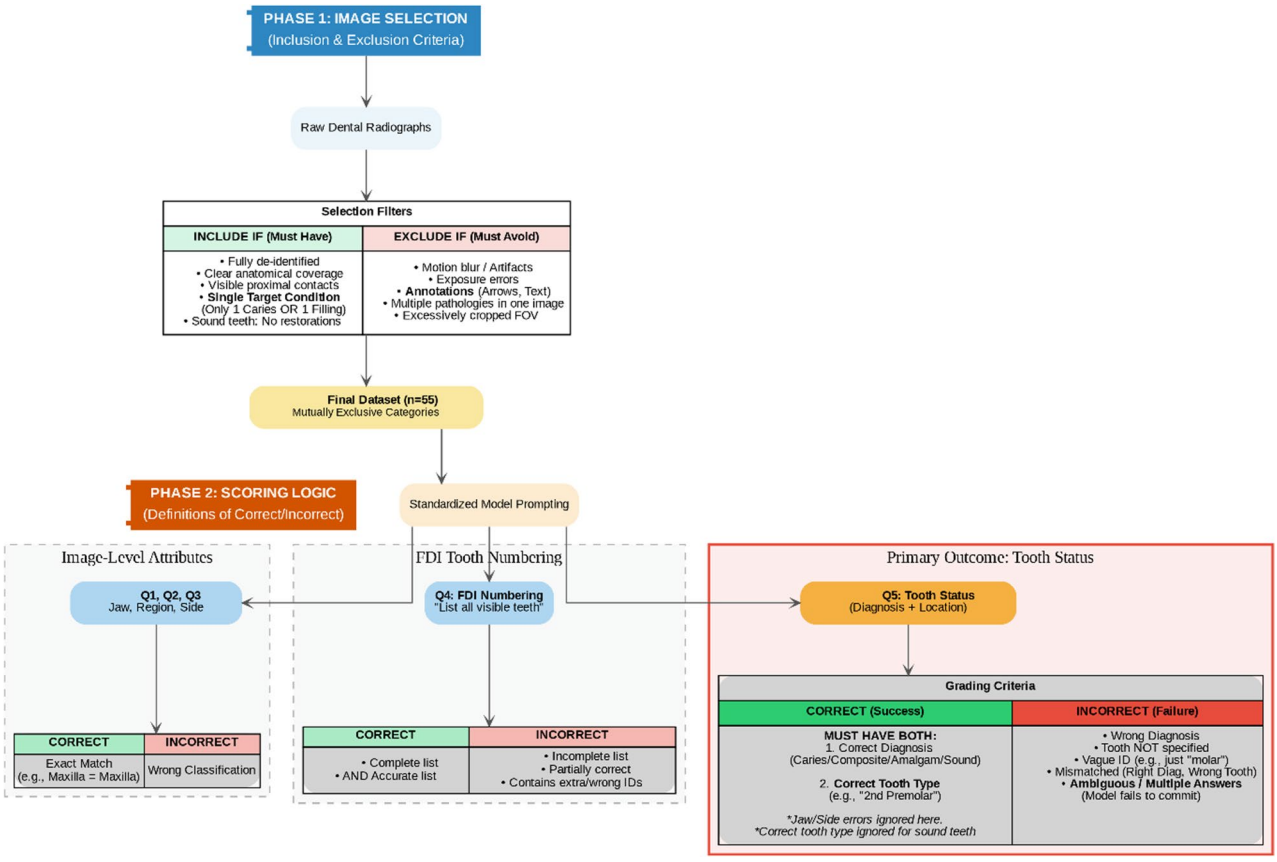


Fig. 2 Study workflow for radiograph selection and scoring

defined as the simultaneous correct identification of the jaw, side, and FDI tooth number. The overall accuracy across all attributes was defined as correct performance across all five directly assessed outcomes. The all-attributes endpoint was intentionally defined as a stringent composite measure of complete response correctness and was intended to complement, rather than replace, the individual task-level outcomes.

The primary outcome of the study was tooth status (diagnosis), which reflected whether the model correctly identified the target condition (caries, composite restoration, or amalgam restoration) and accurately specified the affected tooth with appropriate tooth-type details. The remaining outcomes (jaw, tooth region, side, and FDI tooth numbering) and the two composite measures (localization accuracy and overall accuracy across all attributes) were considered secondary outcomes, providing a more detailed characterization of model performance across increasing levels of anatomical specificity and task complexity.

Randomization and bias control

To reduce potential evaluation bias, the order of radiographs was randomized using a computer-generated random sequence, and the same randomized order was applied across all models. All images were de-identified and reviewed to ensure that they contained no arrows, text labels, annotations, or other external cues that could influence interpretation. Model outputs were recorded verbatim and coded according to predefined scoring rules.

Statistical analysis

All analyses were conducted using a paired study design, in which the same set of dental radiographs was independently evaluated by each MLLM. The model responses were coded as binary outcomes (correct/incorrect) for each task, including tooth-status classification, jaw identification, tooth region identification, side determination, FDI tooth number, composite localization (jaw + region + side + FDI), and complete correctness across all attributes (5/5). Descriptive performance was summarized as accuracy percentages with 95% Wilson confidence intervals, which were chosen because of their robustness for binomial proportions.

To assess whether overall differences in accuracy existed between the MLLMs for each task, Cochran's Q test was applied as an omnibus test suitable for paired binary outcomes across multiple classifiers. When Cochran's Q indicated significant between-model

heterogeneity, further model-based analyses were performed rather than relying solely on multiple pairwise tests. Specifically, generalized estimating equations (GEE) with a logistic link function were used to estimate adjusted odds ratios and marginal (population-averaged) accuracy estimates, accounting for within-case correlation by clustering on the case identifier and assuming an exchangeable correlation structure. Reference tooth status and tooth region were included as covariates to control for potential differences in task difficulty between categories. An exchangeable working correlation structure was selected because each radiograph was evaluated by the same set of MLLMs and the repeated observations within each case were expected to be correlated without a natural temporal or hierarchical ordering among models.

From the fitted GEE models, marginal (adjusted) accuracy estimates were derived for each MLLM by averaging the predicted probabilities over the observed covariate distribution. In addition, pairwise contrasts between MLLMs were computed from the same GEE models to estimate odds ratios (ORs) with 95% confidence intervals (CIs) for selected outcomes. To control for the inflation of type I error due to multiple comparisons, Holm's step-down adjustment was applied to all pairwise p-values. Pairwise comparisons were interpreted cautiously and primarily used to characterize the differences identified by the omnibus tests rather than as the sole basis for inference. GEE model diagnostics were assessed for each endpoint by checking successful model fitting, convergence or numerical warnings, finite parameter estimates, finite robust standard errors, and finite covariance matrices. All fitted GEE models completed successfully without convergence or numerical warnings and produced finite parameter estimates, robust standard errors, and covariance matrices. All statistical analyses were performed using Python-based statistical libraries, and statistical significance was defined as a two-sided p-value < 0.05 after adjustment.

Results

Across the 55 dental radiographs assessed by each model, the performance differed markedly depending on the evaluated task. The accuracy for tooth status (diagnosis) classification was moderate and broadly comparable across MLLMs, ranging from 45.5% to 60.0%. In contrast, all models performed well in identifying the jaw and tooth regions, with several MLLMs, most notably Gemini 3 Pro, ChatGPT-5.2 Thinking, and ChatGPT-o4-mini-high, achieving accuracies exceeding 94%. Performance was lower for tasks requiring greater anatomical specificity. Lower accuracies and greater variability between models were observed for the side, FDI tooth number, and particularly for composite localization combining the jaw, side, and FDI number. The most demanding endpoint, complete correctness across all attributes, showed the widest spread in performance, with Gemini 3 Pro achieving the highest accuracy (43.6%), while the remaining models ranged between 9.1% and 25.5% (Fig. 3). Representative examples of correct and incorrect MLLM outputs, including localization errors, FDI numbering errors, and partially correct responses, are provided in Supplementary Table 2.

Stratified analyses revealed that the tooth-status classification accuracy varied substantially according to the reference tooth status. The classification of amalgam restorations was consistently accurate across models, with several MLLMs achieving perfect or near-perfect performance (70.0–100%) and relatively narrow confidence intervals. In contrast, the results for composite restorations and dental caries were far more heterogeneous. For dental caries, the accuracy ranged from 6.7% to 73.3% across models, whereas a similarly wide range was observed for composite restorations. The performance for sound teeth also varied considerably, with accuracies between 26.7% and 100.0%, underscoring the strong influence of the diagnostic category on model performance (Fig. 4).

When the analyses were stratified by tooth region, the overall accuracy remained within a narrow and moderate range across the models. For anterior teeth, the accuracy values ranged from 40.0% to 66.7%, whereas for premolars, they ranged from 40.0% to 65.0%. The performance for molars was comparable, with accuracies between 45.0% and 65.0%. No single MLLM demonstrated a consistent advantage across all tooth region categories (Fig. 5).

Omnibus testing using Cochran's Q test identified significant between-model differences for most of the evaluated tasks (Table 1). Statistically significant heterogeneity was observed for jaw identification, tooth region, side, FDI tooth number, composite localization, and complete accuracy across all attributes (all Holm-adjusted $p < 0.001$). In contrast, no significant differences were detected in the tooth status classification ($Q = 3.35$, $p = 0.646$).

The results from the GEE models provided adjusted marginal accuracy estimates that accounted for the reference tooth status and tooth region (Table 2). After adjustment, Gemini 3 Pro consistently achieved the highest marginal accuracy across all evaluated tasks, including tooth status classification (60.0%), jaw identification (96.4%), tooth region identification (96.4%), and localization-related outcomes. The differences between the models became increasingly evident as the anatomical specificity increased. While high adjusted accuracies were maintained for jaw and tooth region identification by several MLLMs, substantially lower and more variable performance was observed for side, FDI number, and

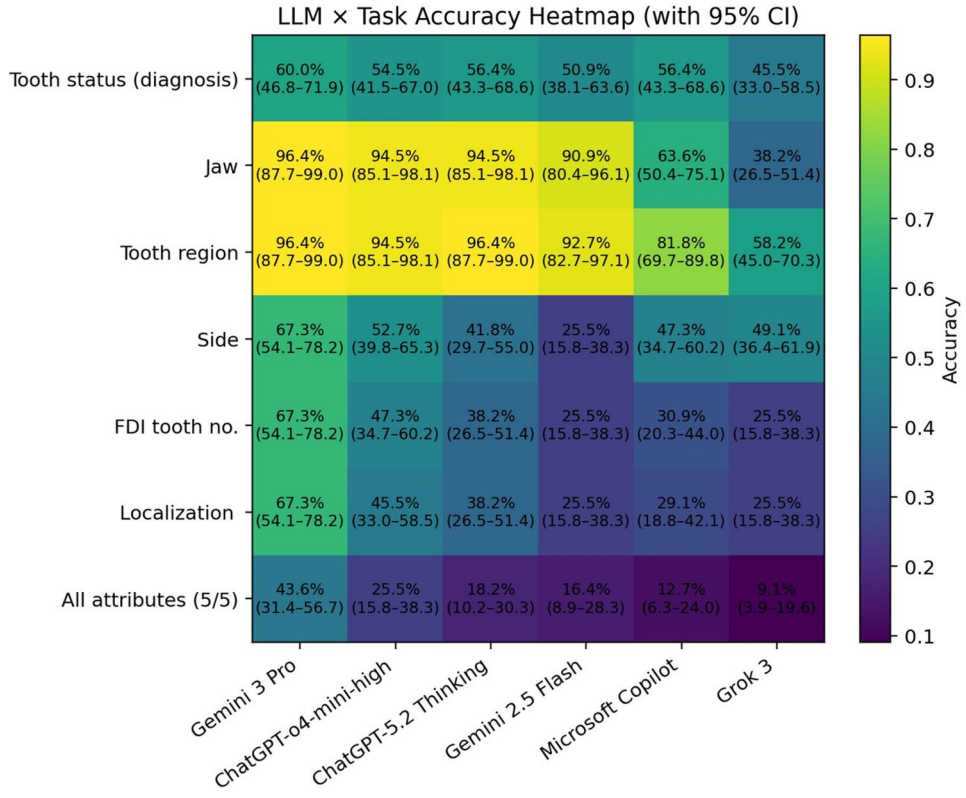


Fig. 3 Task-level diagnostic performance of large language models

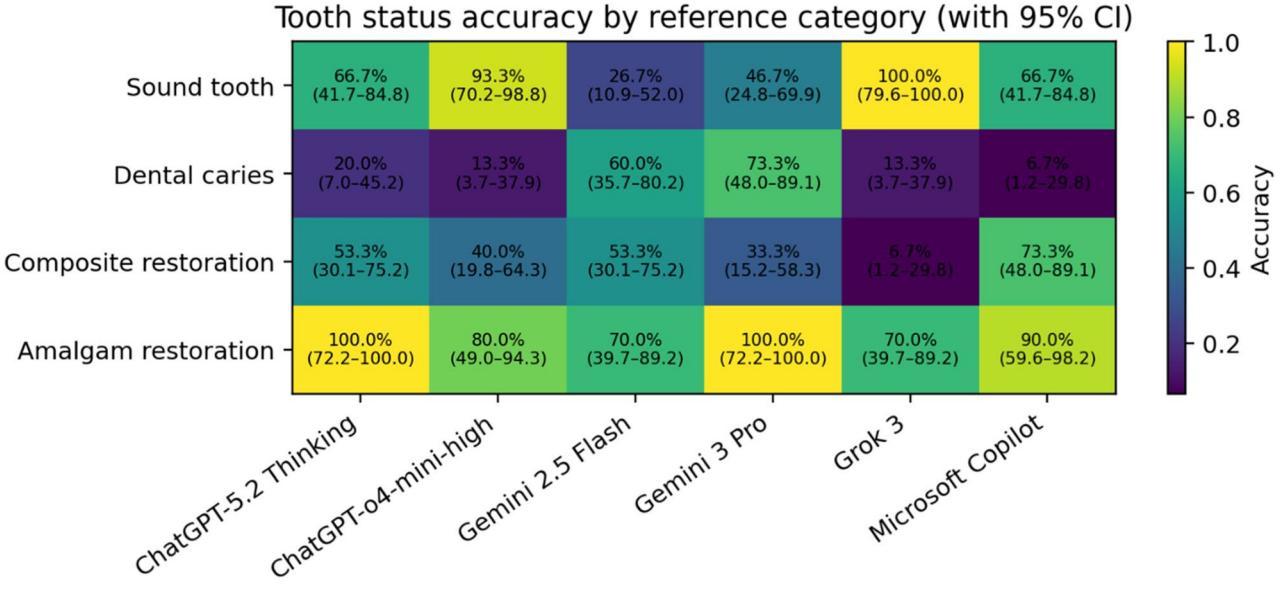


Fig. 4 Tooth status accuracy by reference category

composite localization, with adjusted accuracies ranging from 25.6% to 67.4%. For complete correctness, the adjusted accuracies were uniformly low across the models, ranging from 9.1% to 43.7%, with Gemini 3 Pro again showing the highest performance.

Pairwise comparisons based on the GEE model indicated that tooth status classification did not differ significantly between the individual MLLMs after adjustment for multiple tests (Table 3). Although several contrasts showed numerical differences in odds ratios, all confidence intervals included unity, and none remained

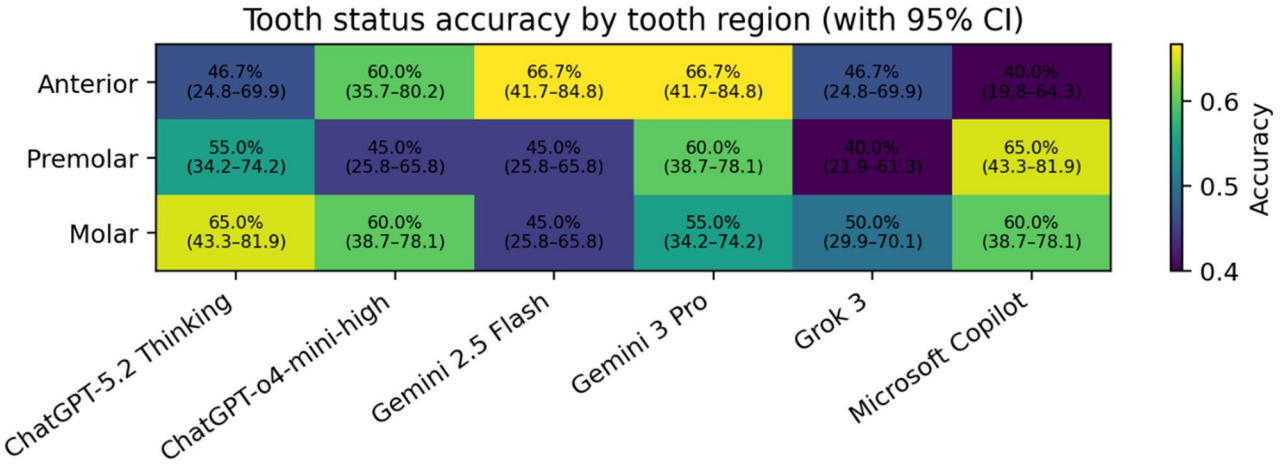


Fig. 5 Tooth status accuracy by tooth region across MLLMs

Table 1 Cochran's Q test for differences in accuracy across LLMs (paired across cases)

Task	Cochran's Q	df	*p*-value	Holm-adjusted *p*
Tooth status (diagnosis)	3.35	5	0.646	0.646
Jaw (maxilla/mandible)	96.77	5	< 0.001	**< 0.001**
Tooth region (anterior/premolar/molar)	52.47	5	< 0.001	**< 0.001**
Side (left/right)	23.03	5	< 0.001	**< 0.001**
FDI tooth number	34.62	5	< 0.001	**< 0.001**
Localization (jaw+side + FDI)	36.25	5	< 0.001	**< 0.001**
All attributes (5/5)	27.91	5	< 0.001	**< 0.001**

significant after Holm correction. In contrast, analyses of localization correctness suggested between-model differences (Table 4), with Gemini 3 Pro showed higher odds of correct localization compared with Grok 3, Gemini 2.5 Flash, Microsoft Copilot, and ChatGPT-5.2 Thinking (all Holm-adjusted $p \leq 0.020$). A similar pattern was observed for complete correctness across all attributes, where Gemini 3 Pro again showed significantly higher odds than the same set of models (Table 5; Holm-adjusted $p \leq 0.041$).

Discussion

In recent years, the use of MLLMs in dentistry has increased substantially, with these systems being widely adopted as tools to support learning in dental education and facilitate patient communication [14, 15]. Periapical radiographs, which are an integral part of everyday clinical practice, represent one of the primary foundations for tooth-level diagnosis and treatment decisions [16]. Nevertheless, a review of the current literature indicates that the ability of MLLMs to interpret dental radiographic images has most often been assessed using a single model or within a limited set of tasks [17]. This hampers systematic comparisons of performance patterns across models with different architectural and multimodal characteristics [18]. In contrast, evaluating multiple language models together allows for a more balanced and comparative identification of their respective strengths and limitations [19]. While many studies have focused primarily on diagnostic labeling or the generation of general radiographic interpretations, clinically critical components, such as tooth-level anatomical precision, side determination, and tooth numbering, have frequently been treated as secondary considerations [14, 17]. Therefore, a comparative analysis of multiple MLLMs within a structured, multicomponent evaluation framework may offer a more

Table 2 GEE-based marginal (adjusted) accuracy across periapical radiograph interpretation tasks ($n = 55$ cases per MLLM)

Rank	LLM	Tooth status (diagnosis)	Jaw	Tooth region	Side	FDI tooth no.	Localization (Jaw+Side + FDI)	All attributes (5/5)
1	Gemini 3 Pro	60.0%	96.4%	96.4%	67.4%	67.4%	67.4%	43.7%
2	Microsoft Copilot	56.3%	63.7%	81.9%	47.4%	31.1%	29.3%	12.8%
3	ChatGPT-5.2 Thinking	56.3%	94.6%	96.4%	41.9%	38.3%	38.4%	18.2%
4	ChatGPT-o4-mini-high	54.5%	94.6%	94.6%	52.8%	47.4%	45.6%	25.5%
5	Gemini 2.5 Flash	50.9%	90.9%	92.8%	25.6%	25.6%	25.6%	16.4%
6	Grok 3	45.4%	38.2%	58.2%	49.2%	25.6%	25.6%	9.1%

Table 3 Full pairwise GEE contrasts for tooth-status correctness

Comparison (A vs. B)	OR (95% CI)	Favours	*p*	p_holm
Grok 3 vs. Gemini 3 Pro	0.49 (0.20–1.19)	Gemini 3 Pro	0.117	1.000
ChatGPT-o4-mini-high vs. Grok 3	1.56 (0.80–3.02)	ChatGPT-o4-mini-high	0.191	1.000
Microsoft Copilot vs. Grok 3	1.70 (0.75–3.84)	Microsoft Copilot	0.200	1.000
Grok 3 vs. ChatGPT-5.2 Thinking	0.59 (0.26–1.33)	ChatGPT-5.2 Thinking	0.201	1.000
Gemini 2.5 Flash vs. Gemini 3 Pro	0.64 (0.29–1.42)	Gemini 3 Pro	0.275	1.000
ChatGPT-o4-mini-high vs. Gemini 3 Pro	0.77 (0.33–1.77)	Gemini 3 Pro	0.533	1.000
Gemini 2.5 Flash vs. ChatGPT-5.2 Thinking	0.77 (0.32–1.83)	ChatGPT-5.2 Thinking	0.549	1.000
Microsoft Copilot vs. Gemini 2.5 Flash	1.30 (0.51–3.31)	Microsoft Copilot	0.578	1.000
Gemini 2.5 Flash vs. Grok 3	1.31 (0.51–3.33)	Gemini 2.5 Flash	0.578	1.000
Gemini 3 Pro vs. ChatGPT-5.2 Thinking	1.20 (0.53–2.72)	Gemini 3 Pro	0.670	1.000
Microsoft Copilot vs. Gemini 3 Pro	0.84 (0.33–2.11)	Gemini 3 Pro	0.706	1.000
ChatGPT-o4-mini-high vs. Gemini 2.5 Flash	1.19 (0.46–3.07)	ChatGPT-o4-mini-high	0.715	1.000
ChatGPT-o4-mini-high vs. ChatGPT-5.2 Thinking	0.92 (0.45–1.88)	ChatGPT-5.2 Thinking	0.809	1.000
Microsoft Copilot vs. ChatGPT-o4-mini-high	1.09 (0.49–2.42)	Microsoft Copilot	0.827	1.000
Microsoft Copilot vs. ChatGPT-5.2 Thinking	1.00 (0.44–2.26)	Tie	1.000	1.000

Table 4 Pairwise GEE contrasts for localization correctness

Comparison (A vs. B)	OR (95% CI)	Favours	*p*	p_holm
Grok 3 vs. Gemini 3 Pro	0.16 (0.07–0.35)	Gemini 3 Pro	**< 0.001**	**< 0.001**
Gemini 2.5 Flash vs. Gemini 3 Pro	0.16 (0.06–0.38)	Gemini 3 Pro	**< 0.001**	**< 0.001**
Microsoft Copilot vs. Gemini 3 Pro	0.17 (0.07–0.40)	Gemini 3 Pro	**< 0.001**	**< 0.001**
Gemini 3 Pro vs. ChatGPT-5.2 Thinking	3.49 (1.60–7.58)	Gemini 3 Pro	**0.002**	**0.020**
ChatGPT-o4-mini-high vs. Gemini 2.5 Flash	2.51 (1.32–4.80)	ChatGPT-o4-mini-high	0.005	0.057
ChatGPT-o4-mini-high vs. Grok 3	2.51 (1.12–5.61)	ChatGPT-o4-mini-high	0.025	0.224
ChatGPT-o4-mini-high vs. Gemini 3 Pro	0.39 (0.17–0.89)	Gemini 3 Pro	0.026	0.224
Microsoft Copilot vs. ChatGPT-o4-mini-high	0.44 (0.22–0.89)	ChatGPT-o4-mini-high	0.022	0.224
Gemini 2.5 Flash vs. ChatGPT-5.2 Thinking	0.54 (0.27–1.08)	ChatGPT-5.2 Thinking	0.084	0.586
Grok 3 vs. ChatGPT-5.2 Thinking	0.54 (0.23–1.27)	ChatGPT-5.2 Thinking	0.158	0.946
Microsoft Copilot vs. Gemini 2.5 Flash	0.60 (0.27–1.31)	Microsoft Copilot	0.198	0.991
Microsoft Copilot vs. Grok 3	1.10 (0.59–2.05)	Microsoft Copilot	0.762	1.000
Microsoft Copilot vs. ChatGPT-5.2 Thinking	1.10 (0.53–2.28)	ChatGPT-5.2 Thinking	0.796	1.000
Gemini 2.5 Flash vs. Grok 3	1.00 (0.49–2.05)	Grok 3	1.000	1.000
ChatGPT-o4-mini-high vs. ChatGPT-5.2 Thinking	1.36 (0.78–2.40)	ChatGPT-o4-mini-high	0.281	1.000

clinically relevant understanding of their potential role in dental radiographic interpretation [20].

The findings of this study suggest that MLLM performance in dental periapical radiograph interpretation was strongly task-dependent. While the models generally performed better in tasks with clearer visual boundaries, such as jaw and tooth region identification, accuracy decreased in tasks requiring side determination, FDI tooth numbering, and integration of multiple anatomical attributes. This pattern supports the view that radiographic interpretation should be evaluated as a multi-component task rather than summarized by a single global accuracy metric [16, 19]. In contrast to our findings, Wu et al. reported high consistency of GPT-4o in tooth numbering and side determination, suggesting that model version, dataset characteristics, prompt design, and task definition may influence reported performance [21]. In the present study, Gemini 3 Pro showed

Table 5 Pairwise GEE contrasts for all attributes correctness

Comparison (A vs. B)	OR (95% CI)	Favours	*p*	p_holm
Grok 3 vs. Gemini 3 Pro	0.12 (0.04–0.33)	Gemini 3 Pro	< 0.001	**< 0.001**
Microsoft Copilot vs. Gemini 3 Pro	0.17 (0.06–0.46)	Gemini 3 Pro	< 0.001	**0.006**
Gemini 2.5 Flash vs. Gemini 3 Pro	0.23 (0.09–0.59)	Gemini 3 Pro	0.002	**0.030**
Gemini 3 Pro vs. ChatGPT-5.2 Thinking	3.75 (1.55–9.10)	Gemini 3 Pro	0.003	**0.041**
ChatGPT-o4-mini-high vs. Grok 3	3.60 (1.06–12.26)	ChatGPT-o4-mini-high	0.040	0.436
ChatGPT-o4-mini-high vs. Gemini 3 Pro	0.42 (0.18–0.96)	Gemini 3 Pro	0.040	0.436
Microsoft Copilot vs. ChatGPT-o4-mini-high	0.41 (0.16–1.06)	ChatGPT-o4-mini-high	0.065	0.589
Microsoft Copilot vs. Grok 3	1.48 (0.44–4.95)	Microsoft Copilot	0.527	1.000
Microsoft Copilot vs. ChatGPT-5.2 Thinking	0.64 (0.23–1.81)	ChatGPT-5.2 Thinking	0.403	1.000
ChatGPT-o4-mini-high vs. Gemini 2.5 Flash	1.80 (0.74–4.34)	ChatGPT-o4-mini-high	0.193	1.000
Microsoft Copilot vs. Gemini 2.5 Flash	0.74 (0.29–1.89)	Gemini 2.5 Flash	0.525	1.000
ChatGPT-o4-mini-high vs. ChatGPT-5.2 Thinking	1.57 (0.65–3.79)	ChatGPT-o4-mini-high	0.314	1.000
Gemini 2.5 Flash vs. ChatGPT-5.2 Thinking	0.88 (0.37–2.08)	ChatGPT-5.2 Thinking	0.763	1.000
Gemini 2.5 Flash vs. Grok 3	2.01 (0.68–5.88)	Gemini 2.5 Flash	0.205	1.000
Grok 3 vs. ChatGPT-5.2 Thinking	0.44 (0.15–1.28)	ChatGPT-5.2 Thinking	0.131	1.000

a relatively more balanced accuracy profile under the tested conditions, whereas Grok 3 and Microsoft Copilot showed greater variability, particularly in tasks involving spatial referencing. However, these differences should be interpreted as observable output-level patterns, because the underlying technical mechanisms could not be directly examined [20].

In this study, the performance of MLLMs in determining tooth status (diagnosis) from dental periapical radiographs was moderate across all evaluated models, with no statistically significant differences observed between them [22]. This may be related to the inherent cognitive and visual challenges associated with diagnostic tasks. Nevertheless, the diagnostic accuracy varies depending on the reference category and model-specific technical characteristics [23]. Gemini 3 Pro demonstrated relatively more consistent accuracy profiles in diagnostically ambiguous categories, such as composite restorations and dental caries, which may be attributed to differences in its ability to integrate visual and textual information [22, 24]. In contrast, the greater variability in accuracy observed for Grok 3 and, in certain tasks, Microsoft Copilot, may reflect limitations in interpreting radiographic findings with low contrast or poorly defined boundaries [25]. When evaluated using the reference category, the consistently high accuracy achieved by all models in identifying amalgam restorations can likely be explained by the high radiopacity and well-defined margins of the material, which provide strong visual discriminative cues [24]. In comparison, the more heterogeneous performance observed for composite restorations and the wide accuracy range seen in dental caries assessments are consistent with the inherently ambiguous and interpretation-dependent radiographic appearance of these conditions. However, in contrast to our findings, a study by Dave et al. reported that GPT-4 achieved higher diagnostic accuracy in dental radiographic interpretation tasks [26], highlighting potential differences related to model architecture, training data or multimodal processing strategies.

The observation that accuracy values stratified by tooth region (anterior, premolar, and molar) largely overlapped across models suggests that the tooth region alone may not serve as a strong discriminator of performance [27]. The substantial overlap in accuracy ranges among the anterior, premolar, and molar teeth indicates that large language models tend to interpret dental radiographs based on more general visual patterns rather than region-specific morphological cues [28, 29]. Model-specific analyses revealed that Gemini 3 Pro largely maintained stable accuracy levels regardless of the tooth region, showing no pronounced fluctuation between the anterior and posterior areas [24, 29]. In contrast, Grok 3 and Microsoft Copilot exhibited greater variability in accuracy, particularly in the molar regions; however, this variability did not translate into a consistent pattern of regional advantage or disadvantage [28, 30]. For ChatGPT-5.2 Thinking and ChatGPT-o4-mini-high, the performance differences associated with the tooth region were minimal, with the overall accuracy remaining relatively consistent across regions [1]. Taken together, these findings suggest that tooth region information may play a secondary role in the diagnostic decision-making processes of these

models, whereas the overall multimodal representation capacity appears to be a more influential determinant of performance [24].

The relatively lower accuracy observed for localization accuracy and the composite endpoint requiring correctness across all attributes may be attributed to the multicomponent and hierarchical nature of these tasks [13, 31]. While more acceptable performance was achieved in tasks focusing on individual anatomical features in isolation, error rates tended to increase in scenarios requiring the simultaneous and coherent integration of multiple pieces of anatomical information [24, 31]. Model-specific comparisons suggested that Gemini 3 Pro tended to show a more consistent accuracy profile for both localization accuracy and the combined evaluation of all attributes, whereas Grok 3 and Microsoft Copilot exhibited greater variability in performance across these endpoints [13, 27]. Case-level observations further suggested that localization errors were more likely to occur in radiographs containing multiple teeth within the same image or featuring symmetrical anatomical structures, where spatial discrimination is inherently more challenging [32].

The performance profiles observed in this study suggest that the reliability of the information provided by MLLMs when interpreting dental periapical radiographs may vary depending on the type of task and the specific model involved [12, 25]. Models that demonstrate more consistent accuracy across different tasks may offer more predictable outputs for identifying basic anatomical features and explaining selected radiographic findings [12]. In contrast, for models showing greater variability in accuracy, particularly in scenarios requiring tooth-level detail and precise localization, the generated responses may need to be interpreted with greater caution by users [23]. From an educational perspective, these findings indicate that such models may serve as supportive tools for learning and discussion during radiographic interpretation training sessions. However, clinical decisions that directly influence tooth-level diagnosis and treatment planning should not be based solely on model-generated outputs [29]. Similarly, when patients seek explanations regarding their dental radiographs, the information provided by these models may create a misleading sense of confidence, especially in cases where an otherwise correct diagnosis is accompanied by uncertain or incorrect localization.

This study examined the performance of different MLLMs in interpreting dental periapical radiographs using a structured evaluation framework that extended beyond diagnostic accuracy alone to include tooth-level anatomical precision. The findings suggest that while these models can identify and describe certain fundamental radiographic features, their performance becomes markedly model-specific in scenarios requiring consistent and comprehensive integration of information at the tooth level. In the existing literature, the performance of MLLMs in dental imaging tasks has most often been assessed through isolated diagnostic tasks, global accuracy metrics, or limited anatomical outputs. Even in studies adopting multimodal approaches, the simultaneous accuracy of multiple anatomical and diagnostic components at the tooth level has frequently been considered a secondary outcome [12, 13, 33]. In this context, the present findings support a task-specific interpretation of MLLM performance in dental radiographic interpretation rather than reliance on a single global accuracy metric. Diagnostic recognition, anatomical localization, FDI numbering, and complete response consistency should be considered as separate output-level dimensions, particularly because the proprietary nature of current MLLMs limits direct conclusions regarding the mechanisms underlying between-model differences [13, 24]. Adopting such an approach may facilitate a more transparent understanding of the strengths and limitations of different model designs when applied to complex radiographic interpretation tasks.

It is important to interpret the present findings in the context of existing computer vision approaches in dental radiology. Task-specific convolutional neural network or deep learning systems trained directly on dental radiographic datasets are generally optimized for narrowly defined image-analysis tasks, such as detecting caries [34], restorations [35], periapical lesions [36], or anatomical structures [37]. Such models may achieve substantially higher performance than general-purpose MLLMs when applied to the same predefined radiographic task, particularly because they are trained and validated using domain-specific imaging data and fixed output labels. In contrast, the MLLMs evaluated in the present study were not task-specific dental radiology algorithms, but general-purpose image-capable language models that generated free-text responses to standardized prompts. Therefore, the present results should not be interpreted as a direct comparison with dedicated dental computer vision systems or as evidence that MLLMs can replace them. Rather, the findings highlight the current limitations of general-purpose MLLMs in tooth-level radiographic interpretation and support their use primarily as educational or explanatory adjuncts rather than standalone diagnostic systems.

These findings should be interpreted in light of the study's methodological framework. Although the dental periapical radiographs included in this study were selected in a balanced manner to enable multimodel and multitask comparisons, they were not intended to represent the full spectrum of morphological variations, rare clinical scenarios, or complex mixed findings encountered in routine clinical practice. In particular, each

radiograph was intentionally restricted to a single target finding category, and images containing multiple carious lesions, multiple restorations, or mixed restorative and carious findings were excluded. This restriction was a deliberate design choice aimed at improving scoring clarity, reducing diagnostic ambiguity, and allowing each MLLM-generated response to be attributed unambiguously to a predefined target condition and anatomical location. However, this approach also reduces ecological validity, because real-world periapical radiographs frequently contain multiple teeth with concurrent restorations, carious lesions, previous treatments, or overlapping findings. Therefore, the present results should be interpreted as reflecting model performance under relatively controlled and unambiguous radiographic conditions rather than fully naturalistic clinical scenarios.

In the existing literature, the performance of MLLMs in dental imaging tasks has most commonly been assessed using isolated diagnostic tasks, global accuracy metrics, or panoramic radiographs. In contrast, structured comparative approaches that prioritize the simultaneous accuracy of multiple diagnostic and anatomical components at the tooth level in periapical radiographs remain relatively limited [10, 13, 38]. To the best of our knowledge, only a few studies have adopted a comparable framework to evaluate multiple MLLMs using the same image set and assessment criteria. Nevertheless, several methodological considerations should be acknowledged. First, although testing all models under an identical prompt structure and using the same image set enhanced comparability, the potential impact of alternative prompting strategies was not examined as an independent variable [39, 40]. Second, the dataset was highly curated and all models were evaluated under standardized prompting conditions. In addition, all radiographs were obtained from a single institutional archive under an ethics approval limited to that center. Therefore, additional external or multicenter radiographic datasets could not be included within the scope of the present approved protocol. This single-center design may limit the generalizability of the findings to other institutions, imaging devices, acquisition protocols, and patient populations. Therefore, the findings may not fully reflect real-world clinical usage environments, where radiographs may vary in quality, contain multiple concurrent findings, be accompanied by variable clinical information, and be submitted using non-standardized prompts or follow-up questions. Future studies should evaluate MLLMs using larger and more naturalistic periapical radiograph datasets that include multiple concurrent findings and mixed clinical scenarios to better estimate their performance under real-world conditions.

Additional limitations relate to model transparency, statistical precision, and response-level characterization. Because the evaluated MLLMs are largely proprietary systems, detailed information regarding their model architectures, training datasets, image-processing pipelines, and multimodal integration mechanisms was not available. Therefore, explanations related to architectural differences, training data, or multimodal integration capacity should be interpreted as plausible hypotheses rather than experimentally verified mechanisms; the present study compared observable model outputs under standardized conditions and cannot determine the underlying technical causes of between-model performance differences [11]. Although the paired design improved statistical efficiency, the overall dataset size remained modest relative to the number of evaluated models, endpoints, and pairwise contrasts. Accordingly, some subgroup analyses and pairwise odds ratio estimates should be interpreted cautiously, as the limited number of cases may have contributed to wide confidence intervals and reduced precision; Holm adjustment controls for multiple testing but does not address imprecision arising from sparse outcomes or limited sample size. Finally, the binary scoring strategy did not capture partially correct responses, repeated responses were not obtained to assess intra-model consistency, and a formal qualitative error taxonomy was not performed. Thus, partially useful outputs, test–retest variability, hallucinated anatomical structures, fabricated or non-visible FDI tooth numbers, ambiguous tooth identifications, and false-positive anatomical or restorative descriptions were not systematically analyzed as separate outcomes. Future studies should incorporate partial-credit scoring, repeated prompting designs, and predefined qualitative error taxonomies to better characterize the reliability and clinical interpretability of MLLM-generated radiographic responses.

Nevertheless, the tooth-level, anatomy-focused approach adopted in this study highlights both the strengths and limitations of MLLMs in interpreting dental periapical radiographs within a structured framework that is clinically relevant but still more controlled than real-world clinical use. While the existing literature suggests that such systems can offer valuable support as learning and discussion tools in dental education, it also emphasizes the importance of critically appraising model outputs, particularly in tasks requiring detailed anatomical assessment and precise localization [10, 13, 38]. The findings of the present study further suggest that MLLM-generated information should be interpreted in a manner that is sensitive to both task characteristics and clinical context. In scenarios requiring tooth-level localization, numbering, and integration of multiple anatomical features, careful and critical evaluation of model outputs by the user remains clinically important [7, 8]. In this respect, the present study provides a balanced evaluation

framework that outlines the potential applications of MLLMs in dental radiographic interpretation while clearly reflecting their current limitations [9].

Conclusion

This study demonstrates that the performance of contemporary MLLMs in interpreting dental periapical radiographs is highly task-dependent. While these models exhibit high accuracy in identifying broad anatomical categories such as the jaw and tooth region, their reliability significantly decreases in tasks requiring precise tooth-level localization, side determination, and FDI tooth numbering. Diagnostic accuracy for tooth status remains moderate and varies substantially depending on the target condition; well-defined features like amalgam restorations are identified more consistently than visually ambiguous conditions like composite restorations or dental caries. Although certain models, notably Gemini 3 Pro, displayed relatively higher and more consistent performance across complex composite endpoints, the overall accuracy for complete anatomical and diagnostic correctness was uniformly low across all evaluated systems. Consequently, current general-purpose MLLMs are not suitable as standalone clinical diagnostic tools for dental radiography. Their clinical and educational utility is presently limited to serving as explanatory adjuncts. To prevent diagnostic errors and miscommunication, strict user oversight and critical appraisal remain essential, particularly for verifying precise anatomical localization and tooth-level findings. Future developments must prioritize the simultaneous integration of multi-component anatomical data to improve the clinical reliability of these models.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12903-026-08870-x.

Supplementary Material 1.

Supplementary Material 2.

Authors' contributions

Ö. H.: Conceptualization, Methodology, Formal analysis, Writing – original draft, Writing – review & editing, Supervision, Project administration. M. B.: Investigation, Data curation. F. P. H.: Writing – original draft, Writing – review & editing.

Funding

No Funding was received.

Data availability

The datasets generated and/or analyzed during the current study are not publicly available considering that we have not required consents to publish this data, but are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

The study protocol was approved by the Recep Tayyip Erdoğan University Non-Interventional Ethics Committee (2025/498). All methods were performed in accordance with the relevant institutional guidelines and regulations and in compliance with the Declaration of Helsinki. The requirement for informed consent was waived by the ethics committee because the study was retrospective in nature and involved fully anonymized radiographic data with no direct patient involvement.

Competing interests

The authors declare no competing interests.

Author details

[1]Department of Restorative Dentistry, Faculty of Dentistry, Recep Tayyip Erdoğan University, Rize, Turkey
[2]Department of Endodontics, Recep Tayyip Erdoğan University, Rize, Turkey

Received: 13 April 2026 / Accepted: 4 June 2026
Published online: 08 June 2026

References

1. Maity S, Saikia MJ. Large Language Models in Healthcare and Medical Applications: A Review. Bioengineering. 2025;12(6):631.
2. Iqbal U, Tanweer A, Rahmanti AR, Greenfield D, Lee LT-J, Li Y-CJ. Impact of large language model (ChatGPT) in healthcare: an umbrella review and evidence synthesis. J Biomed Sci. 2025;32(1):45.
3. Zhou J, Cheng Y, He S, Chen Y, Chen H. Large Language Models for Transforming Healthcare: A Perspective on DeepSeek-R1. MedComm–Future Med. 2025;4(2):e70021.
4. Putra RH, Doi C, Yoda N, Astuti ER, Sasaki K. Current applications and development of artificial intelligence for digital dental radiography. Dentomaxillofacial Radiol. 2022;51(1):20210197.
5. Hamdan MH, Tuzova L, Mol A, Tawil PZ, Tuzoff D, Tyndall DA. The effect of a deep-learning tool on dentists' performances in detecting apical radiolucencies on periapical radiographs. Dentomaxillofacial Radiol. 2022;51(7):20220122.
6. Issa J, Jaber M, Rifai I, Mozdziak P, Kempisty B, Dyszkiewicz-Konwińska M. Diagnostic test accuracy of artificial intelligence in detecting periapical periodontitis on two-dimensional radiographs: a retrospective study and literature review. Medicina. 2023;59(4):768.
7. 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. Am J Orthod Dentofac Orthop. 2025;168(2):220–31. https://doi.org/10.1016/j.ajodo.2025.02.017.
8. Metin U, Goymen M. Information from digital and human sources: a comparison of chatbot and clinician responses to orthodontic questions. Am J Orthod Dentofac Orthop. 2025;168(3):348–57. https://doi.org/10.1016/j.ajodo.2025.04.008.
9. Heo M-S, Kim J-E, Hwang J-J, Han S-S, Kim J-S, Yi W-J. Park I-W: Artificial intelligence in oral and maxillofacial radiology: what is currently possible? Dentomaxillofacial Radiol. 2021;50(3):20200375.
10. Mine Y, Okazaki S, Taji T, Kawaguchi H, Kakimoto N, Murayama T. Benchmarking multimodal large language models on the dental licensing examination: challenges with clinical image interpretation. J Dent Sci. 2025;20(4):2427.
11. Shen Y, Xu Y, Ma J, Rui W, Zhao C, Heacock L, Huang C. Multi-modal large language models in radiology: principles, applications, and potential. Abdom Radiol. 2025;50(6):2745–57.
12. Ma J, Schneider L, Lapuschkin S, Achtibat R, Duchrau M, Krois J, Schwendicke F, Samek W. Towards trustworthy AI in dentistry. J Dent Res. 2022;101(11):1263–8.
13. Dasanayaka C, Dandeniya K, Dissanayake MB, Gunasena C, Jayasinghe R. Multimodal AI and Large Language Models for Orthopantomography Radiology Report Generation and Q&A. Appl Syst Innov. 2025;8(2):39.
14. Samaranayake L, Tuygunov N, Schwendicke F, Osathanon T, Khurshid Z, Boymuradov SA, Cahyanto A. The transformative role of artificial intelligence in dentistry: a comprehensive overview. Part 1: fundamentals of AI, and its

contemporary applications in dentistry. Int Dent J. 2025;75(2):383–96. https://doi.org/10.1016/j.identj.2025.02.005.
15. Tuygunov N, Samaranayake L, Khurshid Z, Rewthamrongsris P, Schwendicke F, Osathanon T, Yahya NA. The transformative role of artificial intelligence in dentistry: a comprehensive overview Part 2: the promise and perils, and the International Dental Federation Communique. Int Dent J. 2025;75(2):397–404. https://doi.org/10.1016/j.identj.2025.02.006.
16. Sohrabniya F, Hassanzadeh-Samani S, Ourang SA, Jafari B, Farzinnia G, Gorjinejad F, Ghalyanchi-Langeroudi A, Mohammad-Rahimi H, Tichy A, Motamedian SR. Exploring a decade of deep learning in dentistry: A comprehensive mapping review. Clin Oral Invest. 2025;29(2):1–18.
17. Pham TD. Classification of pediatric dental diseases from panoramic radiographs using natural language transformer and deep learning models. medRxiv. 2025;2025(2001):2030–25321418.
18. Brinz J, Eslamiamirabadi N, Salamati A, Tresp V, Schwendicke F, Tichy A. Data sharing for responsible artificial intelligence in dentistry: a narrative review of legal frameworks and privacy-preserving techniques. J Dent. 2025;163:106130. https://doi.org/10.1016/j.jdent.2025.106130.
19. Uribe SE, Schwendicke F. A primer on artificial intelligence research in dentistry. In: Schwendicke F, Chaudhari PK, Dhingra K, Uribe SE, Hamdan M, editors. Artificial intelligence for oral health care: applications and future prospects. Cham: Springer International Publishing; 2025:135–50. https://doi.org/10.1007/978-3-031-84047-0_9.
20. Yamaguchi S, Li H, Tichy A, Schwendicke F, Imazato S. Materials Informatics Opens the Door for Dental Materials Development. J Dent Res. 2025;104(11):1165–70.
21. Aşar EM, İpek İ, Bilge K. Customized GPT-4V (ision) for radiographic diagnosis: can large language model detect supernumerary teeth? BMC Oral Health. 2025;25(1):756.
22. Deveci BA, Kocuk M, Şahin M, Yanikoglu F. The performance of AI on the dentistry specialization exam. Int Dent J. 2024;74:S114–5.
23. Ali SA, AlDehlawi H, Jazzar A, Ashi H, Abuzinadah NE, AlOtaibi M, Algarni A, Alqahtani H, Akeel S, Almazrooa S. 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(1):e70566.
24. Quah B, Yong C, Lai C, Islam I. Performance of large language models in oral and maxillofacial surgery examinations. Int J Oral Maxillofac Surg. 2024;53(10):881–6.
25. Koidou VP, Chatzopoulos GS, Tsalikis L, Kaklamanos EG. Large language models in peri-implant disease: how well do they perform? J Prosthet Dent. 2025;134(6):2435–42. https://doi.org/10.1016/j.prosdent.2025.02.008.
26. Dave M, Tattar R, Alafaleg R, Barry S, Ariyaratnam S, Roudsari RV, Patel N. Performance of large language models (ChatGPT4-0, Grok2 and Gemini) in UK dentistry and dental hygiene and therapy assessments. Br Dent J. 2025. https://doi.org/10.1038/s41415-025-8383-2.
27. Albalawi F, Khanagar SB, Iyer K, Alhazmi N, Alayyash A, Alhazmi AS, Awawdeh M, Singh OG. Evaluating the Performance of Artificial Intelligence-Based Large Language Models in Orthodontics—A Systematic Review and Meta-Analysis. Appl Sci. 2025;15(2):893.
28. Chatzopoulos GS, Koidou VP, Tsalikis L, Kaklamanos EG. Evaluation of Large Language Model Performance in Answering Clinical Questions on Periodontal Furcation Defect Management. Dentistry J. 2025;13(6):271.
29. Tokgöz Kaplan T, Cankar M. Evidence-Based Potential of Generative Artificial Intelligence Large Language Models on Dental Avulsion: ChatGPT Versus Gemini. Dent Traumatol. 2025;41(2):178–86.
30. Chatzopoulos GS, Koidou VP, Tsalikis L, Kaklamanos EG. Large language models in periodontology: Assessing their performance in clinically relevant questions. J Prosthet Dent. 2025;134(6):2328–36. https://doi.org/10.1016/j.prosdent.2024.10.020.
31. Nguyen HC, Dang HP, Nguyen TL, Hoang V, Nguyen VA. Accuracy of latest large Language models in answering multiple choice questions in dentistry: A comparative study. PLoS ONE. 2025;20(1):e0317423.
32. Schiaffino S, Zhang T, Mann RM, Pinker K. The role of large language models (LLMs) in breast imaging today and in the near future. J Magn Reson Imaging. 2025;62(5):1296–304. https://doi.org/10.1002/jmri.29807.
33. Wu Y, Zeng L, Hong Y, Li X, Chen X. A real-time interactive restoration system for intraoral digital videos using segment anything model. Digit Health. 2024;10:20552076241269536.
34. Musri N, Christie B, Ichwan SJA, Cahyanto A. Deep learning convolutional neural network algorithms for the early detection and diagnosis of dental caries on periapical radiographs: A systematic review. Imaging Sci dentistry. 2021;51(3):237.
35. Magat G, Altındag A, Pertek Hatipoglu F, Hatipoglu O, Bayrakdar İS, Celik O, Orhan K. Automatic deep learning detection of overhanging restorations in bitewing radiographs. Dentomaxillofacial Radiol. 2024;53(7):468–77.
36. Szabó V, Orhan K, Dobó-Nagy C, Veres DS, Manulis D, Ezhov M, Sanders A, Szabó BT. Deep Learning-Based periapical lesion detection on panoramic radiographs. Diagnostics. 2025;15(4):510.
37. Esmaeili M, Dalili Z, Sadr H, Mousavie A, Faghihi A, Saei R, Nazari M. Hierarchical attention mechanism combined with deep neural networks for accurate semantic segmentation of dental structures in panoramic radiographs. Sci Rep. 2025;15(1):38725.
38. 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(15):1851.
39. Xiong Y-T, Lian W-J, Sun Y-N, Liu W, Guo J-X, Tang W, Liu C. Exploring GPT-4o's multimodal reasoning capabilities with panoramic radiograph: the role of prompt engineering. Clin Oral Invest. 2025;29(9):405.
40. Hassanein FE, Ahmed Y, Maher S, Barbary AE, Abou-Bakr A. Prompt-dependent performance of multimodal AI model in oral diagnosis: a comprehensive analysis of accuracy, narrative quality, calibration, and latency versus human experts. Sci Rep. 2025;15(1):37932.

Publisher's note

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