Title

Detecting Dental Caries Using General-Purpose Large Multimodal Models From Oral Photographs.

Journal

International dental journal, 76(5)

ISSN

0020-6539

Authors

Moharrami, Mohammad
Asadi, Sina
Farooqi, Owais
et al.

Publication Date

2026-07-01

DOI

10.1016/j.identj.2026.109735

Scientific Research Report

Detecting Dental Caries Using General-Purpose Large Multimodal Models From Oral Photographs

Mohammad Moharrami [a,b*], *Sina Asadi* [c], *Owais Farooqi* [b,d,e], *Maryam Amin* [f], *Michael Glogauer* [a], *Carlos Quinonez* [a,g], *Falk Schwendicke* [h]

[a] *Faculty of Dentistry, University of Toronto, Toronto, Ontario, Canada*
[b] *American Academy of Artificial Intelligence in Dentistry (AAAI-D), Los Angeles, USA*
[c] *Topic Group Dental Diagnostics and Digital Dentistry (TG-Dental), Focus Group on Artificial Intelligence for Health (FG-AI4H), ITU–WHO–WIPO, Geneva, Switzerland*
[d] *VA Greater Los Angeles Healthcare System, Los Angeles, CA, USA*
[e] *UCLA School of Dentistry, Los Angeles, CA, USA*
[f] *Mike Petryk School of Dentistry, Faculty of Medicine and Dentistry, University of Alberta, Edmonton, Alberta, Canada*
[g] *Schulich School of Medicine & Dentistry, Western University, Ontario, Canada*
[h] *Conservative Dentistry, Periodontology and Digital Dentistry, LMU Hospital, Munich, Germany*

ARTICLE INFO

Article history:
Received 23 March 2026
Received in revised form
4 June 2026
Accepted 11 June 2026
Available online xxx

Keywords:
Large multimodal models
Dental caries
Generative AI
Few-shot learning
Risk assessment

ABSTRACT

Objectives: To determine whether a general-purpose large multimodal model (LMM) can detect dental caries from intraoral photographs without task-specific training, evaluating image-level classification and tooth-level localisation under zero-shot (no reference examples) and five-shot (five annotated examples) conditions.

Methods: This diagnostic accuracy study used the benchmark test split of 1255 publicly available intraoral photographs. Gemini 3.1 reasoning and instant were queried via Vertex AI API using stateless calls and structured prompts for sequential, tooth-by-tooth scanning. Each model under different configurations was evaluated across 10 independent inference runs. Predictions (bounding boxes) were evaluated against expert annotations using greedy Intersection-over-Union (IoU $\geq$ 0.5). True positives required spatial overlap at the tooth level, or at least one correctly localised lesion at the image level. Performance was summarised using sensitivity, precision, F1-score, and mAP@50.

Results: Performance was strongly dependent on image view and model type, with reliable results observed mainly for the reasoning models on occlusal images. At the image level, zero-shot prompting showed high sensitivity but lower precision, yielding values of 0.95, 0.80, and 0.87 for sensitivity, precision, and F1-score, respectively. Five-shot prompting produced a more balanced profile, with corresponding values of 0.88, 0.87, and 0.87. At the tooth level, zero-shot reasoning showed a similar sensitivity-prioritised pattern, with sensitivity, precision, F1-score, and mAP@50 of 0.88, 0.49, 0.63, and 0.62, respectively. Five-shot prompting improved precision and produced a more balanced localisation profile, with corresponding values of 0.77, 0.57, 0.64, and 0.63.

Conclusions: General-purpose LMMs such as Gemini show potential as a scalable, automated caries screening tool for teledentistry without domain-specific fine-tuning. Although prompt engineering effectively modulates the sensitivity-precision trade-off, the model's tendency toward over-detection requires rigorous clinical validation and governance before real-world deployment as a public screening tool.

* *Corresponding author*. Faculty of Dentistry, University of Toronto, Toronto, ON, Canada, 124 Edward Street, Toronto, Ontario M5G 1X3, Canada.
E-mail address: faraz.moharrami@mail.utoronto.ca (M. Moharrami).

https://doi.org/10.1016/j.identj.2026.109735

Introduction

Dental caries remains one of the most prevalent non-communicable diseases worldwide, affecting individuals across all age groups and imposing a substantial biological and economic burden on society.[1] Despite its preventability, untreated caries can lead to severe complications, including pain, infection, and tooth loss, which exacerbate the strain on healthcare systems.[2] A critical challenge in managing this disease is the disparity in access to oral healthcare; care providers and clinics are often unevenly distributed, leaving populations in remote areas or low-income countries with limited availability to professional services.[3,4] Consequently, there is an urgent need for available, accessible, and cost-effective screening approaches that facilitate early detection and minimally invasive intervention, thereby bridging the gap in care for vulnerable populations.[5,6]

To address these access barriers, teledentistry enables remote screening using intraoral photographs captured by smartphones or professional cameras.[7] This process is increasingly being automated through the integration of artificial intelligence (AI) and deep learning.[8,9] Advanced algorithms, notably task-specific convolutional neural networks (CNNs) and transformers, demonstrate high accuracy in detecting caries and other anomalies from dental imagery.[8,10] These AI-driven systems not only standardise the screening process but also offer the potential for automated, real-time decision support, which is particularly valuable in field settings such as schools or home care environments where specialist expertise may be scarce.[11,12]

However, the widespread adoption of task-specific deep learning models is hindered by significant barriers, primarily the requirement for large, annotated datasets and the prohibitive time and capital required to train, deploy, and maintain separate models for every specific task or population group.[13,14] Furthermore, these systems often suffer from non-intuitive user interfaces and a reliance on unimodal data, typically yielding only visual overlays without integrated textual descriptions.[15] In contrast, Large Multimodal Models (LMMs) such as Gemini offer a transformative advantage by integrating visual analysis with advanced natural language reasoning. These models can generate detailed, context-aware reports and engage in interactive dialogue; this capability reduces the technical burden on end-users,[16–18] and potentially democratises access to AI-based diagnostics by the public without requiring specialised, resource-intensive infrastructure.

This study represents the first investigation into the screening capability of a general-purpose LMM, specifically Gemini, for the detection of dental caries from oral photographs. By moving beyond task-specific training, this research aims to evaluate Gemini's out-of-the-box zero-shot and five-shot learning performance for image-level classification and tooth-level object detection tasks. In doing so, we establish a standardised prompt framework and delineate the critical opportunities and limitations of deploying general LMMs for caries diagnosis.

Methods

This diagnostic accuracy study was prepared in accordance with the STAndards for the Reporting of Diagnostic accuracy studies – Artificial Intelligence (STARD-AI) reporting checklist.[19] A retrospective secondary analysis of an openly available dataset was conducted.

Ethics

No new human data were collected for this secondary analysis. The source dataset was collected under approval of the Ethical Review Committee of The Aga Khan University Hospital (AKUH) (ERC#2024-9433-29648), with written informed consent or assent for open sharing of anonymised data.

Data source

Images and reference annotations were obtained from a well-maintained open dataset published in Nature Scientific Data.[20] The dataset comprised intraoral photographs collected from individuals aged 10-24 years in Mithi, Sindh, Pakistan. The images were annotated using LabelMe and distributed in multiple annotation formats. The dataset includes photographs captured from multiple intraoral views, with and without cheek retractors, covering both mixed and permanent dentitions. Intraoral photographs were captured using a Samsung Galaxy A23 (Android 14.0), covering five views per set (maxillary occlusal, mandibular occlusal, left buccal, right buccal, frontal in habitual occlusion). To reduce variability, the source mobile application incorporated a built-in grid to standardise framing and aspect ratio, and images were captured with participants seated in a dental chair under standardised overhead dental-unit lighting. The dataset provides a benchmark partition with training/validation/test splits of 5011/627/628 images. For the purposes of the current study, the validation and test sets from the original study [20], were combined to create a new test set comprising 1,255 images.

Participants and eligibility criteria

Eligibility for this study was defined by inclusion in the source dataset. The dataset creators report inclusion of adolescents (10-18 years) and young adults (19-24 years) of both sexes. Exclusion criteria in the source dataset were: (1) suboptimal diagnostic image quality (e.g., blurred or obscured), (2) developmental dental anomalies (e.g., cleft lip or palate), (3) limited mouth opening, and (4) lack of consent.[20]

Reference standard (ground truth)

The reference standard consisted of bounding-box annotations identifying teeth with visible decay as a binary outcome. Annotations were performed by two calibrated dentists, with discrepancies resolved by a specialist dentist. Inter-rater reliability was assessed on 10% of the dataset and demonstrated high agreement (Cohen's $\kappa = 0.89$).[20]

Index test

The index test used the general-purpose LMM Gemini reasoning model (*gemini-3.1-pro-preview*) and instant model without reasoning (*gemini-3.1-flash-lite*) accessed via the Google GenAI SDK through a Vertex AI API.[21,22] Generation parameters

were left at the API defaults, including temperature = 1.0 and top-p = 0.95, to approximate the default behavior a typical user would experience in the Gemini Chat interface (https://gemini.google.com/app), while enabling systematic, reproducible, and scalable inference across the full test set. The model was configured to return plain-text outputs.

Zero-shot and five-shot (in-context) evaluations were conducted. For five-shot inference, the prompt included a set of reference images with visually drawn bounding boxes as contextual exemplars prior to the target image; for zero-shot inference, the target image was provided without reference examples. All analyses were executed May 5-20, 2026. The Python implementation used for API-based inference is available at this GitHub repository: https://github.com/fmhr12/LMM_Python_Pipeline and includes the full pipeline (prompts, programmatic image processing, automated extraction of model outputs, and computation of performance metrics).

Inference pipeline and prompt design

All analyses were performed using a Python pipeline that (1) loaded intraoral photographs from the benchmark test split, (2) submitted each image to Gemini via the Vertex AI API, (3) extracted bounding-box detections from the model's plain-text output, and (4) compared predictions against the dataset's reference annotations. To ensure experimental independence, the model was queried using stateless API calls for each image, preventing any context retention or memory carry-over that could influence subsequent predictions. To enable reproducible large-scale inference, images were processed sequentially, and progress was saved to a persistent checkpoint file that stored processed filenames, per-image detections, and cumulative performance statistics.

To prevent data leakage and avoid overfitting the prompt architecture to the test data, we maintained a strict separation between prompt optimisation and final testing. The iterative prompt-development process, which included more than 150 prompt variations to calibrate sensitivity and define the overall criteria, was conducted exclusively on a random subset of 50 occlusal and 50 buccal images drawn from the original dataset's separate training split [20]. Final prompt templates were selected using a mixed-methods approach: prompts were refined qualitatively to minimise false-positive hallucinations, such as misinterpreting calculus as caries, while also being quantitatively optimised for the highest stable F1-score on the training data. The benchmark test split, comprising 1255 images, remained completely isolated during this phase. Once the final zero-shot and five-shot prompt architectures were locked, they were deployed separately for occlusal and buccal views, for 10 independent inference runs on the unseen test set using stateless API calls to account for stochastic variability. Point estimates for all diagnostic accuracy metrics were calculated as the average across these 10 iterations, accompanied by their standard deviations.

For five-shot evaluation, five images with manually drawn bounding boxes were prepended to the request before the target image, providing in-context calibration for caries appearance and bounding-box tightness; for zero-shot evaluation, the target image was provided without exemplars. Prompting was designed to make the model behave as a structured "caries scanner" rather than a conversational assistant: the instruction specified an ICDAS-based rubric (ICDAS 0 = sound; ICDAS 1-6 = caries) for binarised caries detection.[23] required tooth-by-tooth scanning from left to right and enforced immediate reporting of each lesion using a fixed, machine-readable output line ([confidence, ymin, xmin, ymax, xmax]). Because caries prevalence and visual presentation differ by view, the model was run separately for occlusal and buccal views. To support consistent evaluation across variable image resolutions, both predictions and ground-truth annotations were expressed on the same normalised 0-1000 coordinate scale. The complete verbatim prompt templates (zero-shot and five-shot variants), including all role instructions, ICDAS descriptions, formatting constraints, and guardrails for caries detection, are detailed in the code scripts available at this GitHub repository: https://github.com/fmhr12/LMM_Python_Pipeline as well as Appendices 1 and 2.

Performance was summarised at two high levels. At the tooth level, predictions were ranked by confidence and greedily matched one-to-one with ground-truth boxes based on the highest IoU; matches with IoU $\geq$ 0.5 were counted as true positives (TP), unmatched predictions as false positives (FP), and unmatched ground-truth boxes as false negatives (FN). At the image level, images were labelled positive if any ground-truth caries was present; an image was TP only if it contained $\geq$ 1 correctly localised detection ($\geq$ 1 tooth-level TP), otherwise it was FN (including wrong-location predictions), while caries-free images were FP if any prediction occurred and TN if none occurred.

Diagnostic accuracy evaluation

Diagnostic accuracy was quantified from the accumulated TP/FP/FN (and TN for image-level) counts produced by the evaluation pipeline. Sensitivity (recall) was calculated as TP/(TP + FN), precision as TP/(TP + FP), and F1-score as 2*(precision*recall)/(precision + recall), with metrics reported separately for object-level (tooth) detection and image-level classification according to their respective confusion-matrix tallies. In addition, mAP@50 was computed at IoU = 0.5 by pooling all predictions across images, sorting them by confidence, labelling each as TP or FP based on the IoU matching outcome, constructing the resulting precision–recall curve over decreasing confidence thresholds, and integrating the area under the *interpolated* precision–recall curve to obtain average precision (single-class AP, reported as mAP@50).

Results

Of the 1255 images, 486 were occlusal views, of which 385 contained at least one carious tooth. The remaining 769 images were buccal views, of which 61 contained at least one carious tooth.

Image-level classification performance

At the occlusal view, the zero-shot reasoning model demonstrated high diagnostic performance, yielding a

sensitivity of 0.95 ± 0.04, a precision of 0.80 ± 0.05, and an F1-score of 0.87 ± 0.04. In the five-shot configuration (in-context learning), the reasoning model achieved a sensitivity of 0.88 ± 0.04, a precision of 0.87 ± 0.04, and an F1-score of 0.87 ± 0.03.

Similarly, at the occlusal view, the instant model in the zero-shot setting yielded a sensitivity of 0.79 ± 0.05, a precision of 0.78 ± 0.06, and an F1-score of 0.78 ± 0.04. When utilising the five-shot configuration, the instant model achieved a sensitivity of 0.78 ± 0.04, a precision of 0.85 ± 0.04, and an F1-score of 0.81 ± 0.03. The complete image-level performance metrics for occlusal-view caries detection are tabulated in Table 1.

For the buccal-view images, none of the evaluated configurations, including the instant and reasoning models under zero-shot and five-shot settings, achieved a sensitivity above 50% or a precision above 10% (Appendix 3).

Tooth-level object detection performance

The diagnostic capability to localise individual carious lesions was evaluated using IoU matching. At the tooth level, the reasoning model in the zero-shot configuration achieved a mean sensitivity of 0.88 ± 0.06, a mean precision of 0.49 ± 0.07, a mean F1 score of 0.63 ± 0.06, and a mean mAP@50 of 0.62± 0.04. In the five-shot configuration, the Reasoning model displayed a shift in screening behaviour, resulting in a mean sensitivity of 0.77 ± 0.04, a mean precision of 0.57 ± 0.06, a mean F1 score of 0.64 ± 0.05, and a mean mAP@50 of 0.63 ± 0.04.

For the instant model, the zero-shot configuration yielded a mean sensitivity of 0.58 ± 0.06, a mean precision of 0.22 ± 0.08, a mean F1 score of 0.34 ± 0.06, and a mean mAP@50 of 0.29 ± 0.04. Finally, the five-shot configuration for the instant model resulted in a mean sensitivity of 0.60 ± 0.05, a mean precision of 0.38 ± 0.07, a mean F1 score of 0.46 ± 0.05, and a mean mAP@50 of 0.33 ± 0.04. The complete tooth-level performance metrics for occlusal-view caries detection are tabulated in Table 1.

For the buccal-view images, none of the evaluated configurations, including the instant and reasoning models under zero-shot and five-shot settings, achieved a sensitivity above 40% or a precision above 8% (Appendix 3).

Model output

Figure 1A illustrates a representative input image submitted via either the Gemini chat interface (https://gemini.google.com/app) or an API call using the Python code (**GitHub repository**: https://github.com/fmhr12/LMM_Python_Pipeline). Following the prompt specified in Appendix 1, the model generated a text response delineating confidence score and bounding box coordinates for detected caries as follows: *"DETECTED: [0.95, 80, 85, 260, 245] DETECTED: . . . [End Scan]"*. These coordinates could then be visualised either by resubmitting the output string to the chat interface using the secondary prompt in Appendix 4, or by using a separate Python pipeline (**GitHub repository**: https://github.com/fmhr12/LMM_Python_Pipeline) to produce the prediction overlay in Figure 1B. For comparison, Figure 1C displays the ground truth annotations established by the dataset's expert dentists.

Discussion

To the best of our knowledge, this is the first study to successfully adapt a general-purpose LMM, specifically Gemini, for a standardised, reproducible, and scalable bounding-box object detection task in dental photography. A prior related work has explored the use of LMMs for object detection in a different dental imaging context, demonstrating relative success in implant fixture localisation on panoramic radiographs.[24] Precise object detection has traditionally been the exclusive domain of task-specific, supervised deep learning algorithms, such as CNN (e.g., YOLO, Faster R-CNN) or vision transformers, which require extensive, domain-specific training. By contrast, this work demonstrates that general-purpose models can be prompted to perform the localisation task on photographic data without any model fine-tuning or retraining.

Our results indicate that Gemini's performance was shaped by related factors, namely image view, in-context learning, and the intrinsic reasoning capability of the model. First, performance was strongly dependent on image view; results were mainly driven by occlusal images, where carious lesions were more frequent and visually more distinct, whereas buccal views showed poor performance across configurations, likely because smooth-surface lesions were less

Table 1 – Summary of the performance of the general-purpose Gemini models for occlusal caries detection under different configurations.

Model source	Model name	Sensitivity (recall)	Precision	F1-score	mAP @ 0.5 IoU
Image level	Gemini 3.1 Reasoning (Five-Shot)	0.88 ± 0.04	0.87 ± 0.04	0.87 ± 0.03	NA
Image level	Gemini 3.1 Reasoning (Zero-Shot)	0.95 ± 0.04	0.80 ± 0.05	0.87 ± 0.04	NA
Image level	Gemini 3.1 Instant (Five-Shot)	0.78 ± 0.04	0.85 ± 0.04	0.81 ± 0.03	NA
Image level	Gemini 3.1 Instant (Zero-Shot)	0.79 ± 0.05	0.78 ± 0.06	0.78 ± 0.04	NA
Tooth level	Gemini 3.1 Reasoning (Five-Shot)	0.77 ± 0.04	0.57 ± 0.06	0.64 ± 0.05	0.63 ± 0.04
Tooth level	Gemini 3.1 Reasoning (Zero-Shot)	0.88 ± 0.06	0.49 ± 0.07	0.63 ± 0.06	0.62 ± 0.04
Tooth level	Gemini 3.1 Instant (Five-Shot)	0.60 ± 0.05	0.38 ± 0.07	0.46 ± 0.05	0.33 ± 0.04
Tooth level	Gemini 3.1 Instant (Zero-Shot)	0.58 ± 0.06	0.22 ± 0.08	0.34 ± 0.06	0.29 ± 0.04

Abbreviations: IoU, intersection over union; mAP, mean average precision; NA, not applicable.
Image-level mAP is not reported because mAP@0.5 IoU is an object-detection metric calculated for tooth-level localisation.

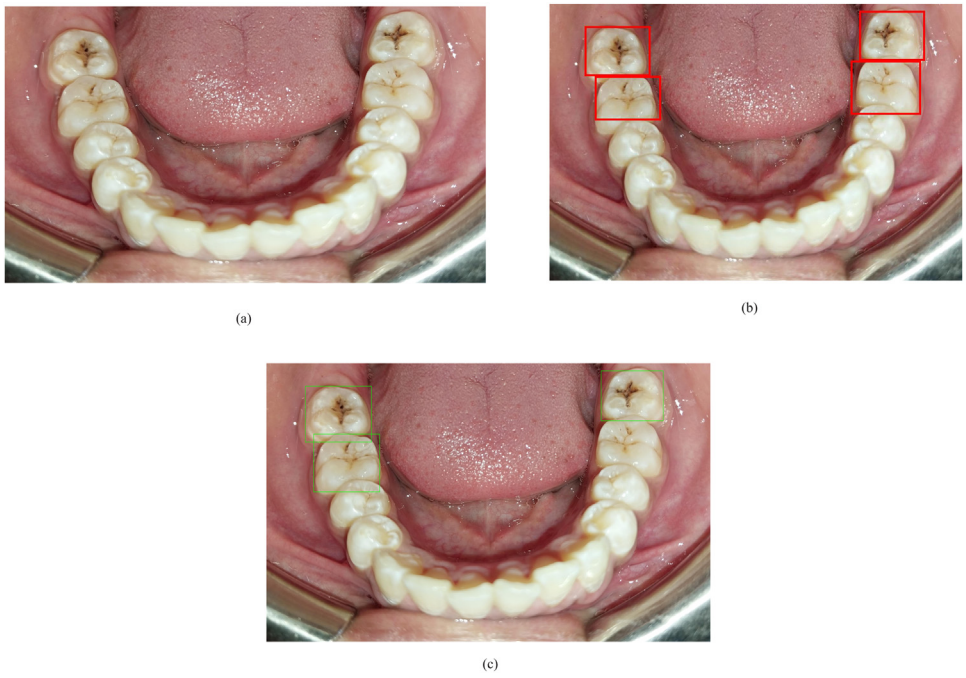

Fig. 1 – Workflow and outputs for automated caries detection using the pre-trained Gemini model.

(A) A representative input intraoral photograph submitted to the model via the Gemini chat interface or an API call. Based on the provided prompt, the model outputs text detailing confidence scores and bounding box coordinates for detected lesions.

(B) A visual prediction overlay demonstrating the bounding box coordinates generated from the model's plain-text output

(C) The corresponding ground truth bounding box annotations established by expert dentists, provided for comparison.

prevalent, more subtle, and more easily confused with plaque, calculus, stains, shadows, or soft-tissue artifacts. Second, five-shot in-context learning narrowed the sensitivity –precision imbalance because the annotated examples gave the model concrete visual references for what should count as caries and how tightly lesions should be localised; this likely shifted the model away from broad over-detection toward more selective predictions, reducing sensitivity but improving precision and producing a more balanced tooth-level detection profile. Third, because the same prompt structure was used across the reasoning and instant models, the observed performance difference is best attributed to the model architecture and inference behaviour rather than prompt design; the reasoning model was able to sustain a systematic tooth-by-tooth search, integrate diagnostic criteria with visual evidence, and avoid some impulsive false-positive detections, whereas the instant model produced weaker and less reliable localisation performance.

The comparison between the best five-shot reasoning Gemini model and task-specific benchmarks for tooth-level object detection on the same dataset provides a more cautious perspective on the current role of general-purpose LMMs in caries detection. The specialised YOLOv8 model achieved the highest F1-score of 0.81,[20] whereas the best Gemini model achieved an F1-score of 0.64 on the relatively easier occlusal-view images, without domain-specific fine-tuning. Although the Gemini model demonstrated relatively high sensitivity, comparable to some task-specific models,[20] this was accompanied by lower precision, indicating a higher false-positive burden. Therefore, its current performance profile appears more consistent with a preliminary screening or triage role than with definitive lesion-level diagnosis. Future work should explore hybrid workflows that leverage LMMs for initial screening and patient communication while relying on task-specific models for precise localisation.

Unlike traditional deep learning models, which are trained on narrowly curated datasets to minimise a specific loss function, such as the binary classification of caries presence, LMMs are built upon the transformer architecture,[25] and pre-trained on vast, internet-scale corpora comprising trillions of tokens of text, code, and images.[26] This extensive pre-training phase allows the model to learn generalised representations, rather than merely memorising task-specific patterns. When processing a dental image, the visual input is divided into fixed-size patches and encoded into a sequence of vectors that share a common embedding space with textual tokens.[27] Through self-attention mechanisms, the model dynamically weighs the relevance of visual features against the provided textual prompt, effectively "attending" to specific regions of interest within an intraoral photo while ignoring irrelevant noise. The decoding process is then generative rather than discriminative; the model predicts the most probable next token in a sequence, constructing a coherent, context-aware response one word at a time. Crucially, this generative capability extends to spatial localisation: the model predicts numerical coordinates as discrete text tokens, quantising the image into integer bins (e.g., [0, 1000]), which effectively allows it to speak the location of a bounding box around a lesion.[28] This shift enables LMMs to perform zero-shot or five-shot reasoning on unseen data without the need

for task-specific retraining, thereby overcoming the rigidity and data scarcity issues that plague conventional CNNs.

Beyond diagnostic metrics, the utilisation of general-purpose LMMs addresses critical economic and infrastructural barriers inherent to current task-specific solutions. While specialised deep learning models demonstrate high efficacy, they are predominantly commercialised as proprietary, subscription-based services targeting well-resourced private practices. This distribution model restricts advanced diagnostic capabilities to clinical environments with sufficient capital, effectively excluding underserved populations and public health initiatives. In contrast, general-purpose LMMs are rapidly evolving into ubiquitous utilities that decouple screening from exclusive software ecosystems. This accessibility facilitates a shift toward direct-to-patient screening and teledentistry, enabling individuals in remote or resource-limited settings to obtain preliminary assessments via standard mobile devices. Moreover, the capacity of these models to provide natural language explanations alongside visual detections enhances health literacy, empowering individuals to seek timely professional care based on automated triage rather than advanced disease symptoms.

The optimisation of the LMM prompt for this specific domain necessitated an extensive iterative process, evolving from simple linear commands to the final structured architecture. Our experimentation highlighted that general-purpose LMMs require explicit, step-by-step navigational instructions, specifically to scan dentition sequentially and isolate individual teeth, to effectively prevent tooth omission and accurately distinguish interproximal boundaries. Regarding diagnostic criteria, while embedding general ICDAS definitions improved performance, accuracy was maximised when using the verbatim descriptions of the simplified 3-stage classification (initial, moderate, extensive) rather than the complex full 6-code ICDAS system.[23] Paradoxically, providing visual cues, such as reference images of the ICDAS rubric, failed to enhance detection capabilities. We further observed that minor prompt adjustments induced substantial shifts in the trade-off between sensitivity and precision, although the overall F1-score remained relatively stable. Qualitative error analysis revealed that false negatives were frequently associated with compromised image quality, whereas a subset of false positives stemmed from localisation errors where the bounding box was placed on the ambiguous border between teeth. Most notably, the integration of a recheck loop within the prompt structure, instructing the model to first screen a tooth with high sensitivity and then immediately self-validate with high precision before confirming, served as a critical mechanism for reducing false positives and improving overall performance.

A primary limitation of this study is the model's susceptibility to hallucinations, particularly in the zero-shot setting, where it demonstrated a tendency to over-interpret shadows or stains as caries. This false-positive tendency has practical clinical implications, as it may increase unnecessary referrals, clinician verification workload, patient anxiety, and loss of trust if outputs are interpreted as diagnoses rather than preliminary screening results. Although the dataset was used only for evaluation and no dental fine-tuning was performed, this study should be interpreted as an independent technical benchmark rather than a comprehensive external clinical validation. The analysis was limited to a single public dataset from one geographic region, and the exclusion of low-quality images; therefore, performance may not generalise to other populations, devices, or real-world image-quality variation.

To address these challenges, future research should test these models across more diverse populations, age groups, and imaging conditions. Future studies could also evaluate LMM-generated explanations using dedicated qualitative or mixed-methods designs to assess their consistency and coherence; however, such post hoc explanations should be interpreted cautiously, as they may not faithfully reflect the model's internal decision process. Reader-performance studies should also compare LMM outputs with those of human screeners, such as dental students or general practitioners, using the same images and standardised blinded assessment conditions. Although IoU ≥ 0.5 was used as the primary localisation threshold, performance estimates may differ under alternative thresholds; lower thresholds would allow more loosely localised detections to count as correct, whereas higher thresholds would require more precise agreement with expert annotations. Future studies should therefore assess LMM performance across multiple IoU thresholds to better characterise localisation robustness. Furthermore, the demonstrated potential of few-shot learning suggests that fine-tuning, combined with advanced prompt engineering strategies such as structured reasoning, may furthur improve the balance between sensitivity and precision. Ultimately, as LMMs become integrated into healthcare, the development of robust regulatory frameworks is essential to ensure transparency, manage liability, and verify safety before these tools are entrusted with patient care in unsupervised settings.

Conclusions

Overall, this study demonstrates that a general-purpose LMM such as Gemini can be prompted without domain-specific fine-tuning to support dental caries screening from intraoral photographs, achieving strong image-level detection while providing moderate tooth-level localisation, mainly for occlusal views. Performance was prompt-dependent, with zero-shot prompting prioritising sensitivity, whereas five-shot in-context exemplars reduced false-positive detections and produced a more balanced precision–recall trade-off, highlighting prompt design as a practical lever for calibrating model behavior. Although these findings suggest a feasible pathway toward scalable, low-burden teledentistry triage and patient-facing decision support, the model's tendency toward over-detection, potentially exacerbated by real-world imaging variability, underscores the need for prospective validation, population- and device-level robustness assessments, and appropriate governance frameworks, including clinical oversight, regulatory compliance, and ethical and data-privacy safeguards, prior to any unsupervised deployment.

Author Contributions

Mohammad Moharrami: Conceptualisation, Methodology, Software, Formal analysis, Data curation, Investigation,

Writing – original draft. **Sina Asadi:** Validation, Formal analysis, Writing – review & editing. **Owais Farooqi:** Validation, Writing – review & editing. **Maryam Amin:** Conceptualisation, Writing – review & editing. **Michael Glogauer:** Resources, Supervision, Writing – review & editing. **Carlos Quinonez:** Supervision, Project administration, Writing – review & editing. **Falk Schwendicke:** Conceptualisation, Methodology, Supervision, Writing – review & editing.

Conflict of interest

None disclosed.

Supplementary materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.identj.2026.109735.

REFERENCES

1. Wen PYF, Chen MX, Zhong YJ, Dong QQ, Wong HM. Global burden and inequality of dental caries, 1990 to 2019. J Dent Res 2022;101:392–9.
2. James SL, Abate D, Abate KH, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. The Lancet 2018;392:1789–858.
3. Peres MA, Macpherson LMD, Weyant RJ, et al. Oral diseases: a global public health challenge. The Lancet 2019;394:249–60.
4. World Health Organization. Global oral health status report: towards universal health coverage for oral health by 2030. Geneva: World Health Organization; 2022.
5. Schwendicke F, Rossi JG, Göstemeyer G, et al. Cost-effectiveness of artificial intelligence for proximal caries detection. J Dent Res 2021;100:369–76.
6. Desai H, Stewart CA, Finer Y. Minimally invasive therapies for the management of dental caries—a literature review. Dent J (Basel) 2021;9:147.
7. Kargozar S, Jadidfard M-P. Teledentistry accuracy for caries diagnosis: a systematic review of in-vivo studies using extra-oral photography methods. BMC Oral Health 2024;24:828.
8. Moharrami M, Farmer J, Singhal S, et al. Detecting dental caries on oral photographs using artificial intelligence: a systematic review. Oral Dis 2024;30:1765–83.
9. Moharrami M, Vahab E, Bagherianlemraski M, et al. Deep learning for detecting dental plaque and gingivitis from oral photographs: a systematic review. Community Dent Oral Epidemiol 2025;53:617–32.
10. Mohammad-Rahimi H, Motamedian SR, Rohban MH, et al. Deep learning for caries detection: a systematic review. J Dent 2022;122:104115.
11. Kim D, Choi J, Ahn S, Park E. A smart home dental care system: integration of deep learning, image sensors, and mobile controller. J Ambient Intell Humaniz Comput 2021. doi: 10.1007/s12652-021-03366-8.
12. Felsch M, Meyer O, Schlickenrieder A, et al. Detection and localization of caries and hypomineralization on dental photographs with a vision transformer model. NPJ Digit Med 2023;6:198.
13. Liu T-Y, Lee K-H, Mukundan A, Karmakar R, Dhiman H, Wang H-C. AI in dentistry: innovations, ethical considerations, and integration barriers. Bioengineering 2025;12:928.
14. Holtkamp A, Elhennawy K, Cejudo Grano de Oro JE, Krois J, Paris S, Schwendicke F. Generalizability of deep learning models for caries detection in near-infrared light transillumination images. J Clin Med 2021;10:961.
15. Mai H, Lee D, Kaenploy J, Kim J, Cho S. Performance of multimodal generative AI models in addressing complex dental inquiries with text, images, and analytical data. J Esthet Restor Dentist 2025;38:166–72.
16. 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;8:39.
17. Arpaci A, Ozturk AU, Okur I, Sadry S. Evaluation of the accuracy of ChatGPT-4 and Gemini's responses to the World Dental Federation's frequently asked questions on oral health. BMC Oral Health 2025;25:1293.
18. Samaranayake L, Tuygunov N, Schwendicke F, et al. 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;75:383–96.
19. Sounderajah V, Guni A, Liu X, et al. The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. Nat Med 2025;31:3283–9.
20. Faizan Ahmed SM, Ghori MH, et al. Annotated intraoral image dataset for dental caries detection. Sci Data 2025;12:1297. doi: 10.5281/zenodo.14827784.
21. Google Cloud. Generative AI SDKs overview. Internet. Vertex AI Generative AI documentation 2026 https://docs.cloud.google.com/vertex-ai/generative-ai/docs/sdks/overview?utm_source=chatgpt.com (Accessed 30 January 2026).
22. Google Cloud. Gemini 3 Pro. Internet. Vertex AI Generative AI documentation 2026 https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/3-pro?utm_source=chatgpt.com (Accessed 30 January 2026).
23. NB Pitts, AI Ismail, S Martignon, et al., ICCMSTM guide for practitioners and educators, (2014).
24. Mine Y, Taji T, Takeda S, et al. Assessing multimodal large language models for localizing dental implant fixtures on panoramic radiographs. J Dent 2026;168:106580.
25. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Adv Neural Inf Process Syst 2017;30.
26. G Team, R Anil, S Borgeaud, et al., Gemini: a family of highly capable multimodal models, ArXiv Preprint ArXiv:2312.11805 (2023).
27. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: transformers for image recognition at scale. Intern Confe Learn Repres 2021.
28. Chen T, Saxena S, Li L, Fleet DJ, Hinton G. Pix2seq: a language modeling framework for object detection. Intern Confe Learn Repres 2022.