applied sciences

MDPI

Article

Exploratory Evaluation of Diagnostic Accuracy and Temporal Reproducibility of Multimodal Large Language Models in the Image-Based Assessment of Oral Mucosal Lesions

Lovro Dumančić [1], Marko Antonio Cug [1], Danica Vidović Juras [1,2], Luís Monteiro [3], Rui Albuquerque [4,5] and Vlaho Brailo [1,2,*]

[1] Department of Oral Medicine, University of Zagreb School of Dental Medicine, 10000 Zagreb, Croatia; ldumancic@sfzg.unizg.hr (L.D.); mcug@sfzg.unizg.hr (M.A.C.); djuras@sfzg.unizg.hr (D.V.J.)
[2] Department of Oral Medicine, Clinic for Dentistry, University Clinical Hospital Centre Zagreb, 10000 Zagreb, Croatia
[3] UNIPRO—Oral Pathology and Rehabilitation Research Unit, University Institute of Health Sciences (IUCS CESPU—Instituto Universitário de Ciências da Saúde, Cooperativa De Ensino Superior Politecnico E Universitario), 4585-116 Gandra, Portugal
[4] Department of Oral Medicine, Guy's and St Thomas NHS Foundation Trust, London SE1 9RT, UK; rui.albuquerque@nhs.net
[5] Faculty of Dentistry, Oral and Craniofacial Sciences, King's College London, London SE1 9RT, UK
* Correspondence: brailo@sfzg.unizg.hr

Abstract

Objective: The aim was to evaluate the diagnostic accuracy and temporal reproducibility of multimodal large language models (LLMs) in the image-based diagnosis of oral mucosal lesions. Materials and Methods: The study included 100 anonymized clinical photographs of oral mucosal conditions obtained from the archive of the Department of Oral Medicine, School of Dental Medicine, University of Zagreb. Images were categorized into four subgroups: physiological variations, benign mucosal lesions, oral potentially malignant disorders, and oral cancer (25 images each). Three multimodal LLMs (ChatGPT-5.1 Plus, Gemini 3 Pro, and Perplexity Pro) analyzed each image using an identical prompt and were required to provide a single most probable diagnosis based solely on visual features. To evaluate temporal reproducibility, the entire evaluation was repeated in three independent testing cycles conducted at one-month intervals. Diagnostic accuracy was compared using chi-square tests, while intra-model agreement across cycles was assessed using Fleiss' kappa. Results: Gemini demonstrated the highest diagnostic accuracy, reaching 78% correct responses in cycles 2 and 3, significantly outperforming ChatGPT (55–57%) and Perplexity (28–31%) ($p < 0.00001$). Subgroup analyses showed similar trends, with Gemini achieving the highest accuracy across most lesion categories. Intra-model agreement across cycles was moderate for ChatGPT ($\kappa = 0.525$), fair for Gemini ($\kappa = 0.338$) and Perplexity ($\kappa = 0.409$). Gemini also showed the highest proportion of responses that remained correct across all three cycles (51%). Conclusions: Multimodal LLMs demonstrate promising diagnostic capabilities in the image-based assessment of oral mucosal lesions; however, variability in reproducibility highlights the need for cautious clinical implementation and further validation.

Keywords: artificial intelligence; multimodal large language models; oral lesion; diagnosis; reproducibility

Check for updates

Academic Editor: Iole Vozza

Received: 26 March 2026
Revised: 15 April 2026
Accepted: 19 April 2026
Published: 21 April 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Clinical assessment in oral medicine fundamentally relies on the visual interpretation of morphological features of oral mucosa, making it an area where artificial intelligence-driven image analysis may offer meaningful diagnostic support [1]. Recent advances in artificial intelligence (AI) have demonstrated promising applications across various dental specialties, with convolutional neural networks (CNNs) achieving high diagnostic accuracy for oral potentially malignant disorders (OPMDs) and oral cancer (OC) in clinical photographs [2]. Meta-analyses have shown that AI-based detection of oral mucosal lesions achieves a pooled sensitivity of 89.9% and specificity of 89.2%, with diagnostic odds ratios ranging from 49 to 155 depending on lesion type [3,4].

Unlike traditional CNNs that are specifically trained on domain-specific datasets, multimodal large language models (LLMs) are foundation models trained on vast amounts of diverse data, enabling them to process both textual and visual information simultaneously [5,6]. In spite of being less specific than traditional CNNs, LLMs are extensively investigated across different areas of medicine due to their wider availability and lower price compared to more specialized AI models [7–11]. LLMs have exhibited promising image interpretation capabilities across multiple medical domains. Recent studies demonstrated LLMs achieving diagnostic accuracy between 58% and 97% when interpreting oral pathology images, with performance varying substantially between different models and clinical contexts [12–14].

Beyond diagnostic accuracy, the successful integration of AI into clinical workflow requires consistent, reliable, and stable performance over time. In real-world clinical settings, automation and workflow stability are critical, as variability in model outputs may directly influence clinical decision-making and user trust. Increasing evidence suggests that the clinical impact of AI systems depends not only on their performance under controlled conditions but also on their robustness, reproducibility, and ability to operate reliably within dynamic healthcare environments [5,6,15,16]. Moreover, recent studies evaluating multimodal LLMs have highlighted performance variability and temporal fluctuations as important limitations, particularly in everyday diagnostic applications [8,10,11]. These challenges are further emphasized in broader discussions on the clinical implementation and regulation of AI systems, which underline the importance of transparency, consistency, and ongoing validation in healthcare settings [17–19]. Such considerations are especially relevant for continuously evolving systems such as LLMs, where probabilistic output generation and ongoing model updates may introduce temporal variability in clinical use.

Despite these promising results, the diagnostic reliability of multimodal LLMs in oral mucosal lesion assessment remains limited. Most existing AI research in oral medicine has focused on task-specific CNNs rather than general-purpose multimodal models [2,20,21]. Furthermore, a significant gap exists in understanding the reproducibility and stability of LLM-generated diagnostic outputs over time. While traditional machine learning models should theoretically produce identical outputs when presented with the same input, the stochastic nature of LLM architectures and potential model updates raise concerns about consistency [13,14]. Studies in oral pathology have reported variable intra-model consistency, with Cohen's kappa values ranging from 0.740 to 0.918 across different LLMs, yet evaluation of temporal reproducibility remains insufficiently investigated [13,14].

Therefore, the aim of this study was to evaluate the potential application of selected LLMs as supportive diagnostic tools in everyday clinical settings by assessing both their diagnostic accuracy and the reproducibility of their outputs over time.

2. Materials and Methods

The study protocol was approved by the Institutional Review Board of the University of Zagreb School of Dental Medicine on 12 November 2025 (Approval No. KLASA: 03-01/25-05/13; No: 251-60-4/119-7). All procedures were conducted in accordance with the Declaration of Helsinki.

This study sample included 100 clinical images of oral mucosal lesions obtained from the image archive of the Department of Oral Medicine, School of Dental Medicine, University of Zagreb. Images were collected in a period from 1 January 2010 to 31 October 2025. A formal a priori sample size calculation was not performed, as the study was designed as an exploratory evaluation of diagnostic accuracy and temporal reproducibility rather than hypothesis testing. However, the sample size was comparable to similar studies in the literature [13,14,22,23].

Clinical photographs were captured with two devices—camera (Konica Minolta DiMAGE A2, Konica Minolta Inc., Tokyo, Japan) and mobile phone (Samsung A54, Samsung Electronics Inc., Suwon-si, South Korea) while maintaining consistent lighting and centering the oral lesions. All images were fully anonymized prior to analysis, and no identifying patient information was visible. Only images of adequate clarity (i.e., the entity of interest in the focus) were included. No other technical restrictions were applied to the photograph selection. This strategy was purposefully adopted to better reflect hypothetical real-world scenarios in which clinicians capture images of oral lesions with different devices and then consult LLM for diagnostic support.

Each image corresponded to a single oral mucosal lesion. For every case, the reference diagnosis was determined independently by two senior oral medicine specialists from the Department of Oral Medicine, School of Dental Medicine, University of Zagreb, each with more than 20 years of clinical experience. Complete agreement was achieved between the two examiners before the images were entered into the AI programs. These consensus diagnoses were used as the reference standard for comparison with the AI-generated results.

Photographs were classified into four subgroups of oral mucosal conditions: subgroup 1 (S1): physiological or morphological variations (n = 25); subgroup 2 (S2): non-malignant oral mucosal lesions (n = 25); subgroup 3 (S3): oral potentially malignant disorders (OPMDs) (n = 25); and subgroup 4 (S4): oral cancer (n = 25).

Three commercially available multimodal artificial intelligence systems were included in the analysis: ChatGPT-5.1 Plus (OpenAI Inc., San Francisco, CA, USA), Perplexity Pro (Perplexity AI Inc., San Francisco, CA, USA), and Gemini 3 Pro (Google Inc., Mountain View, CA, USA). All models were accessed using their professional versions at the time of testing (1 November 2025–31 January 2026) and operated under default system settings. Settings permitting the use of user-submitted data for model improvement were disabled to ensure that the uploaded clinical images were not incorporated into future learning processes. This precaution helped ensure that the evaluation reflected the models existing capabilities rather than potential adaptations resulting from the study data. An identical protocol was applied across all three AI systems to ensure methodological consistency. Each clinical image was analyzed within a newly initiated and independent chat session to prevent contextual memory effects and minimize potential bias resulting from prior interactions.

A single standardized prompt was used throughout the evaluation process. Each LLM was instructed to provide one single most probable definitive diagnosis based solely on the clinical image of the oral mucosal lesion, without including differential diagnoses or additional explanations. The following prompt modified from the study of Vaira et al. [23] was used:

> "I am an oral medicine specialist with a decade of experience in oral mucosal lesions. A colleague has sent me an image of an oral mucosal lesion in a patient,

asking for my help with the diagnosis. You are the world's leading expert in oral medicine. Analyze the attached image, assess whether mucosal lesions are present, and, if so, provide only one diagnosis that you consider the most probable. If no lesions are detected, state it explicitly. Use professional language. The information must be based on the most recent and solid scientific evidence."

No clinical history, symptom description, or additional contextual information was provided during AI-based image evaluation in order to standardize assessment conditions and ensure that the analysis relied exclusively on visual morphological features.

To evaluate the temporal stability and reproducibility of AI-generated outputs, the complete evaluation using the same prompt was repeated in three independent analytical cycles conducted at one-month intervals. In each cycle, all 100 images were re-analyzed and each image was evaluated in a newly initiated and independent chat session using the same model versions and identical prompting protocol. This design enabled assessment of potential variability in AI-generated diagnostic outputs over time.

Statistical Analysis

All responses were registered and subsequently coded in an Excel spreadsheet (Microsoft Excel®, Microsoft Inc., Redmond, WA, USA). Statistical package for social sciences version 29 (SPSS® ver. 29, IBM Inc., Armonk, NY, USA) was used for statistical analysis.

Data are presented as absolute number (N) and proportion of correct answers (%) with 95% confidence interval (95% CI). The chi square test was used to assess the difference in the proportion of correct responses between different AI models in each cycle of testing. In case of a significant difference between models, post hoc pairwise comparisons with Bonferroni correction were applied. Cramer's V was used as a measure of an effect size as follows: 0.07–0.21—small effect; 0.21–0.35—moderate effect; >0.35 large effect.

To assess the temporal stability and reproducibility of the responses, two sets of analysis were performed. First, the intra-model agreement was assessed using Fleiss kappa statistics as follows: $\leq$0 as no agreement; 0.01–0.20 none to slight agreement; 0.21–0.40 fair agreement; 0.41–0.60 moderate agreement; 0.61–0.80 substantial agreement; and 0.81–1.00 almost perfect agreement [24]. Second, the proportion of answers which remained correct in all three cycles of testing was calculated and their difference between models was assessed by a chi square test. In all analyses, p values lower than 0.05 ($p < 0.05$) were considered statistically significant.

A formal a priori sample size calculation was not performed, as the study was designed as an exploratory evaluation rather than a hypothesis testing analysis. However, a post hoc estimation using G*power 3.1 software [25,26] was conducted. The estimation was based on a two-sided comparison of proportions (Z-test), assuming a significance level of 0.05 and a power of 80%. Based on the observed differences in diagnostic accuracy between models, approximately 70–90 cases per group would be required to detect a difference of around 20%. Given that each model was evaluated on 100 images, the study was adequately powered to detect moderate-to-large differences in performance. Additionally, the repeated-measures design for the assessment of reproducibility (100 images $\times$ 3 models $\times$ 3 cycles) further increased the robustness of the analysis.

3. Results

3.1. Diagnostic Accuracy Testing

Results of a diagnostic accuracy testing are presented in Table 1. Overall, Gemini had the highest proportion of correct responses (63% (52.8–72.4%)–78% (68.6–85.7%)) compared to ChatGPT (55% (44.7–65%)–58% (47.7–67.8%)) and Perplexity (31% (22.1–41%)–50% (39.8–60.2%)). No significant difference in the proportion of correct responses between AI

models was observed in the cycle 1. In cycles 2 and 3, Gemini had a significantly higher proportion of correct responses compared to ChatGPT and Perplexity (78% (68.6–85.7%) vs. 55% (44.7–65%)/57% (46.7–66.9%) and 28% (19.5–37.9%)/31% (22.1–41%); $p < 0.00001$; V = 0.410/0.387) (Table 1).

Table 1. Diagnostic accuracy of tested large language models (LLMs).

		ChatGPT	Gemini	Perplexity	*p*	Effect Size	Post Hoc Bonferroni Correction
Normal anatomy and variations (S1)							
Cycle 1	N	17	15	12	0.352	0.167	
	% (95% CI)	68% (46.5–85.1%)	60% (38.7–78.9%)	48% (27.8–68.7%)			
Cycle 2	N	15	23	10	0.001 *	0.446	Gemini > ChatGPT; $p < 0.0001$ * Gemini > Perplexity; $p < 0.0001$ *
	% (95% CI)	60% (38.7–78.9%)	92% (74–99%)	40% (21.1–61.3%)			
Cycle 3	N	16	22	11	0.005 *	0.378	Gemini > Perplexity; $p < 0.0001$ *
	% (95% CI)	64% (42.5–82%)	88% (68.8–97.5%)	44% (24.4–65.1%)			
Benign mucosal lesions (S2)							
Cycle 1	N	15	15	12	0.614	0.114	
	% (95% CI)	60% (38.7–78.9%)	60% (38.7–78.9%)	48% (27.8–68.7%)			
Cycle 2	N	18	18	5	0.0001 *	0.492	Gemini > Perplexity; $p < 0.0001$ * ChatGPT> Perplexity; $p < 0.0001$ *
	% (95% CI)	72% (50.6–87.9%)	72% (50.6–87.9%)	20% (6.8–40.7%)			
Cycle 3	N	12	17	6	0.008 *	0.361	Gemini > Perplexity; $p < 0.0001$ *
	% (95% CI)	48% (27.8–68.7%)	68% (46.5–85.1%)	24% (9.4–45.1%)			
Oral potentially malignant disorders (S3)							
Cycle 1	N	13	16	12	0.497	0.137	
	% (95% CI)	52% (31.3–72.2%)	64% (42.5–82%)	48% (27.8–68.7%)			
Cycle 2	N	11	17	5	0.003 *	0.395	Gemini > Perplexity; $p = 0.0027$ *
	% (95% CI)	44% (24.4–65.1%)	68% (46.5–85.1%)	20% (6.8–40.7%)			
Cycle 3	N	12	17	6	0.008 *	0.361	Gemini > Perplexity; $p = 0.0051$ *
	% (95% CI)	48% (27.8–68.7%)	68% (46.5–85.1%)	24% (9.4–45.1%)			

Table 1. *Cont.*

		ChatGPT	Gemini	Perplexity	*p*	Effect Size	Post Hoc Bonferroni Correction
Oral cancer (S4)							
Cycle 1	N	13	17	14	0.489	0.138	
	% (95% CI)	52% (31.3–72.2%)	68% (46.5–85.1%)	56% (34.9–75.6%)			
Cycle 2	N	11	20	8	0.002 *	0.408	Gemini > Perplexity; $p = 0.0007$ *
	% (95% CI)	44% (24.4–65.1%)	80% (59.3–93.2%)	32% (14.9–53.5%)			
Cycle 3	N	14	21	7	0.0004 *	0.461	Gemini > Perplexity; $p = 0.0005$ *
	% (95% CI)	56% (34.9–75.6%)	84% (63.95.5%)	28% (12.1–49.4%)			
Total N; % (95% CI)							
Cycle 1	N	58	63	50	0.173	0.108	
	% (95% CI)	58% (47.7–67.8%)	63% (52.8–72.4%)	50% (39.8–60.2%)			
Cycle 2	N	55	78	28	$p < 0.0001$ *	0.410	ChatGPT > Perplexity; $p < 0.0001$ * Gemini > Perplexity, $p < 0.0001$ * Gemini > ChatGPT; $p < 0.0001$ *
	% (95% CI)	55% (44.7–65%)	78% (68.6–85.7%)	28% (19.5–37.9%)			
Cycle 3	N	57	78	31	$p < 0.0001$ *	0.387	ChatGPT > Perplexity; $p < 0.0001$ * Gemini > Perplexity, $p < 0.0001$ * Gemini > ChatGPT; $p < 0.0001$ *
	% (95% CI)	57% (46.7–66.9%)	78% (68.6–85.7%)	31% (22.1–41%)			

* significant difference ($p < 0.05$).

Subgroup analysis displayed a similar trend. In cycle 1 no significant difference in correct responses between AI models was seen for any of the four subgroups. In cycle 2, Gemini had the highest proportion of correct responses in three out of the four subgroups (S1—92% (74–99%); S3—68% (46.5–85.1%); and S4 80% (59.3–93.2%)) while in S2 both Gemini and ChatGPT had a similar proportion of correct responses (72% (50.6–87.9%)) (Table 1).

In cycle 2, Gemini had a significantly higher proportion of correct responses compared to Perplexity in all four subgroups. In S1, Gemini had a significantly higher proportion of correct responses compared to ChatGPT, as well. In S2, ChatGPT also had a significantly higher proportion of correct responses compared to Perplexity (Table 1).

In cycle 3, Gemini had the highest proportion of correct responses in all four subgroups (S1—88% (68.8–97.5%); S2—68% (46.5–85.1%); S3—68% (46.5–85.1%); and S4 84% (63.9–95.5%)). The difference between Gemini and Perplexity was statistically significant in all four subgroups (Table 1).

3.2. Reliability Assessment

Results of the reliability assessment are presented in Table 2. Overall intra-model agreement between three cycles of testing was moderate for ChatGPT (0.525; (0.412–0.638))

and fair for Gemini (0.338; (0.223–0.452)) and Perplexity (0.409; (0.296–0.522)). For S1 intra-model agreement was moderate for ChatGPT (0.537; (0.311–0.763)) and Perplexity (0.513; (0.287–0.739)), and fair for Gemini (0.167; (−0.06–0.393)). For S2 intra-model agreement was moderate for ChatGPT (0.479; (0.253–0.705)) and Gemini (0.542; (0.311–0.773)), and fair for Perplexity (0.247; (0.021–0.474)). For S3, intra-model agreement was substantial for ChatGPT (Fleiss Kappa 0.679; (0.453–0.906)), moderate for Gemini (0.542; (0.354–0.806)) and fair for Perplexity (0.247; (0.021–0.474)). For S4 intra-model agreement was fair for ChatGPT (0.360; (0.134–0.586)), poor for Gemini (0.011; (−0.215–0.237)) and moderate for Perplexity (0.438; (0.211–0.664)) (Table 2).

Table 2. Reliability assessment of tested large language models (LLMs).

	ChatGPT	***p***	**Gemini**	***p***	**Perplexity**	***p***
Intra-model agreement between three cycles of testing (Fleiss Kappa; (95% CI))						
Normal anatomy and variations (S1)	0.537; (0.311–0.763)	$p < 0.0001$ *	0.167; (−0.06–0.393)	0.169	0.513; (0.287–0.739)	$p < 0.0001$ *
Benign mucosal lesions (S2)	0.479; (0.253–0.705)	$p < 0.0001$ *	0.542; (0.311–0.773)	$p < 0.0001$ *	0.247; (0.021–0.474)	$p < 0.0001$ *
Oral potentially malignant disorders (S3)	0.679; (0.453–0.906)	$p < 0.0001$ *	0.542 (0.354–0.806)	$p < 0.0001$ *	0.247; (0.021–0.474)	0.032
Oral cancer (S4)	0.360; (0.134–0.586)	0.002 *	0.011; (−0.215–0.237)	0.923	0.438; (0.211–0.664)	$p < 0.0001$ *
Total	0.525; (0.412–0.638)	$p < 0.0001$	0.338; (0.223–0.452)	$p < 0.0001$ *	0.409; (0.296–0.522)	$p < 0.0001$ *

Correct responses in all three cycles of testing		ChatGPT	Gemini	Perplexity	*p*	Effect size	Post hoc Bonferroni correction
Normal anatomy and variations (S1)	N	12	14	6	0.059	0.275	
	% (95% CI)	48% (27.8–68.7%)	56% (34.9–75.6%)	24% (9.4–45.1%)			
Benign mucosal lesions (S2)	N	11	13	3	0.008 *	0.360	Gemini > Perplexity; $p = 0.0019$ *
	% (95% CI)	44% (24.4–65.1%)	52% (31.3–72.2%)	12% (2.5–31.2%)			
Oral potentially malignant disorders (S3)	N	9	13	2	0.003 *	0.390	Gemini > Perplexity; $p = 0.0014$ *
	% (95% CI)	36% (18–57.5%)	52% (31.3–72.2%)	8% (1–26%)			
Oral cancer (S4)	N	7	11	5;	0.173	0.216	
	% (95% CI)	28% (12.1–49.4%)	44% (24.4–65.1%)	20% (6.6–40.7%)			
Total	N	39	51	16	$p < 0.0001$ *	0.304	ChatGPT > Perplexity; $p < 0.0001$ * Gemini > Perplexity, $p < 0.0001$ *
	% (95% CI)	39% (29.4–49.3%)	51% (40.8–61.1%)	16% (9.4–24.7%)			

* significant difference ($p < 0.05$).

Gemini had the highest proportion of the responses that were correct in all three cycles of testing (51% (40.8–61.1%)). Both ChatGPT and Gemini had a significantly higher proportion of the responses that were correct in all three cycles of testing compared to

Perplexity (39% (29.4–49.3%) and 51% (40.8–61.1%) vs. 16% (9.4–24.7%); $p < 0.00001$; V = 0.304) (Table 2).

Gemini also had the highest proportion of the responses that were correct in all three cycles of testing across all four subgroups. No significant difference in the proportion of the responses that were correct in all three cycles of testing was found in S1 and S4. Significant difference in the proportion of the responses that were correct in all three cycles of testing was found between Gemini and Perplexity in S2 (52% (31.3–72.2%) vs. 12% (2.5–31.2%); $p = 0.00190$; V = 0.360) and S3 (52% (31.3–72.2%) vs. 8% (1–26%); $p = 0.00137$; V = 0.390). (Table 2).

Interestingly, not a single case of oral cancer was recognized by all three LLMs in all three cycles of testing. Representative examples of correctly and incorrectly classified oral mucosal lesions are presented in Figure 1.



Figure 1. Representative examples of correctly (**a**,**b**) and incorrectly (**c**,**d**) classified oral mucosal lesions by multimodal large language models: (**a**) a case of oral leukoplakia correctly identified by all 3 multimodal large language models in all 3 cycles of testing; (**b**) a case of oral recurrent herpes infection correctly identified by all 3 multimodal large language models in all 3 cycles of testing; (**c**) a case of oral mucocele incorrectly identified by all 3 multimodal large language models in all 3 cycles of testing; (**d**) a case of oral cancer incorrectly identified by all 3 multimodal large language models in all 3 cycles of testing;.

4. Discussion

Our study evaluated the diagnostic accuracy and temporal reproducibility of three commercially available multimodal large language models in the image-based diagnosis of oral mucosal lesions. The results demonstrate that Gemini achieved the highest diagnostic accuracy (78% in cycles 2 and 3), significantly outperforming ChatGPT (55–57%) and Perplexity (28–31%). However, temporal reproducibility was only fair to moderate across all models. Notably, Gemini showed fair intra-model agreement ($\kappa = 0.338$) despite its superior accuracy, whereas ChatGPT demonstrated moderate agreement ($\kappa = 0.525$) despite lower diagnostic performance. These findings highlight both the potential and the limitations of multimodal LLMs as diagnostic support tools in oral medicine.

The diagnostic accuracy observed in this study is consistent with the emerging literature on multimodal LLMs in oral pathology, although direct comparisons remain challenging due to methodological heterogeneity. Recent studies have reported LLM diagnostic accuracy ranging from 58% to 97% for oral lesions, with substantial variation depending on the model, prompt structure, and clinical context [12,13,23]. Our findings fall within this range but remain lower than the 94–97% accuracy reported for histopathology-based LLM diagnosis, suggesting that the interpretation of clinical photographs presents greater challenges than analysis of histological images [13]. The observed inter-model differences may be attributed to variations in training data sets, multimodal processing architecture, and optimization strategies. These inherent differences between AI systems may influence their ability to interpret subtle morphological features of oral mucosal lesions, including surface texture, color variation, and lesion borders.

An additional finding of this study is the absence of statistically significant differences in diagnostic accuracy between models in the first testing cycle, followed by clear and significant differences in cycles 2 and 3. Several factors may explain this pattern. First, multimodal LLMs generate outputs probabilistically, which may lead to variability in responses even when identical inputs are provided, potentially contributing to greater performance variability during the evaluation. Furthermore, given the one-month intervals between testing cycles, it is possible that ongoing model updates or modifications in underlying system parameters influenced performance over time, resulting in improved or more stable outputs in certain models. Overall, these findings support the notion that LLM-based diagnostic performance is not static but may evolve over time, with important implications for their clinical reliability and ongoing evaluation.

Subgroup analysis revealed distinct patterns in model performance across lesion categories. Gemini demonstrated particularly strong performance in identifying physiological variations (88–92%) and oral cancer (68–84%), while showing lower accuracy for benign lesions (60–72%) and oral potentially malignant disorders (64–68%). This pattern likely reflects differences in visual distinctiveness between lesion types. Normal anatomical variations and advanced malignancies often exhibit more characteristic morphological features, whereas benign lesions and early potentially malignant disorders display greater phenotypic overlap and diagnostic ambiguity [2]. From a clinical perspective, this suggests that current LLMs may be more reliable for recognizing clearly defined conditions or advanced disease than for differentiating between subtle mucosal alterations. However, one must keep in mind that the sample size in each subgroup was small ($n = 25$), and the results need to be interpreted with caution.

The temporal reproducibility findings represent a particularly important contribution of this study, as this dimension remains underexplored. The fair to moderate intra-model agreement observed across three testing cycles ($\kappa = 0.338$–0.525) raises important concerns regarding the reliability and stability of LLM-generated diagnoses over time. While Alshammari et al. [13] reported higher intra-model consistency ($\kappa = 0.740$–0.918) in the

context of histopathological image interpretation, their analysis was conducted over a shorter timeframe (n = 2 weeks) and may not have captured temporal variability associated with routine clinical use over time. In contrast, the present study, incorporating repeated evaluations at one-month intervals, enables assessment of longitudinal stability and suggests that observed variability may be influenced not only by stochastic model behavior but also by underlying system modifications. Commercial LLM platforms undergo continuous updates and refinements, potentially introducing temporal variation in model outputs that is not transparent to end users. Such temporal changes in internal model parameters (parameter drift) are inherent to constantly evolving AI systems. This variability has important clinical implications, as reproducibility is a fundamental requirement for reliable diagnostic support tools.

Several limitations of this study should be acknowledged. First, the dataset consisted of 100 clinical photographs from a single center, which may limit generalizability to broader clinical populations. Although the sample size was relatively small, based on our post hoc estimations, it was sufficient to detect moderate-to-large differences in diagnostic accuracy. However, smaller differences between models may not have been reliably identified, which should be considered when interpreting non-significant findings. However, the primary aim of this study was to assess the reproducibility of AI-generated outputs, which requires fewer observations than conventional diagnostic accuracy studies. The repeated-measures design, in which each image was evaluated across three independent cycles, increased the total number of observations (100 images $\times$ 3 cycles $\times$ 3 models = 900 observations) and allowed for robust estimation of agreement metrics. Nevertheless, we acknowledge that future studies should be performed with larger, multicenter datasets and formal a priori power calculations to confirm these findings and detect smaller effect sizes.

Apart from image clarity (i.e., the lesion of interest in focus), no additional technical standardization of clinical photographs was applied. While standardized imaging conditions might have improved diagnostic accuracy, this approach was intentionally chosen to reflect real-world clinical scenarios in which clinicians (or patients) capture images and seek LLMs' diagnostic support. Variability in image acquisition conditions—including lighting, shadows, reflections, and differences between imaging devices—may have influenced diagnostic accuracy by altering the visual representation of key morphological features such as color, texture, and lesion borders. However, as the same set of images was used across all testing cycles, these factors remained constant and are unlikely to have contributed to the observed temporal variability in model outputs. Rather, the variability in reproducibility is more likely attributable to intrinsic characteristics of the models, such as probabilistic output generation and potential system-level updates over time.

In addition, AI models were evaluated based solely on clinical images, without accompanying clinical history or symptomatology. This approach does not fully reflect real-world diagnostic decision-making and may have contributed to less reliable or more speculative outputs. However, it was intentionally adopted in order to isolate the visual diagnostic capabilities of multimodal LLMs. The inclusion of clinical context would introduce additional variability and could confound direct comparisons between models. Future studies should incorporate structured clinical information to better approximate real-world clinical scenarios and to evaluate the combined effect of visual and contextual inputs on diagnostic performance. However, isolating image-based performance represents a necessary first step in this process, as it establishes a baseline for further evaluation of LLMs' diagnostic capability. Such a stepwise approach is essential for understanding the individual and combined contributions of visual and contextual inputs to the overall diagnostic performance of LLMs.

Finally, only three commercially available LLMs were evaluated at specific time points, and ongoing system updates, as well as sensitivity to prompt design and model parameters, may influence future performance.

5. Conclusions

In conclusion, multimodal LLMs demonstrate promising potential for image-based assessment of oral mucosal lesions, particularly in recognizing visually distinct conditions such as normal anatomical variations and advanced malignancies. However, the observed variability in diagnostic reproducibility across time highlights a critical limitation that must be addressed before these systems can be reliably integrated into clinical practice. At present, LLMs should be considered supportive tools rather than independent diagnostic systems, and their outputs should be interpreted in association with clinical expertise. Further research is required to improve model stability, enhance performance in challenging lesion categories, and ensure consistent behavior over time.

Author Contributions: Conceptualization, V.B., L.D. and M.A.C.; methodology, V.B., R.A. and L.M.; validation, R.A., L.M. and D.V.J.; formal analysis, V.B.; investigation, L.D. and M.A.C.; resources, V.B. and D.V.J.; data curation, V.B. and D.V.J.; writing—original draft preparation, L.D. and M.A.C.; writing—review and editing, R.A., L.M. and D.V.J.; visualization, V.B.; supervision, V.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study protocol was approved by the Ethics Committee of the University of Zagreb School of Dental Medicine on 12 November 2025 (Approval No. KLASA: 03-01/25-05/13; No: 251-60-4/119-7). All procedures were conducted in accordance with the Declaration of Helsinki.

Informed Consent Statement: Informed consent was obtained from all patients before capturing the clinical photographs used in this study.

Data Availability Statement: Data are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial intelligence
LLM	Large language model
OPMD	Oral potentially malignant disorders

References

1. Feng, Q.J.; Harte, M.; Carey, B.; Alqarni, A.; Monteiro, L.; Diniz-Freitas, M.; Fricain, J.C.; Lodi, G.; Brailo, V.; Andreoletti, M.; et al. The risks of artificial intelligence: A narrative review and ethical reflection from an Oral Medicine group. *Oral Dis.* **2025**, *31*, 348–353. [CrossRef]
2. Warin, K.; Limprasert, W.; Suebnukarn, S.; Jinaporntham, S.; Jantana, P.; Vicharueang, S. AI-based analysis of oral lesions using novel deep convolutional neural networks for early detection of oral cancer. *PLoS ONE* **2022**, *17*, e0273508. [CrossRef]
3. Rokhshad, R.; Mohammad-Rahimi, H.; Price, J.B.; Shoorgashti, R.; Abbasiparashkouh, Z.; Esmaeili, M.; Sarfaraz, B.; Rokhshad, A.; Motamedian, S.R.; Soltani, P.; et al. Artificial intelligence for classification and detection of oral mucosa lesions on photographs: A systematic review and meta-analysis. *Clin. Oral Investig.* **2024**, *28*, 88. [CrossRef]
4. Li, J.; Kot, W.Y.; McGrath, C.P.; Chan, B.W.A.; Ho, J.W.K.; Zheng, L.W. Diagnostic accuracy of artificial intelligence assisted clinical imaging in the detection of oral potentially malignant disorders and oral cancer: A systematic review and meta-analysis. *Int. J. Surg.* **2024**, *110*, 5034–5046. [CrossRef]

5. Rajpurkar, P.; Lungren, M.P. The Current and Future State of AI Interpretation of Medical Images. *N. Engl. J. Med.* **2023**, *388*, 1981–1990. [CrossRef]
6. Bhayana, R. Chatbots and Large Language Models in Radiology: A Practical Primer for Clinical and Research Applications. *Radiology* **2024**, *310*, e232756. [CrossRef]
7. Alvarez-Silberberg, V.I.; Alvarez-Silberberg, C.P.; Galletti, C.; Flores-Fraile, J.; Galletti, C.; Ramirez, V.; Palma, C.B.; Gil-Manich, V.; Fiorillo, L.; Mehta, V.; et al. Comparative analysis of large language models as decision support tools in oral pathology. *Sci. Rep.* **2026**, *Epub ahead of printing*. [CrossRef]
8. Hou, B.; Mukherjee, P.; Batheja, V.; Wang, K.C.; Summers, R.M.; Lu, Z. One Year On: Assessing Progress of Multimodal Large Language Model Performance on RSNA 2024 Case of the Day Questions. *Radiology* **2025**, *316*, e250617. [CrossRef] [PubMed]
9. Nguyen, V.A.; Vuong, T.Q.T.; Nguyen, V.H. Benchmarking large-language-model vision capabilities in oral and maxillofacial anatomy: A cross-sectional study. *PLoS ONE* **2025**, *20*, e0335775. [CrossRef] [PubMed]
10. Zhou, Y.; Wang, W.; Wang, P.; Hu, K. Diagnostic performance of large language models on the NEJM image challenge: A comparative study with human evaluators and the impact of prompt engineering. *Front. Med.* **2026**, *12*, 1709413. [CrossRef] [PubMed]
11. Gupta, M.; Virostko, J.; Kaufmann, C. Large language models in radiology: Fluctuating performance and decreasing discordance over time. *Eur. J. Radiol.* **2025**, *182*, 111842. [CrossRef]
12. Kim, K.; Kim, B.C. Diagnostic Performance of Large Language Models in Multimodal Analysis of Radiolucent Jaw Lesions. *Int. Dent. J.* **2025**, *75*, 103910. [CrossRef]
13. Suárez, A.; Freire, Y.; Suárez, M.; Díaz-Flores García, V.; Andreu-Vázquez, C.; Thuissard Vasallo, I.J.; Castillo Varón, A.I.; Martín, C. Diagnostic Performance of Multimodal Large Language Models in the Analysis of Oral Pathology. *Oral Dis.* **2025**, *Epub ahead of printing*. [CrossRef]
14. Alshammari, A.F.; Madfa, A.A.; Anazi, B.A. Human versus artificial intelligence in oral pathology diagnosis: A comparative study of ChatGPT, Grok, and MANUS. *Sci. Rep.* **2026**, *Epub ahead of printing*. [CrossRef]
15. Kelly, C.J.; Karthikesalingam, A.; Suleyman, M.; Corrado, G.; King, D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* **2019**, *17*, 195. [CrossRef]
16. Wiens, J.; Saria, S.; Sendak, M.; Ghassemi, M.; Liu, V.X.; Doshi-Velez, F.; Jung, K.; Heller, K.; Kale, D.; Saeed, M.; et al. Do no harm: A roadmap for responsible machine learning for health care. *Nat. Med.* **2019**, *25*, 1337–1340. [CrossRef] [PubMed]
17. Topol, E.J. High-performance medicine: The convergence of human and artificial intelligence. *Nat. Med.* **2019**, *25*, 44–56. [CrossRef] [PubMed]
18. Ibrahim, H.; Liu, X.; Rivera, S.C.; Moher, D.; Chan, A.W.; Sydes, M.R.; Calvert, M.J.; Denniston, A.K. Reporting guidelines for clinical trials of artificial intelligence interventions: The SPIRIT-AI and CONSORT-AI guidelines. *Trials* **2021**, *22*, 11. [CrossRef]
19. Palaniappan, K.; Lin, E.Y.T.; Vogel, S. Global Regulatory Frameworks for the Use of Artificial Intelligence (AI) in the Healthcare Services Sector. *Healthcare* **2024**, *12*, 562. [CrossRef] [PubMed]
20. Dubuc, A.; Zitouni, A.; Thomas, C.; Kémoun, P.; Cousty, S.; Monsarrat, P.; Laurencin, S. Improvement of Mucosal Lesion Diagnosis with Machine Learning Based on Medical and Semiological Data: An Observational Study. *J. Clin. Med.* **2022**, *11*, 6596. [CrossRef]
21. Al-Ali, A.; Hamdi, A.; Elshrif, M.; Isufaj, K.; Shaban, K.; Chauvin, P.; Madathil, S.; Daer, A.; Tamimi, F.; Ba-Hattab, R. CLASEG: Advanced multiclassification and segmentation for differential diagnosis of oral lesions using deep learning. *Sci. Rep.* **2025**, *15*, 23016. [CrossRef]
22. Vaira, L.A.; Lechien, J.R.; Maniaci, A.; De Vito, A.; Mayo-Yáñez, M.; Troise, S.; Consorti, G.; Chiesa-Estomba, C.M.; Cammaroto, G.; Radulesco, T.; et al. Diagnostic Performance of ChatGPT-4o in Analyzing Oral Mucosal Lesions: A Comparative Study with Experts. *Medicina* **2025**, *61*, 1379. [CrossRef]
23. Hassanein, F.E.A.; Hussein, R.R.; Almalahy, H.G.; Sarhan, S.; Ahmed, Y.; Abou-Bakr, A. Vision-based diagnostic gain of ChatGPT-5 and gemini 2.5 pro compared with human experts in oral lesion assessment. *Sci. Rep.* **2025**, *15*, 43279. [CrossRef]
24. McHugh, M.L. Interrater reliability: The kappa statistic. *Biochem. Med.* **2012**, *22*, 276–282. [CrossRef]
25. Faul, F.; Erdfelder, E.; Lang, A.G.; Buchner, A. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* **2007**, *39*, 175–191. [CrossRef] [PubMed]
26. Faul, F.; Erdfelder, E.; Buchner, A.; Lang, A.G. Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behav. Res. Methods* **2009**, *41*, 1149–1160. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.