BUKTI KORESPONDENSI
ARTIKEL JURNAL INTERNASIONAL BEREPUTASI

Judul artikel : Accuracy of Orthodontic Malocclusion Detection Using Multiple AI Models: A Comparative Study

Jurnal : Healthcare Informatics Research, Volume 32, Issue 2, Mei 2026. DOI: 10.4258/hir.2026.32.2.166

No	Perihal	Tanggal
1	Bukti konfirmasi submit artikel dan artikel yang disubmit	4 Januari 2026
2	Bukti kegagalan pemeriksaan teknis dan permintaan resubmit artikel	8 Januari 2026
3	Bukti konfirmasi submit revisi dan artikel yang diresubmit	15 Januari 2026
4	Bukti permintaan revisi	19 Maret 2026
5	Bukti konfirmasi submit revisi, respon kepada reviewer, dan artikel yang diresubmit	27 Maret 2026
6	Bukti konfirmasi artikel accepted	21 April 2026
7	Bukti permintaan proofreading	24 April 2026
8	Bukti penerimaan proofreading serta artikel yang telah dilakukan proofreading	24 April 2026

1. Bukti konfirmasi submit artikel dan artikel yang disubmit

4 Januari 2026

UNIVERSITAS TRISAKTI

Joko Kusnoto <joko.k@trisakti.ac.id>

Submission Confirmation for Accuracy of Orthodontic Malocclusion Detection Using Multiple AI Models: A Comparative Study

1 message

Healthcare Informatics Research <em@editorialmanager.com> Sun, Jan 4, 2026 at 3:30 AM
Reply-To: Healthcare Informatics Research <hir@kosmi.org>
To: Joko Kusnoto <joko.k@trisakti.ac.id>

CC: "Hillda Herawati" hillda.herawati@lecture.unjani.ac.id, "Indrayadi Gunardi" indrayadi@trisakti.ac.id, "Anggit Wirasto" anggitwirasto@uhb.ac.id, "Tri Erri Astoeti" tri.erri@trisakti.ac.id

Dear Dr Kusnoto,

Your submission entitled "Accuracy of Orthodontic Malocclusion Detection Using Multiple AI Models: A Comparative Study" has been received by journal Healthcare Informatics Research

You will be able to check on the progress of your paper by logging on to Editorial Manager as an author. The URL is https://www.editorialmanager.com/hir/.

Your manuscript will be given a reference number once an Editor has been assigned.

Thank you for submitting your work to this journal.

Kind regards,

Healthcare Informatics Research

COVER LETTER

January 3rd, 2026

To:
Editor in Chief of Healthcare Informatics Research

Dear Editor in Chief,

We would like to submit our manuscript entitled “Accuracy of Orthodontic Malocclusion Detection Using Multiple AI Models: A Comparative Study” for consideration for publication in Healthcare Informatics Research.

This cross-sectional, observational study evaluates the accuracy with which multiple artificial intelligence models (ChatGPT 4.0, Gemini 3, Claude 4.5 Sonnet, and Microsoft Copilot) can detect orthodontic malocclusion features in standardized, multi-view, intraoral photographs. The study uses orthodontist assessment as the reference standard. The study focuses on the classification of eight clinically relevant malocclusion parameters in children aged 9–12 years using images.
This work's primary contribution lies in its healthcare informatics perspective. Instead of proposing a new diagnostic system, the study compares how different general-purpose, multimodal AI models interpret the same clinical image inputs. It highlights patterns of agreement, discrepancy, and model-specific limitations. Accuracy metrics are reported to support the comparative analysis and inform the development of future task-specific AI systems.
To our knowledge, this is one of the first studies to directly compare multiple large language models across modalities using standardized intraoral photographic data in an orthodontic context. Our findings offer practical insights into the current capabilities and limitations of AI-driven image interpretation for clinical decision support.
This manuscript has not been published or under consideration elsewhere. The authors declare no conflicts of interest. We believe this work aligns with the scope of healthcare informatics research because it provides evidence regarding clinical data readiness, AI model evaluation, and the responsible integration of artificial intelligence into healthcare practices.

Thank you for your consideration.

Best regards
Joko Kusnoto, DDS, MS, PhD

Accuracy of Orthodontic Malocclusion Detection Using Multiple AI Models: A Comparative Study

Hillda Herawati 1, Joko Kusnoto 2, Indrayadi Gunardi 3, Anggit Wirasto 4, Tri Erri Astoeti 5

1 Doctoral Candidate, Faculty of Dentistry, Universitas Trisakti, Jakarta, Indonesia.
2 Department of Orthodontics, Faculty of Dentistry, Universitas Trisakti, Jakarta, Indonesia.
3 Department of Oral Medicine, Faculty of Dentistry, Universitas Trisakti, Jakarta, Indonesia.
4 Informatics Study Program, Faculty of Science and Technology, Universitas Harapan Bangsa, Purwokerto, Indonesia.
5 Department of Dental Public Health and Preventive Dentistry, Faculty of Dentistry, Universitas Trisakti, Jakarta, Indonesia.

*Correspondence:
Name: Joko Kusnoto
Address: Department of Orthodontics, Faculty of Dentistry, Universitas Trisakti
Jl. Kyai Tapa no. 260. Jakarta 11440. Indonesia
Phone: +62-21-5672731
E-mail:joko.k@trisakti.ac.id

Abstract

Objective: This study aimed to evaluate and compare the accuracy of multiple AI models (ChatGPT 5.2 Pro, Gemini 3 Fast, Claude 4.5 Sonnet, and Microsoft Copilot) in detecting orthodontic malocclusion features in standardized, multi-view, intraoral photographs. The reference standard was an orthodontist assessment. **Methods:** A cross-sectional observational study was conducted using five standardized intraoral photographs (frontal, right lateral, left lateral, maxillary occlusal, and mandibular occlusal) obtained from 50 children aged 9-12 years. The following eight malocclusion parameters were assessed: anterior crowding, diastema, overjet, overbite, molar relationship, canine relationship, crossbite, and dental-arch symmetrical. The diagnostic accuracy and agreement between each AI model and the orthodontist was evaluated using Cohen's kappa and AUC. **Results:** The agreement between the AI models and the orthodontist ranged from poor to moderate across all orthodontic domains, with Cohen's κ -0.15 to 0.63. Visually prominent alignment features, including anterior crowding and diastema, demonstrated comparatively higher agreement (κ 0.00-0.63) and discriminatory performance, with ROC-AUC values ranging from 0.56 to 0.85. In contrast, parameters requiring precise spatial interpretation such as sagittal relationships, overbite, crossbite, and arch morphology showed consistently low agreement (κ -0.15 to 0.38) and poor to near-random classification performance, with AUC values predominantly between 0.41 and 0.70, and in some cases approaching 0.50. **Conclusion:** Current multimodal AI models demonstrate limited and parameter-dependent accuracy in detecting orthodontic malocclusions using intraoral photographs. These results emphasize the limitations of general-purpose AI systems for orthodontic decision support and highlight the necessity of task-specific models trained on clinically annotated datasets.

Keywords: accuracy, artificial intelligence, detection, malocclusion, orthodontic

I. Introduction

Malocclusion is a prevalent oral health problem among children and adolescents, with prevalence rates reported to be as high as 30-40% in some populations [1] . Early detection is crucial for preventing functional impairment and the need for more complex orthodontic interventions later in life [1,2]. Traditional orthodontic screening methods largely rely on direct clinical examinations by trained specialists, which can limit accessibility to necessary dental care, especially in community settings like schools [2,3].

Recent advancements in artificial intelligence (AI) have shown promise in improving diagnostic accuracy within orthodontics. This includes the application of machine learning and deep learning techniques to various imaging modalities. For instance, convolutional neural networks (CNN) have been effectively used in cephalometric analysis and panoramic image interpretation, yielding accurate automated orthodontic diagnoses [4,5]. However, there remains a gap in the utilization of intraoral photographs, which offer non-invasive, low-cost clinical records suitable for early screening and tele orthodontics [6–8] . The potential accuracy of intraoral photographs for diagnosing dental conditions has been recognized, with studies indicating strong diagnostic outcomes [6,9] .

In recent years, large language models (LLMs), such as ChatGPT and similar AI frameworks, have surfaced, demonstrating the capability to interpret both textual and visual inputs. Studies evaluating LLM responses in orthodontics reveal moderate to high consistency but significant variability in accuracy compared to human orthodontic specialists [10,11] . While these models can generate fluent and structured answers, they often provide incomplete or misleading information, thus necessitating a cautious approach to clinical interpretation [11–

14] . This emphasizes the importance of continuous evaluation and possible integration of LLMs in orthodontic practice, serving as an aid rather than a replacement for expert judgment.

Comparative studies indicate discrepancies between LLM-generated responses and those rendered by orthodontic experts, especially in malocclusion classification and treatment decision-making processes [11,12] . Despite the potential of LLMs for preliminary educational contributions, existing evidence suggests they cannot supplant the nuanced assessments made by experienced clinicians [1,4,14]. This aspect highlights the need for further investigations to rigorously validate these AI models' outputs against established clinical practices.

To evaluate the accuracy of malocclusion classification from clinical images via AI models, including advanced versions like (ChatGPT 5.2 Pro, Gemini 3 Fast, Claude 4.5 Sonnet, and Microsoft Copilot), standardized inputs are essential for comparison against expert evaluations. As of now, there is limited research focused on the performance of multimodal LLMs in analyzing standardized multi-view intraoral photographs against comprehensive orthodontic diagnostic criteria. Future studies are necessary to explore this domain [9,12,15].

II. Method

1. Study design

This observational cross-sectional study was designed to evaluate the diagnostic performance of multimodal large language models (LLMs) in detecting orthodontic malocclusion features from standardized intraoral photographs. The study focused exclusively on image-based diagnostic interpretation without incorporating clinical examination, radiographic analysis, or treatment planning. The study was conducted using retrospective clinical photographs obtained from 50 children aged 9-12 years in Cimahi, West Java, Indonesia. No clinical intervention, treatment modification, or follow-up assessment was performed as part of this study.

2. Study population

The study included intraoral photographic records from 50 children aged 9-12 years. This age range was selected to represent mixed to early permanent dentition stages, during which orthodontic screening and early diagnosis are clinically relevant. Patients were included if complete sets of standardized intraoral photographs (frontal, lateral right, lateral left, maxillary occlusal, and mandibular occlusal views) were available and of sufficient quality for diagnostic interpretation. Records with blurred images, incomplete views, or significant photographic artifacts were excluded.

3. Ethical approval and Consent to participate

All participants and their legal guardians provided written informed consent prior to the initiation of the study. Ethical approval for the study protocol was granted by the Ethics Committee of the Faculty of Dentistry, Universitas Trisakti (No. 1006/S3/KEPK/FKG/9/2025).

4. Image collection

Five images were acquired for each participant, comprising frontal, left lateral, right lateral, maxillary, and mandibular occlusal views of the teeth (in Figure 1). All images were captured using an iPhone 15 (Apple Inc., Cupertino, CA, USA) with a resolution of 6048 × 4032 pixels (24 megapixels). A single investigator, who was calibrated for this study, acquired all images according to standardized intraoral photography protocols to ensure consistent angulation, lighting, and field of view. Baseline orthodontic parameters were established following recognized diagnostic criteria, including Angle’s classification for sagittal relationships and

the Dental Health Component (DHC) of the Index of Orthodontic Treatment Need (IOTN), which assesses occlusal traits [16,17].

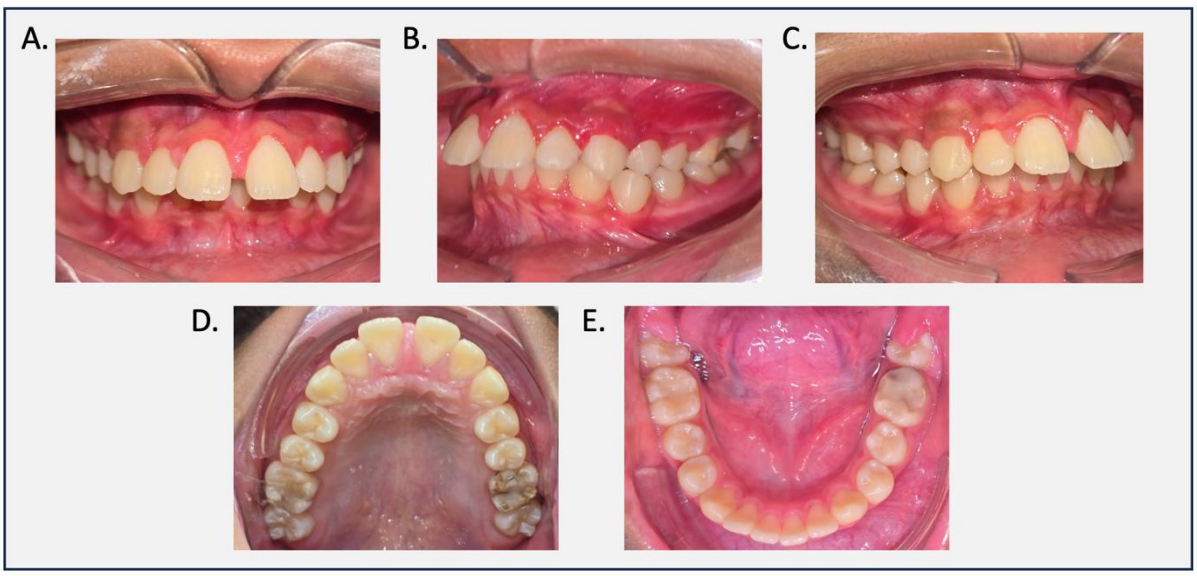


Figure 1. Sample of clinical images obtained from orthodontic patients. A. Frontal; B. Lateral left; C. Lateral right, D. Maxillary occlusal view, and E. Mandibular occlusal view.

5. Image preparation and preprocessing

All images were anonymized prior to analysis by removing information that could identify patients. No digital enhancement, filtering, contrast adjustment, or color correction was applied. However, images were cropped and resized due to the presence of other objects such as hands and mirrors during image acquisition. Images were analyzed at their original resolution to preserve clinically relevant visual features, such as tooth arrangement, occlusal relationships, and arch morphology.

6. Orthodontic diagnostic framework (IOTN-DHC based criteria)

Orthodontic assessment within this study employed a structured diagnostic framework based on established clinical standards, including Proffit's definitions of malocclusion, Angle's classification of sagittal relationships, and the Dental Health Component (DHC) of the Index of Orthodontic Treatment Need (IOTN) [18,19]. The IOTN-DHC framework is favoured for its clinically validated, internationally recognized method for categorizing malocclusion severity by utilizing observable occlusal and functional characteristics [18,19].
For analytical consistency and effective image-based evaluation, orthodontic parameters were systematically organized into five diagnostic domains that reflect varying levels of spatial complexity:

1. **Alignment Domain**: This domain included the assessment of anterior crowding and diastema in both the maxillary and mandibular arches. These features are commonly recognized indicators in orthodontic screening [20].
2. **Sagittal Relationship Domain**: This aspect comprised evaluating overjet, molar relationship, canine relationship, and overall Angle classification, focusing on anteroposterior dental relationships crucial for orthodontic diagnoses [21].

3. **Vertical Relationship Domain**: This domain emphasized overbite assessments categorized as normal, deep bite, or open bite, which denote significant vertical discrepancies in function [21].
4. **Transverse Relationship Domain**: In this domain, the presence and type of crossbite (anterior or posterior, unilateral or bilateral) were assessed, capturing transverse discrepancies often subtle in two-dimensional imaging [22].
5. **Arch Morphology Domain**: This domain evaluated dental arch symmetry through occlusal views, highlighting structural characteristics shaped by underlying skeletal patterns [23].

All parameters were systematically evaluated under this unified framework by both orthodontists and AI models, ensuring a structured approach. For statistical analysis, outcomes were dichotomized into normal and abnormal categories, facilitating a direct comparison of diagnostic agreement and discrimination [24].

7. Generative AI Multimodal large language models

In this study, four publicly available multimodal large language models (LLMs) with image-text processing capabilities were evaluated. The models included ChatGPT 5.2 Pro (OpenAI), Claude 4.5 Sonnet (Anthropic), Gemini 3 Fast (Google), and Microsoft Copilot (Microsoft) [11,25] . These models were accessed through their official web-based interfaces, adhering to their default system configurations without any application programming interfaces (APIs), external tools, or model fine-tuning procedures [11]. This ensures that the evaluation reflects a true representation of how these LLMs would operate in clinical contexts.

ChatGPT 5.2 Pro[26]

ChatGPT was utilized via its web interface, leveraging its built-in vision capability for image-based reasoning. Intraoral photographs were uploaded directly, and diagnostic outputs were generated solely based on the visual input and standardized prompts [11]. The application of this model aligns with current trends in artificial intelligence (AI) usage in dentistry, which emphasize enhancing diagnostic accuracy through automated systems [27].

Claude 4.5 Sonnet[28]

Claude 4.5 Sonnet was accessed through the MoniCA platform, known for its structured image interpretation capabilities. Each image was analyzed independently in its default configuration, reflecting the model's design for independent evaluations without conversational memory across different cases [25] . The independent nature of analysis enables clearer comparisons of outcomes and showcases the model's potential in diagnostic contexts [11].

Gemini 3 Fast[29]

Gemini was evaluated through its dedicated web application, allowing for direct image uploads and prompt-based analyses without iterative refinement or follow-up questioning. This straightforward approach reflects the model's capabilities and is consistent with research advocating the use of AI for rapid diagnostics in clinical dentistry [25,30].

Microsoft Copilot[31]

Finally, Microsoft Copilot was accessed via its web interface, utilizing its integrated multimodal image analysis features. The model was employed without advanced enterprise features or external integrations, which ensures it was evaluated under conditions akin to general clinical usage [11].

8. Model testing workflow and prompting strategy

Within this current research, model evaluation was done using a structured process to compare all of the multimodal large language models (LLMs). Every individual intraoral image was

evaluated without prior images and without prior model output to avoid both carryover effects and to make certain that every individual diagnostic result was based solely upon visual information contained within an image itself.

A parameter-specific prompt was also used for each parameter of orthodontics. The prompts were carefully designed to model the cognitive process of analyzing typical of an orthodontist as they specifically ask the model to stress key anatomical landmarks, use uniform diagnostic criteria, and have a systematic diagnostic statement. Each prompt was written in Indonesian, following a uniform syntax, level of detail, and command format to reduce variability from the prompt and ensure prompt-induced bias in performance. The prompts were not altered or refined during the test procedure. Every image and corresponding prompt was tested separately for each of these models. The results generated by these models were then explicitly mapped to a predetermined orthodontic diagnostic category.

9. Standardized prompting framework

The prompting framework was aligned with the orthodontic diagnostic rubric and organized by diagnostic domain and image view. Table 1 summarizes the standardized prompts used for each orthodontic parameter.

Table 1. Standardized prompts used for multimodal LLM evaluation by orthodontic domain

Orthodontic domain	**Image view**	**Diagnostic focus**	**Standardized prompt**
Alignment	Frontal intraoral	Anterior crowding	"Analyze this frontal intraoral photograph. Focus on anterior crowding. Identify whether crowding is present in the maxillary and/or mandibular anterior teeth, specify the affected teeth, classify severity as mild, moderate, or severe, and provide a diagnostic conclusion."
Alignment	Frontal intraoral	Diastema	"Analyze this frontal intraoral photograph. Focus on diastema. Determine whether diastema is present, identify its location (maxillary and/or mandibular), and provide a diagnostic conclusion."
Sagittal relationship	Lateral intraoral (right/left)	Molar relationship (M1)	"Analyze this lateral intraoral photograph. Focus on the relationship between the maxillary first molar and mandibular first molar. Classify the molar relationship as Class I, II, or III. Report findings separately for the right and left sides."
Sagittal relationship	Lateral intraoral (right/left)	Canine relationship (C)	"Analyze this lateral intraoral photograph. Focus on the relationship between the maxillary canine and mandibular canine. Classify the canine relationship as Class I, II, or III. Report findings separately for the right and left sides."
Sagittal relationship	Lateral intraoral	Overjet	"Analyze this lateral intraoral photograph. Focus on overjet. Classify the overjet as normal, increased, edge-to-edge, or reverse, and provide a diagnostic conclusion."
Vertical relationship	Frontal intraoral	Overbite	"Analyze this frontal intraoral photograph. Focus on overbite. Classify the vertical overlap as normal, deep bite, or open bite, and provide a diagnostic conclusion."
Transverse relationship	Frontal intraoral	Crossbite	"Analyze this frontal intraoral photograph. Focus on transverse relationships. Identify the presence of crossbite, specify whether it is unilateral or bilateral, and provide a diagnostic conclusion."
Arch morphology	Maxillary occlusal	Arch symmetry	"Analyze this maxillary occlusal photograph. Assess the arch is symmetrical or asymmetrical, and provide a diagnostic conclusion."

Arch morphology	Mandibular occlusal	Arch symmetry	"Analyze this mandibular occlusal photograph. Assess the arch is symmetrical or asymmetrical, and provide a diagnostic conclusion."

10. Output handling and categorization

AI-generated outputs were reviewed and mapped to predefined orthodontic categories based on the diagnostic rubric. For statistical analysis, all parameters were subsequently dichotomized into normal and abnormal categories. No interpretation, correction, or clinical judgment was applied beyond this categorical mapping.

11. Statistical analysis

The Kappa test was employed to assess the level of agreement between the artificial intelligence models and the orthodontist evaluations. This method is widely recognized in dental research to quantify inter-rater reliability, particularly in studies examining diagnostic discrepancies [9,32] . By establishing a comparative framework, the Kappa test provides insight into how effectively AI models align with expert clinical assessments, thus enhancing the understanding of AI's potential in orthodontics [33] . To evaluate the accuracy of each artificial intelligence model, all diagnostic variables were dichotomized into two response categories: (1) normal, none, or Class I; and (2) abnormal, present, or any classification other than Class I. This classification is significant, as it allows for a clearer identification of malocclusion types, aligning with established orthodontic standards and facilitating comparisons across studies [9,33] .

The categorization is crucial in the context of malocclusion assessment, where variations in classification can impact treatment planning and outcomes. For instance, studies have demonstrated that Class I malocclusion is the most prevalent finding in children, with varied implications for subsequent orthodontic treatment pathways [34,35]. By utilizing this dichotomous approach, the research effectively contributes to the ongoing discourse regarding the efficacy of AI diagnostics compared to traditional clinical evaluation methods [9,12].

III. Results

A total of 50 childrens were included in this study, consisting of 20 males and 30 females. The participants' ages ranged from 9 to 12 years (Table 2). For each subject, five standardized intraoral photographs were obtained, comprising frontal, left lateral, right lateral, maxillary occlusal, and mandibular occlusal views. In total, 250 intraoral images were analyzed and used as input for the artificial intelligence models.

Table 2. subject demographic results

Variable	Value
Gender [n, %]	
Male	20 (40)
Female	30 (60)
Age (year) [mean, SD]	10.56 (0.73)

Within the alignment domain (Table 3), anterior crowding was highly prevalent in both arches. In the maxilla, mild to moderate crowding accounted for the majority of cases (72%), while severe crowding was observed in 16% of patients. Similarly, the mandible showed a predominance of moderate crowding (42%), followed by mild (32%) and severe (14%) presentations. Diastema was more frequently identified in the mandible (24%) than in the maxilla (8%). These findings indicate substantial variability in alignment-related malocclusion patterns within the dataset.

Table 3. Alignment domain

Parameter	**n (%)**
Maxillary crowding	
None	6 (12)
Mild	23 (46)
Moderate	13 (26)
Severe	8 (16)
Mandibular crowding	
None	6 (12)
Mild	16 (32)
Moderate	21 (42)
Severe	7 (14)
Diastema	
Maxilla	4 (8)
Mandible	12 (24)

Sagittal relationships demonstrated a predominance of normal occlusal patterns, although clinically relevant deviations were frequently observed (Table 4). Normal overjet was present in 62% of patients, while increased overjet was identified in 30%. Edge-to-edge and reverse overjet configurations were less common, each accounting for 4% of cases.

Angle Class I molar and canine relationships predominated on both sides (Table 4); however, a considerable proportion of Class II relationships was observed bilaterally. Overall Angle classification revealed that more than half of the patients were classified as Class I (54%), while Class II malocclusion accounted for 42% of the sample. Class III malocclusion was relatively rare (4%). These distributions highlight the presence of clinically meaningful sagittal discrepancies suitable for comparative AI analysis.

Table 4. Sagittal relationship domain

Parameter	**n (%)**
Overjet	
Normal	31 (62)

Increased	15 (30)
Edge-to-edge	2 (4)
Reverse	2 (4)
Molar relationship (right)	
Class I	31 (62)
Class II	18 (36)
Class III	1 (2)
Molar relationship (left)	
Class I	30 (60)
Class II	18 (36)
Class III	2 (4)
Canine relationship (right)	
Class I	28 (56)
Class II	22 (44)
Canine relationship (left)	
Class I	30(60)
Class II	19 (38)
Class III	1(2)
Angle classification	
Class I	27 (54)
Class II	21 (42)
Class III	2 (4)

Assessment of the vertical relationship (Table 5) indicated that most patients exhibited a normal overbite (62%). Deep bite was identified in more than one-third of the sample (36%), whereas open bite was rare, occurring in only one patient (2%). This distribution reflects a predominance of vertical patterns characterized by increased rather than reduced vertical overlap.

Table 5. Vertical relationship domain

Parameter	**n (%)**
Overbite	
Normal	31 (62)
Open bite	1 (2)
Deep bite	18 (36)

Transverse discrepancies were relatively uncommon within the study population (Table 6). The majority of patients showed no crossbite on either the left or right side (82%). When present, crossbites were predominantly unilateral and anterior in nature. Posterior and combined anterior-posterior crossbites occurred infrequently. These findings suggest limited transverse variability compared with other orthodontic domains.

Table 6. Transverse relationship domain

Parameter	**n (%)**
Crossbite (left)	
None	41 (82)
Anterior	6 (12)
Posterior	2 (4)
Anterior + posterior	1 (2)
Crossbite (right)	
None	41 (82)
Anterior	6 (12)
Posterior	3 (6)

Despite variability in arch form, most arches were classified as symmetrical (Table 7), with symmetry observed in 78% of maxillary arches and 80% of mandibular arches. This combination of morphological diversity and predominant symmetry provides a heterogeneous yet clinically representative dataset for AI-based analysis.

Table 7. Arch morphology domain

Parameter	**n (%)**
Arch symmetry	
Maxilla (symmetrical)	39 (78)
Mandible (symmetrical)	40 (80)

Following the descriptive analysis, receiver operating characteristic (ROC) analysis was performed to evaluate the ability of the multimodal AI models to discriminate between normal and abnormal orthodontic findings across all diagnostic domains, independent of decision thresholds. Figure 2 presents the integrated receiver operating characteristic (ROC) curves for all evaluated orthodontic parameters across the five diagnostic domains. Overall, the ROC analysis demonstrated clear domain-dependent differences in the discriminatory performance of the multimodal AI models.

Alignment-related parameters, particularly maxillary and mandibular crowding as well as maxillary diastema, showed the most favourable ROC profiles, with curves deviating further from the diagonal reference line, indicating better discrimination between normal and abnormal

conditions. Among the evaluated models, ChatGPT and Gemini generally demonstrated higher discriminatory performance for these visually prominent features.

In contrast, parameters within the sagittal relationship domain, including molar and canine relationships as well as overall Angle classification, exhibited moderate discrimination. The ROC curves for these parameters were closer to the diagonal, reflecting limited ability to reliably distinguish normal from abnormal sagittal relationships across all AI models.

The poorest discriminatory performance was observed for parameters within the vertical, transverse, and arch morphology domains. ROC curves for overbite, crossbite, and dental arch symmetry clustered near the diagonal line, indicating near-random classification performance. These patterns were consistent across all evaluated AI models.

Taken together, the integrated ROC curves illustrate that current multimodal AI models perform substantially better in detecting orthodontic features with strong global visual cues than in interpreting complex spatial relationships that require precise geometric assessment.

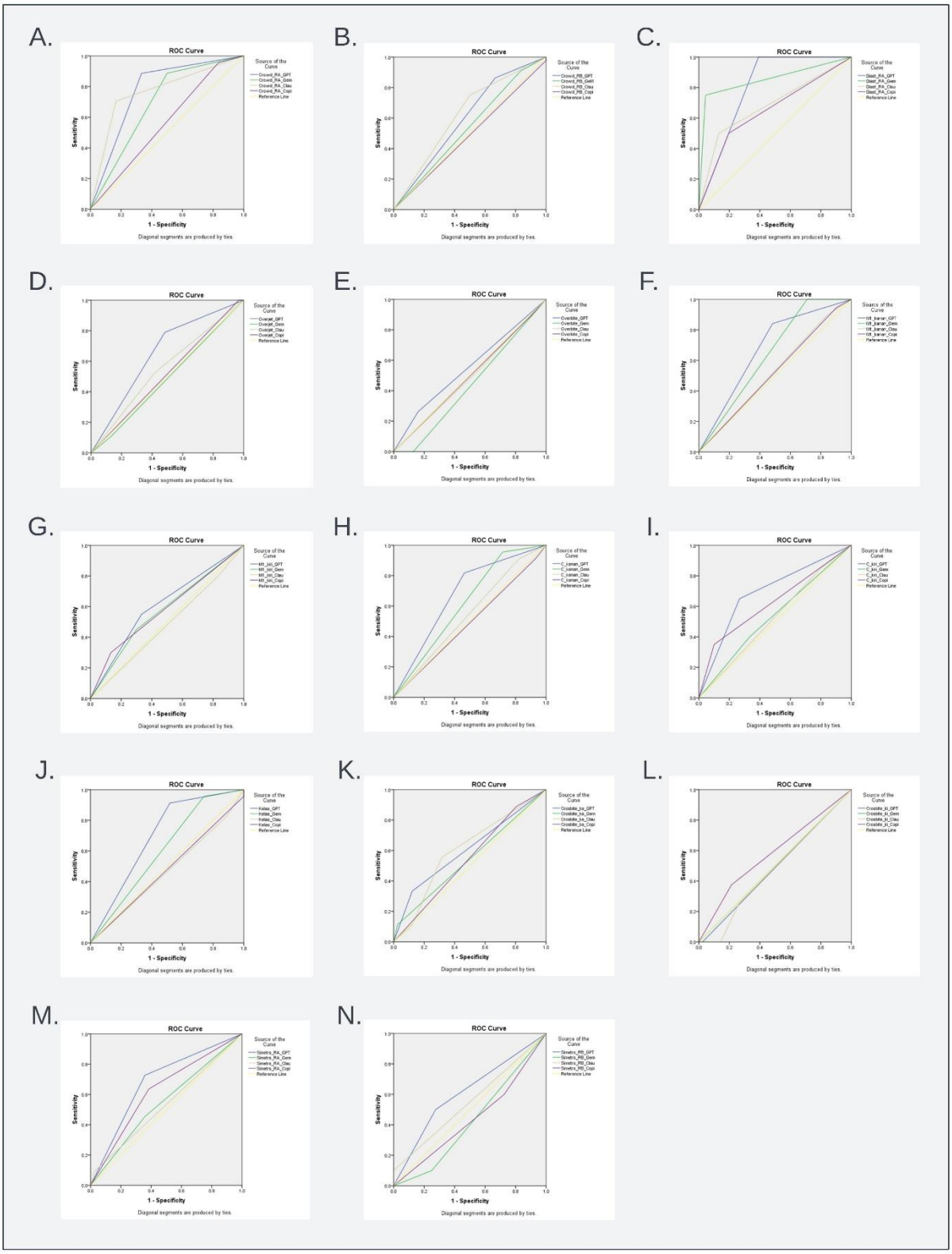


Figure 2. Integrated receiver operating characteristic (ROC) curves of multimodal AI models for orthodontic parameters. Panels A-O represent ROC curves for individual orthodontic parameters evaluated from standardized multi-view intraoral photographs. Panel A: maxillary crowding; B: mandibular crowding; C: maxillary diastema; D: overjet; E: overbite; F: right first molar relationship; G: left first molar relationship; H: right canine relationship; I: left canine relationship; J: Angle classification; K: right crossbite; L: left crossbite; M: maxillary arch symmetry; N: mandibular arch symmetry. Each panel compares the discriminatory

performance of ChatGPT 5.2 Pro, Gemini 3 Fast, Claude 4.5 Sonnet, and Microsoft Copilot against the orthodontist reference standard.

As summarized in Table 8, the discriminatory performance of the multimodal AI models showed clear domain-dependent variation. The alignment domain consistently demonstrated the strongest diagnostic performance, with Gemini achieving the highest AUC value of 0.85, indicating good discrimination for visually salient features such as anterior crowding and diastema. ChatGPT also performed well in this domain, with AUC values approaching 0.80, whereas Claude showed slightly lower but comparable performance. Copilot exhibited the lowest discriminatory ability among the models in the alignment domain.

In the sagittal relationship domain, discriminatory performance was more limited. ChatGPT achieved the highest AUC values, reaching approximately 0.70, suggesting moderate discrimination for molar and canine relationships as well as overall Angle classification. The remaining models demonstrated lower AUC values, indicating reduced ability to reliably differentiate normal from abnormal sagittal relationships.

In contrast, the vertical relationship domain showed poor discrimination across all AI models, with AUC values remaining close to 0.50, reflecting near-random classification performance for overbite assessment. Similarly, transverse relationship parameters, including unilateral crossbite, demonstrated limited discriminatory ability, with ROC curves clustering near the diagonal reference line.

The arch morphology domain exhibited the weakest overall performance, with AUC values failing to reach thresholds indicative of meaningful discrimination. These findings indicate that current multimodal AI models struggle to interpret complex dental arch characteristics and symmetry from intraoral photographs.

Table 8. Integrated ROC-AUC performance of multimodal AI models across orthodontic domains

Orthodontic domain	**Parameters included**	**ChatGPT 5.2 Pro (AUC)**	**Gemini 3 Fast (AUC)**	**Claude 4.5 Sonnet (AUC)**	**Microsoft Copilot (AUC)**	**Overall interpretation**
Alignment	Crowding (Max, Mand), Diastema (Max)	0.60-0.80	0.69-0.85	0.68-0.77	0.56-0.65	Good
Sagittal relationships	Overjet, M1 (R/L), Canine (R/L), Angle classification	0.65-0.70	0.53-0.65	0.47-0.55	0.48-0.52	Moderate
Vertical relationship	Overbite	0.55	0.44	0.49	0.50	Poor
Transverse relationships	Crossbite (R/L)	0.49-0.61	0.50-0.54	0.48-0.60	0.54-0.58	Poor
Arch morphology	Symmetry (Max, Mand)	0.41-0.68	0.43-0.61	0.47-0.60	0.44-0.63	Poor

Note: Max maxillary; Mand mandibular; R right; L left; AUC area under curve

Cohen's kappa analysis (Table 9) revealed overall poor to moderate agreement between the AI models and orthodontist assessment across orthodontic domains. The highest level of agreement was observed in the alignment domain, where Gemini achieved a maximum κ value of 0.63, representing moderate agreement for alignment-related parameters. ChatGPT and Claude demonstrated lower but comparable agreement levels, while Copilot showed minimal concordance with orthodontist assessment.

In the sagittal relationship domain, agreement remained limited across all models. ChatGPT demonstrated the highest agreement, with κ values reaching up to 0.38 for sagittal parameters, including molar and canine relationships. However, this level of agreement was insufficient to support consistent clinical classification, highlighting challenges in AI-based interpretation of sagittal occlusal relationships.

Agreement in the vertical and transverse relationship domains was uniformly poor, with κ values clustering around zero or negative values, indicating minimal agreement with orthodontist assessment. Similarly, the arch morphology domain demonstrated poor agreement, with κ values failing to reach thresholds indicative of meaningful concordance.

Taken together, the combined ROC-AUC and kappa analyses indicate that although multimodal AI models can moderately discriminate certain orthodontic features, particularly within the alignment domain, their agreement with expert orthodontic assessment remains limited, especially for parameters requiring precise spatial interpretation.

Table 9. Summary of agreement (Cohen's kappa) between AI models and orthodontist assessment by orthodontic domains

Orthodontic domain	**Parameters included**	**ChatGPT 5.2 Pro (κ)**	**Gemini 3 Fast (κ)**	**Claude 4.5 Sonnet (κ)**	**Microsoft Copilot (κ)**	**Overall agreement**
Alignment	Crowding (Max, Mand), Diastema (Max)	0.11-0.33	0.15-0.63	0.16-0.25	0.00-0.17	Poor-Moderate
Sagittal relationships	Overjet, M1 (R/L), Canine (R/L), Angle classification	0.20-0.38	0.08-0.21	−0.03-0.07	−0.04-0.23	Poor-Moderate
Vertical relationship	Overbite	0.13	−0.15	0.05	−0.01	Poor
Transverse relationships	Crossbite (R/L)	−0.02-0.13	0.00-0.05	0.09-0.11	0.03-0.07	Poor
Arch morphology	Symmetry (Max, Mand)	0.07-0.27	−0.15-0.15	0.07-0.15	−0.07-0.19	Poor

Note: Max maxillary; Mand mandibular; R right; L left; AUC area under curve

Figure 3 illustrates the comparative diagnostic performance of the four multimodal AI models using both receiver operating characteristic (ROC) curves and a radar chart summarizing key classification metrics. The ROC curve analysis demonstrates clear differences in overall discriminatory ability among the evaluated models. ChatGPT exhibited the most pronounced deviation from the diagonal reference line, indicating superior discrimination between normal and abnormal conditions compared with the other models. Gemini showed comparable but slightly lower performance, whereas Claude and Microsoft Copilot demonstrated more modest discrimination, with ROC curves closer to the diagonal.

The radar chart further summarizes model performance across five diagnostic metrics: sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and overall accuracy. Across these metrics, ChatGPT consistently demonstrated the most balanced and robust performance profile, achieving higher sensitivity and accuracy while maintaining favourable specificity and predictive values. Gemini showed competitive performance, particularly in sensitivity and PPV, but exhibited slightly reduced balance across all metrics compared with ChatGPT.

Claude and Copilot demonstrated comparatively lower overall performance, characterized by reduced sensitivity and accuracy, despite maintaining moderate specificity and PPV. Notably, differences among models were more pronounced for sensitivity and overall accuracy than for specificity, suggesting that false-negative detection remains a key limitation for several AI models.

Taken together, the combined ROC and radar chart analyses indicate that although all evaluated multimodal AI models possess some capacity to discriminate orthodontic conditions, their diagnostic performance varies substantially. Models with stronger ROC performance also tended to exhibit more balanced classification metrics, highlighting the importance of evaluating both discrimination and predictive reliability when assessing AI-based orthodontic diagnostic tools.

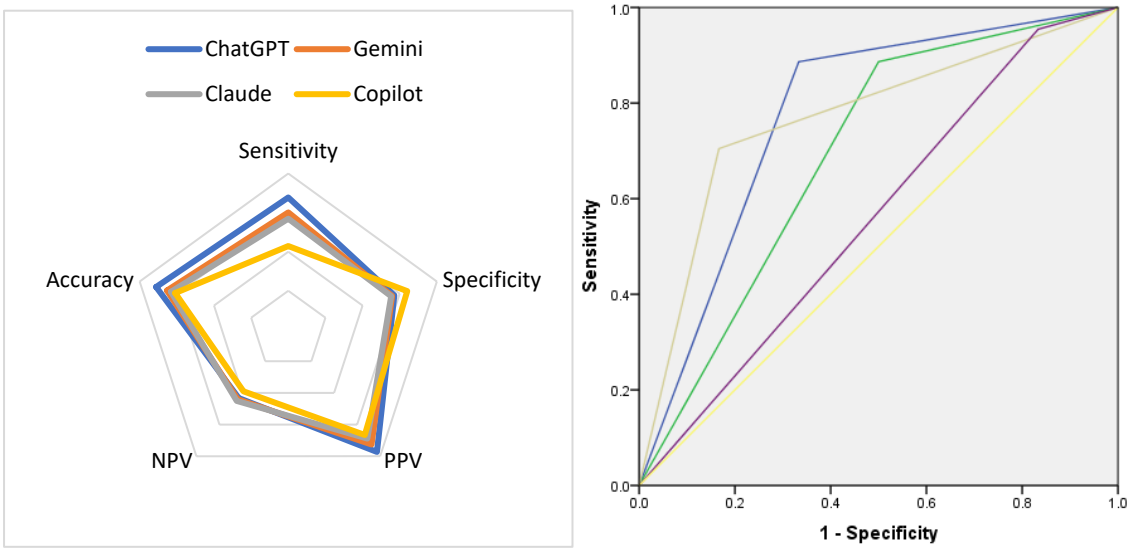


Figure 3. Comparative diagnostic performance of AI models based on classification metrics and ROC curve. Each panel compares the discriminatory performance of ChatGPT 5.2 Pro, Gemini 3 Fast, Claude 4.5 Sonnet, and Microsoft Copilot.

Overall, the results demonstrate a consistent pattern of domain-dependent performance across all evaluated multimodal AI models. Visually prominent orthodontic features, particularly those within the alignment domain, were associated with higher discriminatory

ability and moderate agreement with orthodontist assessment. In contrast, parameters requiring precise spatial interpretation, including sagittal, vertical, transverse, and arch morphology features, showed limited discrimination and poor agreement across models. These findings highlight the strengths and inherent limitations of general-purpose multimodal AI systems in orthodontic image interpretation and provide a foundation for the subsequent discussion on their clinical applicability and methodological constraints.

IV. Discussion

This study offers an exploratory analysis regarding the clinical readiness of the multiview intraoral photography collection and the diagnostic ability of multimodal large language models in assessing orthodontic malocclusion. This dataset we have describes a high level of clinical variability in terms of alignment, sagittal, vertical, transverse, and arch morphology, aspects that are crucial for effectively analyzing artificial intelligence (AI) tools [1]. Significantly, the data sample consisted of both male and female subjects of different ages, supplemented with multi-angle images, allowing for clear investigation of both simple and complex orthodontic details [6].

Looking across these domains, a clear trend is present. Those parameters associated with anterior alignment, such as crowding and diastema, showed greater agreement (κ 0.00-0.63) and accuracy (0.56-0.85). Recent studies have found that geometrically precise parameters, namely molars and canines, and the shapes of the dental arches showed suboptimal performance in the form of low AUC values, according to [36]. The implications of these findings support the hypothesis that general multi-modal learning models and LLMs can be useful in preliminary orthodontic analysis, but there are quite striking variations in some areas of orthodontic expertise [1]. This is supported by a recent assessment of LLMs using large-scale benchmarks, which shows that high performance in broad benchmarks does not necessarily translate to effective reasoning capabilities in scientific and clinical domains [37].

Our results suggest that the evaluative performance of large language models is stronger for anterior than for posterior relationships. These results are consistent with previous studies that have examined AI capabilities in orthodontics. For example, Hack et al. (2024) showed that AI models are superior at locating conspicuously visible attributes like front tooth position but are challenged by complex occlusal relationships that demand strong spatial references [1]. Similarly, Stetzel et al. (2024) found that deep learning models are successful at predicting aesthetic parts of the Index of Orthodontic Treatment Need (IOTN), which illustrate AI capabilities in visually-oriented analyses [38]. Notwithstanding, this outcome conflicts with that demonstrated in this study, particularly concerning sagittal and transverse parameter assessments, which are consistent with previous assessments that argued AI models are challenged when assessing parameters that depend on strong spatial references [36].

This study also demonstrated variations in agreement and accuracy across different large language models (Tables 8 and 9). These inter-model differences may be attributed to variations in underlying architectures and training paradigms. For example, Transformer-based LLMs, such as Vision Transformers (ViT), take a holistic approach to images, efficiently uncovering general irregularities in dental alignment, although they are challenged by making subtle distinctions in densely dependent orthodontic categories such as overjet and molar intercuspation [39]. Once again, this is consistent with previous reviews showing that CNNs trained using architecture-specific data excel at tasks requiring anatomical specificity compared to general AI systems such as LLMs [9]. The observed differences in outputs among the large language models appear to be attributable to variations in visual abstraction, depth of reasoning, and internal representational capacities, rather than true conceptual understanding [40]. The aggregated results are thus likely to represent the representational constraints

presented by their non-specialized training data [41]. Notably, cross-model analysis in scientific artificial intelligence research has revealed high error correlations among leading large language models (LLMs), indicating shared inductive biases rather than independent failure modes [37].

In terms of healthcare informatics, the assessment of several LLMs together can be seen to represent the "hive mind." While each one has different abilities and capacities, certain broad trends become apparent: strong abilities to detect anomalies in alignment and very limited abilities to detect complex spatial characteristics [1,42]. These findings point to the need to learn from all models together and also to recognize the limitations of AI and the importance of human input [41]. The observed convergence across models further suggests that aggregating general-purpose LLMs may offer limited benefit for inherently spatial clinical tasks, reinforcing the need for task-specific AI systems ([37]. Future studies should concentrate on building task-focused orthodontic AI systems rather than on testing general-purpose LLMs. The next important area would be to develop appropriate deep learning models based on standardized intraoral photographs with multiple views, which would be appropriately labeled by specialists in the orthodontic field [36]. This would make available appropriate data for building a supervised learning model that would employ spatial reasoning at a finer scale with clinically interpretable outputs [38] that require joint effort on the part of orthodontists and IT personnel.

This study has several limitations. Image quality variability, reliance on a single orthodontist for reference annotation, and the relatively small sample size may have influenced the performance estimates. Additionally, some AI models generated unsolicited treatment suggestions rather than purely diagnostic outputs, reflecting the limitations of output control. Notably, none of the models explicitly acknowledged diagnostic uncertainty or provided clinical disclaimers, which raises important considerations regarding ethical AI deployment in healthcare.

Funding

This research received no external funding

Conflict of interest

The authors declare no conflict of interest

Reference

1. Monill-González A, Rovira-Calatayud L, d'Oliveira NG, Ustrell Torrent JM. Artificial Intelligence in Orthodontics: Where Are We Now? A Scoping Review. Orthod Craniofac Res. 2021;24 Suppl 2:6-15. https://doi.org/10.1111/ocr.12517
2. Albalawi F, Alamoud K. Trends and Application of Artificial Intelligence Technology in Orthodontic Diagnosis and Treatment Planning—A Review. Applied Sciences. 2022;12(22):11864. https://doi.org/10.3390/app122211864
3. Dimberg L, Lennartsson B, Arnrup K, Bondemark L. Prevalence and change of malocclusions from primary to early permanent dentition: A longitudinal study. Angle Orthodontist. 2015 Sep 1;85(5):728–34. https://doi.org/10.2319/080414-542.1
4. Subramanian AK, Chen Y, Almalki A, Sivamurthy G, Kafle D. Cephalometric Analysis in Orthodontics Using Artificial Intelligence - A Comprehensive Review. Biomed Res Int. 2022;2022:1880113. https://doi.org/10.1155/2022/1880113

5. Hwang H, Moon J, Kim M, Donatelli RE, Lee S. Evaluation of Automated Cephalometric Analysis Based on the Latest Deep Learning Method. Angle Orthod. 2021;91(3):329-335. https://doi.org/10.2319/021220-100.1
6. Ryu J, Kim YH, Kim T, Jung S. Evaluation of Artificial Intelligence Model for Crowding Categorization and Extraction Diagnosis Using Intraoral Photographs. Sci Rep. 2023;13(5177):1-10 https://doi.org/10.1038/s41598-023-32514-7
7. Hack M, Drăgulin B, Hack L, ElSaafin M, Dumitrescu I, Stan D, Păcurar M. Comparative study on the results of orthodontic diagnostics by using algorithms generated by Artificial Intelligence and simple algorithms. Med Pharm Rep. 2024;97(2):215–21. https://doi.org/10.15386/mpr-2702
8. Liu J, Zhang C, Shan Z. Application of Artificial Intelligence in Orthodontics: Current State and Future Perspectives. Vol. 11, Healthcare (Switzerland). Multidisciplinary Digital Publishing Institute (MDPI). 2023;11(20):2760. https://doi.org/10.3390/healthcare11202760
9. Zhang R, Zhang L, Zhang D, Wang Y, Huang YS, Wang D, Li X. Development and Evaluation of a Deep Learning Model for Occlusion Classification in Intraoral Photographs. PeerJ. 2025; ;13:e20140. https://doi.org/10.7717/peerj.20140
10. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the Feasibility of ChatGPT in Healthcare: An Analysis of Multiple Clinical and Research Scenarios. J Med Syst. 2023;47(1):33. https://doi.org/10.1007/s10916-023-01925-4
11. Giannakopoulos K, Kavadella A, Salim AA, Stamatopoulos V, Kaklamanos EG. Evaluation of the Performance of Generative AI Large Language Models ChatGPT, Google Bard, and Microsoft Bing Chat in Supporting Evidence-Based Dentistry: Comparative Mixed Methods Study. J Med Internet Res. 2023;25:e51580. https://doi.org/10.2196/51580
12. Zheng J, Ding X, Pu JJ, Chung SM, H. Ai QY, Hung KF, Shan Z. Unlocking the Potentials of Large Language Models in Orthodontics: A Scoping Review. Bioengineering. 2024;11(11):1145. https://doi.org/10.3390/bioengineering11111145
13. Wu X, Huang Y, He Q. A large language model improves clinicians' diagnostic performance in complex critical illness cases. Crit Care. 2025 Dec 1; 29(1):230.https://doi.org/10.1186/s13054-025-05468-7
14. Hatia A, Doldo T, Parrini S, Chisci E, Cipriani L, Montagna L, Lagana G, Guenza G, Agosta E, Vinjolli F, Hoxha M, D'Amelio C, Favaretto N, Chisci G. Accuracy and Completeness of ChatGPT-Generated Information on Interceptive Orthodontics: A Multicenter Collaborative Study. J Clin Med. 2024 Feb 1;13(3). https://doi.org/10.3390/jcm13030735
15. Arısan A. Orthodontic Biomechanical Reasoning With Multimodal Language Models: Performance and Clinical Utility. Bioengineering. 2025; 12(11):1165. https://doi.org/10.3390/bioengineering12111165
16. Sun L, Wong HM, McGrath C. A Cohort Study of Factors That Influence Oral Health-Related Quality of Life From Age 12 to 18 in Hong Kong. Health Qual Life Outcomes. 2020;18(1):65 https://doi.org/10.1186/s12955-020-01317-z
17. Bayat JT, Huggare J, Akrami N. Distinguishing Between Global and Dental Self-Esteem in Evaluating Malocclusions. Acta Odontol Scand. 2019;77(6):452-456. https://doi.org/10.1080/00016357.2019.1588371
18. Choi S, Cha J, Lee K -J., Yu H, Hwang C. Changes in Psychological Health, Subjective Food Intake Ability and Oral Health-related Quality of Life During Orthodontic Treatment. J Oral Rehabil. 2017;44(11):860-869. https://doi.org/10.1111/joor.12556
19. Jasim ES, Kadhom ZM, Al-Groosh D. Comparative evaluation of orthodontic treatment needs index and the dental aesthetic index to assess the need for orthodontic treatment from the participants' perspective: A cross-sectional study. J Orthod Sci. 2025 Sep 1;14(1). https://doi.org/10.4103/jos.jos_24_25

20. Peter H. Brook, William C. Shaw. The development of an index of orthodontic treatment priority. Eur J Orthod. 1989 Aug 1;11(3):309–20. https://doi.org/10.1093/oxfordjournals.ejo.a035999
21. Nguyen MS, Nguyen MK, Saag M, Jagomägi T. The Need for Orthodontic Treatment Among Vietnamese School Children and Young Adults. Int J Dent. 2014; 2014:132301. https://doi.org/10.1155/2014/132301
22. Kozanecka A, Sarul M, Kawala B, Antoszewska-Smith J. Objectification of Orthodontic Treatment Needs: Does the Classification of Malocclusions or a History of Orthodontic Treatment Matter? Adv Clin Exp Med. 2016;25(6):1303-1312. https://doi.org/10.17219/acem/62828
23. Stanley Braun, William P. Hnat, Dana E. Fender, Harry L. Legan. The form of the human dental arch. Angle Orthod. 1998 Feb;68:29–36.
24. Huynh N, Zhang J, Pliska BT, Amin R, Narang I, Chadha NK, Cholette M, Kirk V, Montpetit A, Vézina K, Jacob S V, Laberge S, Hamoda MM, Almeida FR. Prevalence of Altered Craniofacial Morphology in Children With OSA. J Sleep Res. 2025;34(5):e70060. https://doi.org/10.1111/jsr.70060
25. Shah NH, Entwistle DA, Pfeffer MA. Creation and Adoption of Large Language Models in Medicine. JAMA. 2023;330(9):866-869. https://doi.org/10.1001/jama.2023.14217
26. Open AI. GPT-5.2 [Internet]. 2025 [cited 2025 Dec 26]. p. 1–1. Available from: https://openai.com/id-ID/index/introducing-gpt-5-2/
27. Carrillo-Pérez F, Pecho ÓE, Morales JC, Paravina RD, Bona Á Della, Ghinea R, Pulgar R, Mar Gómez M del, Herrera LJ. Applications of Artificial Intelligence in Dentistry: A Comprehensive Review. J Esthet Restor Dent. 2022 Jan;34(1):259-280. https://doi.org/10.1111/jerd.12844
28. Anthropic PBC. Claude sonnet 4.5 [Internet]. 2025 [cited 2025 Dec 26]. p. 1–1. Available from: https://www.anthropic.com/claude/sonnet
29. Google. Gemini 3 Fast [Internet]. 2025 [cited 2025 Dec 26]. p. 1–1. Available from: https://ai.google.dev/gemini-api/docs/gemini-3?hl=id
30. Ossowska A, Kusiak A, Świetlik D. Artificial Intelligence in Dentistry—Narrative Review. Int J Environ Res Public Health. 2022;19(6):3449. https://doi.org/10.3390/ijerph19063449
31. Microsoft. Microsoft Copilot [Internet]. 2025 [cited 2025 Dec 26]. p. 1–1. Available from: https://www.microsoft.com/id-id/microsoft-copilot/for-individuals/?form=MA13YT
32. Kirschneck C, Kuhr K, Ohm C, Baudisch NF, Jordan AR. Comparison of Orthodontic Treatment Need and Malocclusion Prevalence According to KIG, ICON, and mIOTN in German 8- To 9-Year-Old Children of the Sixth German Oral Health Study (DMS 6). J Orofac Orthop. 2023 Jan;84(Suppl 1):26-35. https://doi.org/10.1007/s00056-023-00446-6
33. Masood Y, Masood M, Binti Zainul NN, Abdul Araby NB, Hussain SF, Newton T. Impact of Malocclusion on Oral Health Related Quality of Life in Young People. Health Qual Life Outcomes. 2013;11:25. https://doi.org/10.1186/1477-7525-11-25
34. Perrotta S, Bucci R, Simeon V, Martina S, Michelotti A, Valletta R. Prevalence of Malocclusion, Oral Parafunctions and Temporomandibular Disorder-pain in Italian Schoolchildren: An Epidemiological Study. J Oral Rehabil. 2019;46(7):611-616. https://doi.org/10.1111/joor.12794
35. Cenzato N, Nobili A, Maspero C. Prevalence of Dental Malocclusions in Different Geographical Areas: Scoping Review. Dent J (Basel). 2021;9(10):117. https://doi.org/10.3390/dj9100117
36. Vishwanathaiah S, Fageeh HN, Khanagar SB, Maganur PC. Artificial Intelligence Its Uses and Application in Pediatric Dentistry: A Review. Biomedicines. 2023;11(3):788. https://doi.org/10.3390/biomedicines11030788

37. Song Z, Lu J, Du Y, Yu B, Pruyn TM, Huang Y, et al. Evaluating Large Language Models in Scientific Discovery. 2025 Dec 17; https://doi.org/10.48550/arXiv.2512.15567
38. Stetzel L, Foucher F, Jang SJ, Wu TH, Fields H, Schumacher F, Richmond S, Ko CC. Artificial Intelligence for Predicting the Aesthetic Component of the Index of Orthodontic Treatment Need. Bioengineering. 2024 Sep 1;11(9). https://doi.org/10.3390/bioengineering11090861
39. 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. 2020; https://doi.org/10.48550/arxiv.2010.11929
40. Azizi S. Applications of Artificial Intelligence in Diagnosis and Treatment Planning of Orthodontics: A Narrative Review. Saudi Dent J. 2025;37(7-9):70 https://doi.org/10.1007/s44445-025-00077-0
41. Ahmed AE, Aljohani WF, Abu Rukbah LK, Rajhi SA, Najmi NK, Zughlul MK, et al. The Role of Artificial Intelligence in Stroke Imaging in Emergency Settings: A Systematic Review. Cureus. 2025;17(10) https://doi.org/10.7759/cureus.93941
42. Jiang L, Chai Y, Li M, Liu M, Fok R, Dziri N, et al. Artificial Hivemind: The Open-Ended Homogeneity of Language Models (and Beyond). 2025 Oct 27; https://doi.org/10.48550/arXiv.2510.22954

2. Bukti kegagalan pemeriksaan teknis dan permintaan resubmit artikel

8 Januari 2026

UNIVERSITAS TRISAKTI

Joko Kusnoto <joko.k@trisakti.ac.id>

Technical Check failure: Your submission entitled Accuracy of Orthodontic Malocclusion Detection Using Multiple AI Models: A Comparative Study

2 messages

Healthcare Informatics Research <em@editorialmanager.com> Thu, Jan 8, 2026 at 9:23 AM
Reply-To: Healthcare Informatics Research <hir@kosmi.org>
To: Joko Kusnoto <joko.k@trisakti.ac.id>

Dear Dr Kusnoto,

Your submission entitled "Accuracy of Orthodontic Malocclusion Detection Using Multiple AI Models: A Comparative Study" has been received by the HIR, however, it is being returned to you for the following reason(s):

- **Main text length:** Must be under 3,000 words
- Number of references : **30 or fewer**
- **Reference format:** Please refer to the Instructions for Authors (https://e-hir.org/authors/authors.php#:~:text=Examples%20of%20reference%20formats)
- Number of tables and figures : **10 in total**

Please address the above issue(s) prior to resubmitting your manuscript.
Thank you for submitting your work to Healthcare Informatics Research.

Kind regards,
Editorial office
Healthcare Informatics Research

3. Bukti konfirmasi submit revisi dan artikel yang diresubmit

15 Januari 2026

UNIVERSITAS TRISAKTI

Joko Kusnoto <joko.k@trisakti.ac.id>

Technical Check failure: Your submission entitled Accuracy of Orthodontic Malocclusion Detection Using Multiple AI Models: A Comparative Study

2 messages

Healthcare Informatics Research <em@editorialmanager.com> Thu, Jan 8, 2026 at 9:23 AM
Reply-To: Healthcare Informatics Research <hir@kosmi.org>
To: Joko Kusnoto <joko.k@trisakti.ac.id>

Dear Dr Kusnoto,

Your submission entitled "Accuracy of Orthodontic Malocclusion Detection Using Multiple AI Models: A Comparative Study" has been received by the HIR, however, it is being returned to you for the following reason(s):

- **Main text length:** Must be under 3,000 words
- Number of references : **30 or fewer**
- **Reference format:** Please refer to the Instructions for Authors (https://e-hir.org/authors/authors.php#:~:text=Examples%20of%20reference%20formats)
- Number of tables and figures : **10 in total**

Please address the above issue(s) prior to resubmitting your manuscript.
Thank you for submitting your work to Healthcare Informatics Research.

Kind regards,
Editorial office
Healthcare Informatics Research

Joko Kusnoto <joko.k@trisakti.ac.id> Thu, Jan 15, 2026 at 12:11 AM
To: Healthcare Informatics Research <hir@kosmi.org>

Dear Editorial Office Healthcare Informatics Research,

We would like to inform you that we have already revised our manuscript "Accuracy of Orthodontic Malocclusion Detection Using Multiple AI Models: A Comparative Study" according to the instructions given and have already resubmitted our manuscript through HIR website. Hopefully our revision satisfies your requirement and our manuscript can get into the HIR review process. Thank you for your kind attention.

Sincerely yours,
Joko Kusnoto

[Quoted text hidden]

Abstract

Objective: This study aimed to evaluate and compare the accuracy of multiple AI models (ChatGPT 5.2 Pro, Gemini 3 Fast, Claude 4.5 Sonnet, and Microsoft Copilot) in detecting orthodontic malocclusion features in standardized, multi-view, intraoral photographs. The reference standard was an orthodontist assessment. **Methods:** A cross-sectional observational study was conducted using five standardized intraoral photographs (frontal, right lateral, left lateral, maxillary occlusal, and mandibular occlusal) obtained from 50 children aged 9-12 years. The following eight malocclusion parameters were assessed: anterior crowding, diastema, overjet, overbite, molar relationship, canine relationship, crossbite, and dental-arch symmetrical. The diagnostic accuracy and agreement between each AI model and the orthodontist was evaluated using Cohen's kappa and AUC. **Results:** The agreement between the AI models and the orthodontist ranged from poor to moderate across all orthodontic domains, with Cohen's κ -0.15 to 0.63. Visually prominent alignment features, including anterior crowding and diastema, demonstrated comparatively higher agreement (κ 0.00-0.63) and discriminatory performance, with ROC-AUC values ranging from 0.56 to 0.85. In contrast, parameters requiring precise spatial interpretation such as sagittal relationships, overbite, crossbite, and arch morphology showed consistently low agreement (κ -0.15 to 0.38) and poor to near-random classification performance, with AUC values predominantly between 0.41 and 0.70, and in some cases approaching 0.50. **Conclusion:** Current multimodal AI models demonstrate limited and parameter-dependent accuracy in detecting orthodontic malocclusions using intraoral photographs. These results emphasize the limitations of general-purpose AI systems for orthodontic decision support and highlight the necessity of task-specific models trained on clinically annotated datasets.

Keywords: accuracy, artificial intelligence, detection, malocclusion, orthodontic

III. Introduction

Malocclusion is a prevalent oral health problem among children and adolescents, with prevalence rates reported to be as high as 30-40% in some populations [1] . Early detection is crucial for preventing functional impairment and the need for more complex orthodontic interventions later in life [1,2]. Traditional orthodontic screening methods largely rely on direct clinical examinations by trained specialists, which can limit accessibility to necessary dental care, especially in community settings like schools [2].

Recent advancements in artificial intelligence (AI) have shown promise in improving diagnostic accuracy within orthodontics. This includes the application of machine learning and deep learning techniques to various imaging modalities. For instance, convolutional neural networks (CNN) have been effectively used in cephalometric analysis and panoramic image interpretation, yielding accurate automated orthodontic diagnoses [3]. However, there remains a gap in the utilization of intraoral photographs, which offer non-invasive, low-cost clinical records suitable for early screening and tele orthodontics [4,5] . The potential accuracy of intraoral photographs for diagnosing dental conditions has been recognized, with studies indicating strong diagnostic outcomes [6,7] .

In recent years, large language models (LLMs), such as ChatGPT and similar AI frameworks, have surfaced, demonstrating the capability to interpret both textual and visual inputs. Studies evaluating LLM responses in orthodontics reveal moderate to high consistency but significant variability in accuracy compared to human orthodontic specialists [8] . While these models can generate fluent and structured answers, they often provide incomplete or misleading information, thus necessitating a cautious approach to clinical interpretation [9–

11] . This emphasizes the importance of continuous evaluation and possible integration of LLMs in orthodontic practice, serving as an aid rather than a replacement for expert judgment.

Comparative studies indicate discrepancies between LLM-generated responses and those rendered by orthodontic experts, especially in malocclusion classification and treatment decision-making processes [9,12] . Despite the potential of LLMs for preliminary educational contributions, existing evidence suggests they cannot supplant the nuanced assessments made by experienced clinicians [1,3,11]. This aspect highlights the need for further investigations to rigorously validate these AI models' outputs against established clinical practices.

To evaluate the accuracy of malocclusion classification from clinical images via AI models, including advanced versions like (ChatGPT 5.2 Pro, Gemini 3 Fast, Claude 4.5 Sonnet, and Microsoft Copilot), standardized inputs are essential for comparison against expert evaluations. As of now, there is limited research focused on the performance of multimodal LLMs in analyzing standardized multi-view intraoral photographs against comprehensive orthodontic diagnostic criteria. Future studies are necessary to explore this domain [7,9].

IV. Method

1. Study design and population

This observational cross-sectional study was designed to evaluate the diagnostic performance of multimodal large language models (LLMs) in detecting orthodontic malocclusion features from standardized intraoral photographs. The study focused exclusively on image-based diagnostic interpretation without incorporating clinical examination, radiographic analysis, or treatment planning. The study was conducted using retrospective clinical photographs obtained from 50 children aged 9-12 years in Cimahi, West Java, Indonesia. No clinical intervention, treatment modification, or follow-up assessment was performed as part of this study.

2. Ethical approval and Consent to participate

All participants and their legal guardians provided written informed consent prior to the initiation of the study. Ethical approval for the study protocol was granted by the Ethics Committee of the Faculty of Dentistry, Universitas Trisakti (No. 1006/S3/KEPK/FKG/9/2025).

3. Image collection

Five images were acquired for each participant, comprising frontal, left lateral, right lateral, maxillary, and mandibular occlusal views of the teeth (in Figure 1). All images were captured using an iPhone 15 (Apple Inc., Cupertino, CA, USA) with a resolution of 6048 × 4032 pixels (24 megapixels). A single investigator, who was calibrated for this study, acquired all images according to standardized intraoral photography protocols to ensure consistent angulation, lighting, and field of view.

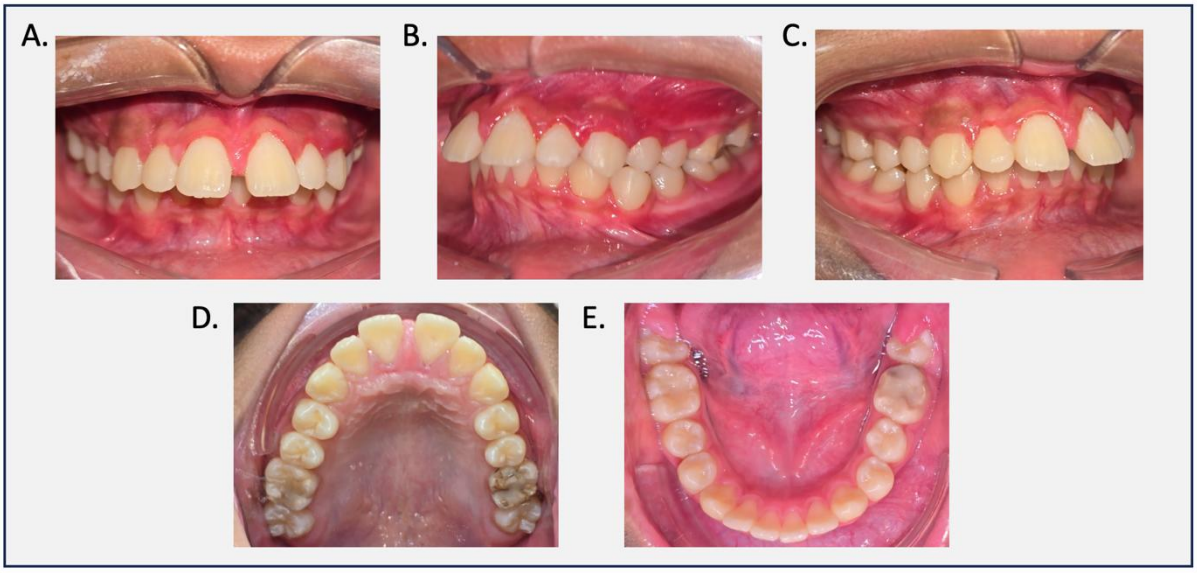


Figure 4. Sample of clinical images obtained from orthodontic patients. A. Frontal; B. Lateral left; C. Lateral right, D. Maxillary occlusal view, and E. Mandibular occlusal view.

4. Image preparation and preprocessing

All images were anonymized prior to analysis by removing information that could identify patients. No digital enhancement, filtering, contrast adjustment, or color correction was applied. However, images were cropped and resized due to the presence of other objects such as hands and mirrors during image acquisition. Images were analyzed at their original resolution to preserve clinically relevant visual features, such as tooth arrangement, occlusal relationships, and arch morphology.

5. Orthodontic diagnostic framework

Orthodontic assessment within this study employed a structured diagnostic framework based on established clinical standards, including Proffit's definitions of malocclusion, Angle's classification of sagittal relationships, and the Dental Health Component (DHC) of the Index of Orthodontic Treatment Need (IOTN) [13–15].

For analytical consistency and effective image-based evaluation, orthodontic parameters were systematically organized into five diagnostic domains that reflect varying levels of spatial complexity:

1. **Alignment Domain**: This domain included the assessment of anterior crowding and diastema in both the maxillary and mandibular arches. These features are commonly recognized indicators in orthodontic screening [16].
2. **Sagittal Relationship Domain**: This aspect comprised evaluating overjet, molar relationship, canine relationship, and overall Angle classification, focusing on anteroposterior dental relationships crucial for orthodontic diagnoses [17].
3. **Vertical Relationship Domain**: This domain emphasized overbite assessments categorized as normal, deep bite, or open bite, which denote significant vertical discrepancies in function [17].
4. **Transverse Relationship Domain**: In this domain, the presence and type of crossbite (anterior or posterior, unilateral or bilateral) were assessed, capturing transverse discrepancies often subtle in two-dimensional imaging [18].

5. **Arch Morphology Domain**: This domain evaluated dental arch symmetry through occlusal views, highlighting structural characteristics shaped by underlying skeletal patterns [19].

All parameters were systematically evaluated under this unified framework by both orthodontists and AI models, ensuring a structured approach. For statistical analysis, outcomes were dichotomized into normal and abnormal categories, facilitating a direct comparison of diagnostic agreement and discrimination [20].

6. Generative AI Multimodal large language models

In this study, four publicly available multimodal large language models (LLMs) with image-text processing capabilities were evaluated. The models included ChatGPT 5.2 Pro (OpenAI), Claude 4.5 Sonnet (Anthropic), Gemini 3 Fast (Google), and Microsoft Copilot (Microsoft) [12,21] . These models were accessed through their official web-based interfaces, adhering to their default system configurations without any application programming interfaces (APIs), external tools, or model fine-tuning procedures [12]. This ensures that the evaluation reflects a true representation of how these LLMs would operate in clinical contexts.

7. Model testing workflow and prompting strategy

For model evaluation, a structured approach to model evaluation was used, as follows in Figure 2 and listed in Table 1. Every intraoral image will be rated independently by each multimodal model, in a manner that gave no access to previous intraoral images or to outputs from previous models. This ensured that each model rating strictly depended on information contained within each intraoral image itself. Parameter-specific prompts were used for each category in orthodontics in accordance with an identical syntax and structure which aimed to capture the manner in which an orthodontist would logically reason. All prompts are in Indonesian language and have been identical for the whole study duration. Any outputs from each model are to be identified in predefined categories for orthodontics and reduced to two groups, normal or abnormal.

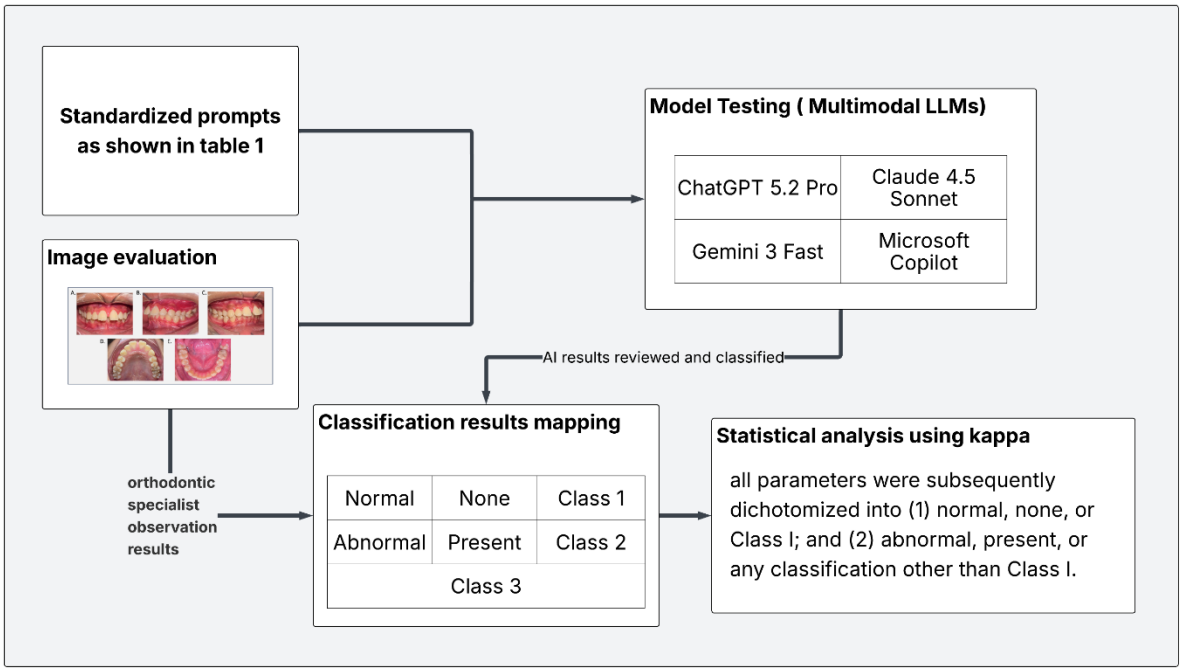


Figure 5. Schematic overview of multimodal LLM evaluation, result mapping, and agreement analysis.

Table 10. Standardized prompts used for multimodal LLM evaluation by orthodontic domain

Orthodontic domain	**Image view**	**Diagnostic focus**	**Standardized prompt**
Alignment	Frontal intraoral	Anterior crowding	"Analyze this frontal intraoral photograph. Focus on anterior crowding. Identify whether crowding is present in the maxillary and/or mandibular anterior teeth, specify the affected teeth, classify severity as mild, moderate, or severe, and provide a diagnostic conclusion."
Alignment	Frontal intraoral	Diastema	"Analyze this frontal intraoral photograph. Focus on diastema. Determine whether diastema is present, identify its location (maxillary and/or mandibular), and provide a diagnostic conclusion."
Sagittal relationship	Lateral intraoral (right/left)	Molar relationship (M1)	"Analyze this lateral intraoral photograph. Focus on the relationship between the maxillary first molar and mandibular first molar. Classify the molar relationship as Class I, II, or III. Report findings separately for the right and left sides."
Sagittal relationship	Lateral intraoral (right/left)	Canine relationship (C)	"Analyze this lateral intraoral photograph. Focus on the relationship between the maxillary canine and mandibular

			canine. Classify the canine relationship as Class I, II, or III. Report findings separately for the right and left sides."
Sagittal relationship	Lateral intraoral	Overjet	"Analyze this lateral intraoral photograph. Focus on overjet. Classify the overjet as normal, increased, edge-to-edge, or reverse, and provide a diagnostic conclusion."
Vertical relationship	Frontal intraoral	Overbite	"Analyze this frontal intraoral photograph. Focus on overbite. Classify the vertical overlap as normal, deep bite, or open bite, and provide a diagnostic conclusion."
Transverse relationship	Frontal intraoral	Crossbite	"Analyze this frontal intraoral photograph. Focus on transverse relationships. Identify the presence of crossbite, specify whether it is unilateral or bilateral, and provide a diagnostic conclusion."
Arch morphology	Maxillary occlusal	Arch symmetry	"Analyze this maxillary occlusal photograph. Assess the arch is symmetrical or asymmetrical, and provide a diagnostic conclusion."
Arch morphology	Mandibular occlusal	Arch symmetry	"Analyze this mandibular occlusal photograph. Assess the arch is symmetrical or asymmetrical, and provide a diagnostic conclusion."

8. Statistical analysis

The Kappa test was employed to assess the level of agreement between the artificial intelligence models and the orthodontist evaluations. This method is widely recognized in dental research to quantify inter-rater reliability, particularly in studies examining diagnostic discrepancies [7,22] . By establishing a comparative framework, the Kappa test provides insight into how effectively AI models align with expert clinical assessments, thus enhancing the understanding of AI's potential in orthodontics [23] . To evaluate the accuracy of each artificial intelligence model, all diagnostic variables were dichotomized into two response categories: (1) normal, none, or Class I; and (2) abnormal, present, or any classification other than Class I. This classification is significant, as it allows for a clearer identification of malocclusion types, aligning with established orthodontic standards and facilitating comparisons across studies [7,23] .

III. Results

A total of 50 childrens were included in this study. The participants' ages ranged from 9 to 12 years (Table 2). For each subject, five standardized intraoral photographs were obtained, comprising frontal, left lateral, right lateral, maxillary occlusal, and mandibular occlusal views. In total, 250 intraoral images were analyzed and used as input for the artificial intelligence models.

Table 11. subject demographic results

Variable	Value
Gender [n, %]	

Male	20 (40)
Female	30 (60)
Age (year) [mean, SD]	10.56 (0.73)

For all orthodontic fields, the data showed a high degree of clinical variability but represented a typical range of malocclusions (Table 3). For the alignment field of orthodontics, both arches showed a predominance of anterior crowding, mostly mild-to-moderate in nature. Diastemas were found more often in the mandible than the maxilla. Sagittal field examination showed a predominance of normal overjet and Angle Class I relationships, however a high proportion of individuals with Class II malocclusions and high overjet values was observed. Also in the sagittal field, a predominance of normal overbite was found with the deep bite as primary anomaly and open bite in rare occurrence. Crossbites are rare in the transverse field since the largest proportion of the population had no crossbites. When occurred, crossbites are mostly unilateral. The variability in arch type showed a preponderance of symmetry in the maxillary and mandibular arches. Taken together, the data represent a clinically variable but typical set appropriate for the comparative study of the performance of multi-model AI systems.

Table 12. Integrated distribution of orthodontic diagnostic characteristics across all domains.

Domain	**Parameter**	**n (%)**
Alignment	Maxillary crowding	
	None	6 (12)
	Mild	23 (46)
	Moderate	13 (26)
	Severe	8 (16)
	Mandibular crowding	
	None	6 (12)
	Mild	16 (32)
	Moderate	21 (42)
	Severe	7 (14)
	Diastema	
	Maxilla	4 (8)
	Mandible	12 (24)
Sagittal relationship	Overjet	
	Normal	31 (62)
	Increased	15 (30)
	Edge-to-edge	2 (4)
	Reverse	2 (4)
	Molar relationship (right)	
	Class I	31 (62)
	Class II	18 (36)

	Class III	1 (2)
	Molar relationship (left)	
	Class I	30 (60)
	Class II	18 (36)
	Class III	2 (4)
	Canine relationship (right)	
	Class I	28 (56)
	Class II	22 (44)
	Canine relationship (left)	
	Class I	30 (60)
	Class II	19 (38)
	Class III	1 (2)
	Angle classification	
	Class I	27 (54)
	Class II	21 (42)
	Class III	2 (4)
Vertical relationship	Overbite	
	Normal	31 (62)
	Open bite	1 (2)
	Deep bite	18 (36)
Transverse relationship	Crossbite (left)	
	None	41 (82)
	Anterior	6 (12)
	Posterior	2 (4)
	Anterior + posterior	1 (2)
	Crossbite (right)	
	None	41 (82)
	Anterior	6 (12)
	Posterior	3 (6)
Arch morphology	Arch symmetry	
	Maxilla (symmetrical)	39 (78)
	Mandible (symmetrical)	40 (80)

After the descriptive analysis, Receiver Operating Characteristic (ROC) analysis was performed to evaluate the discriminatory powers of multimodal models for normal vs. abnormal instances of orthodontic manifestations. Figure 3 highlights that there are significant domain-specific variations in model performance based upon the ROC curves. Parameters related to alignment had the highest discriminatory powers with larger distances of their respective ROC curves from the reference diagonal, whereas sagittal relationship parameters had moderate discriminatory powers. However, vertical, transverse, and arch morphology had nearly random classification accuracies with clustering of their respective ROC curves close to

the reference diagonal. Based on these results, it is concluded that multimodal models are efficient for visually more prominent orthodontic manifestations compared to complex spatial relationships.

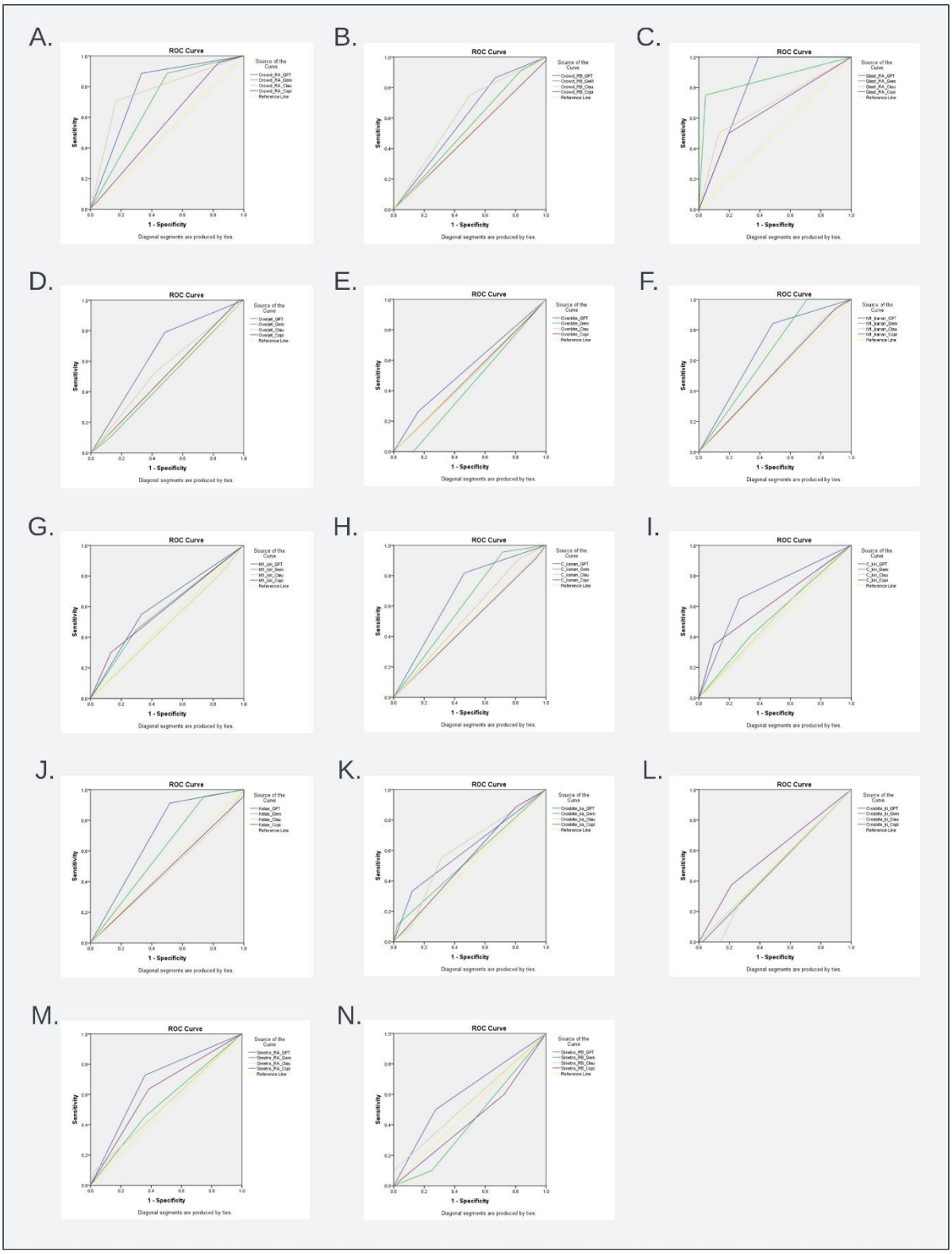


Figure 6. Integrated receiver operating characteristic (ROC) curves of multimodal AI models for orthodontic parameters. Panels A-O represent ROC curves for individual orthodontic parameters evaluated from standardized multi-view intraoral photographs. Panel A: maxillary

crowding; B: mandibular crowding; C: maxillary diastema; D: overjet; E: overbite; F: right first molar relationship; G: left first molar relationship; H: right canine relationship; I: left canine relationship; J: Angle classification; K: right crossbite; L: left crossbite; M: maxillary arch symmetry; N: mandibular arch symmetry. Each panel compares the discriminatory performance of ChatGPT 5.2 Pro, Gemini 3 Fast, Claude 4.5 Sonnet, and Microsoft Copilot against the orthodontist reference standard.

As shown in Table 4, the degree of discrimination achieved by the multimodal AI models varied significantly across the orthodontics tasks. The alignment domain had the highest degree of discrimination, with Gemini having the highest area under the curve (AUC) value of 0.85, followed by ChatGPT with AUC values approaching 0.80, which illustrated high discrimination abilities related to visually observable aspects like crowding and diastema. Although the values were similar, they were marginally lower for Claude, while Copilot had the lowest discrimination ability.

The ability to discriminate reduced in the sagittal relationship domain, in which the highest AUC values of up to 0.70 were recorded for ChatGPT, representing a fair ability to discriminate, and in which the remaining models performed worse. Conversely, a poor ability to discriminate was observed in both vertical and transverse relationship domains for all models, reflected in AUC values close to the random classification threshold of 0.50. Lastly, the ability to discriminate in the arch morphology domain was found to be the worst, in which AUC values did not reach levels associated with fair discriminant ability, reflecting the continued limitations of multimodal models in accurately processing complex dental geometry features.

Table 13. Integrated ROC-AUC performance of multimodal AI models across orthodontic domains

Orthodontic domain	**Parameters included**	**ChatGPT 5.2 Pro (AUC)**	**Gemini 3 Fast (AUC)**	**Claude 4.5 Sonnet (AUC)**	**Microsoft Copilot (AUC)**	**Overall interpretation**
Alignment	Crowding (Max, Mand), Diastema (Max)	0.60-0.80	0.69-0.85	0.68-0.77	0.56-0.65	Good
Sagittal relationships	Overjet, M1 (R/L), Canine (R/L), Angle classification	0.65-0.70	0.53-0.65	0.47-0.55	0.48-0.52	Moderate
Vertical relationship	Overbite	0.55	0.44	0.49	0.50	Poor
Transverse relationships	Crossbite (R/L)	0.49-0.61	0.50-0.54	0.48-0.60	0.54-0.58	Poor
Arch morphology	Symmetry (Max, Mand)	0.41-0.68	0.43-0.61	0.47-0.60	0.44-0.63	Poor

Note: Max maxillary; Mand mandibular; R right; L left; AUC area under curve

Data analysis of Cohen's κ statistics for both the multimodal models and orthodontist assessment showed poor to moderate agreement for all domains (Table 5). For alignment domains, there was generally poor agreement between models and orthodontist assessment, with moderate agreement for model Gemini and κ values of 0.63 at best for orthodontic alignment domains, demonstrating moderate agreement in orthodontic alignment domains by model Gemini compared to lower agreement but similar to that of models ChatGPT and Claude, in addition to minimal agreement for model Copilot. While there was also poor agreement for models in sagittal relationship domains, ChatGPT demonstrated relatively better agreement with κ values of 0.38 at best in orthodontic sagittal relationship domains compared to models Gemini and Claude, in addition to minimal agreement for models Copilot and ChatGPT. Conversely, there was poor agreement in vertical domains for all models compared to agreement in transverse domains of similar κ values near zero or lower for both vertical and transverse domains, similar minimal agreement for domains of arch morphology regarding agreement between models and assessment by human experts. Overall, the combined ROC–AUC and κ findings suggest that while multimodal AI models can moderately discriminate visually salient alignment features, their agreement with orthodontist evaluation remains limited, particularly for parameters requiring precise spatial interpretation.

Table 14. Summary of agreement (Cohen's kappa) between AI models and orthodontist assessment by orthodontic domains

Orthodontic domain	**Parameters included**	**ChatGPT 5.2 Pro (κ)**	**Gemini 3 Fast (κ)**	**Claude 4.5 Sonnet (κ)**	**Microsoft Copilot (κ)**	**Overall agreement**
Alignment	Crowding (Max, Mand), Diastema (Max)	0.11-0.33	0.15-0.63	0.16-0.25	0.00-0.17	Poor-Moderate
Sagittal relationships	Overjet, M1 (R/L), Canine (R/L), Angle classification	0.20-0.38	0.08-0.21	−0.03-0.07	−0.04-0.23	Poor-Moderate
Vertical relationship	Overbite	0.13	−0.15	0.05	−0.01	Poor
Transverse relationships	Crossbite (R/L)	−0.02-0.13	0.00-0.05	0.09-0.11	0.03-0.07	Poor
Arch morphology	Symmetry (Max, Mand)	0.07-0.27	−0.15-0.15	0.07-0.15	−0.07-0.19	Poor

Note: Max maxillary; Mand mandibular; R right; L left; AUC area under curve

Figure 4 illustrates a comparison between the discrimination abilities of the four AI models, which use Receiver Operating Characteristic (ROC) curves and radar charts. ROC analysis

indicates significant differences in discrimination ability among the AI models, whereby ChatGPT has the highest discrimination ability, followed by Gemini, while the abilities of Claude and Microsoft Copilot are limited, as shown by the ROC curves close to the reference line.

From the radar chart, the performance of the model with the best balance for sensitivity, specificity, positive predictive value, negative predictive value, and accuracy is the ChatGPT model. The model with the best competing sensitivity and positive and negative predictive values with somewhat reduced balance with regards to the indices used is the Gemini model. Finally, the model with the least overall performance regarding sensitivity and accuracy with somewhat reduced specificity is Claude and Copilot.

Thus, multimodal AI models demonstrate domain-specific performance with high discrimination for visually salient alignment features and low discrimination and agreement values for parameters that need accurate spatial interpretation, thereby emphasizing domain-specific limitations of general-purpose AI in the analysis of dental images.

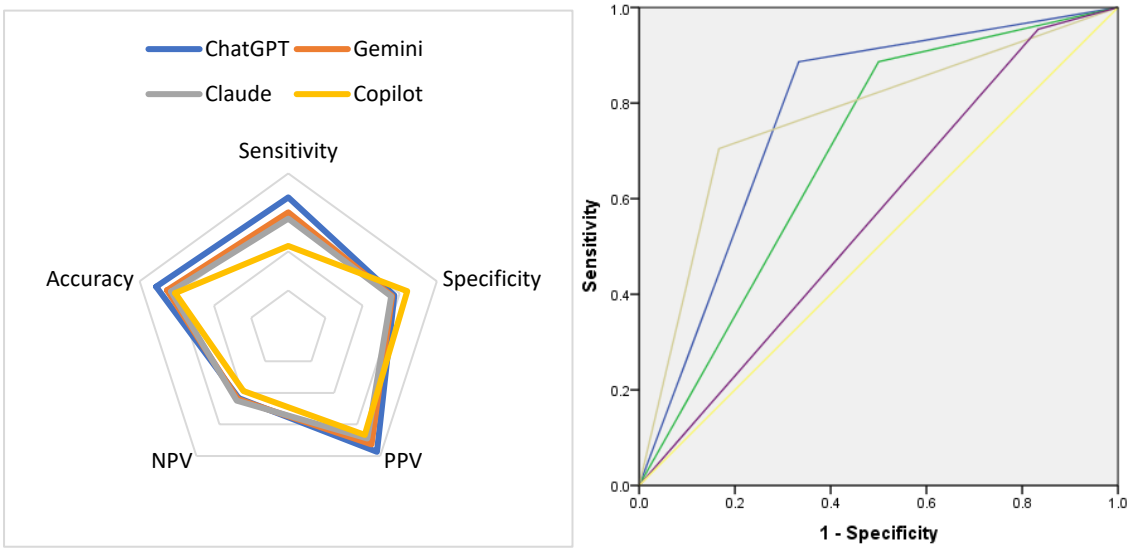


Figure 7. Comparative diagnostic performance of AI models based on classification metrics and ROC curve. Each panel compares the discriminatory performance of ChatGPT 5.2 Pro, Gemini 3 Fast, Claude 4.5 Sonnet, and Microsoft Copilot.

IV. Discussion

This study offers an exploratory analysis regarding the clinical readiness of the multiview intraoral photography collection and the diagnostic ability of multimodal large language models in assessing orthodontic malocclusion. This dataset we have describes a high level of clinical variability in terms of alignment, sagittal, vertical, transverse, and arch morphology, aspects that are crucial for effectively analyzing artificial intelligence (AI) tools [1]. Significantly, the data sample consisted of both male and female subjects of different ages, supplemented with multi-angle images, allowing for clear investigation of both simple and complex orthodontic details [6].

Looking across these domains, a clear trend is present. Those parameters associated with anterior alignment, such as crowding and diastema, showed greater agreement (κ 0.00-0.63) and accuracy (0.56-0.85). Recent studies have found that geometrically precise parameters, namely molars and canines, and the shapes of the dental arches showed suboptimal performance in the form of low AUC values, according to [24]. The implications of these

findings support the hypothesis that general multi-modal learning models and LLMs can be useful in preliminary orthodontic analysis, but there are quite striking variations in some areas of orthodontic expertise [1]. This is supported by a recent assessment of LLMs using large-scale benchmarks, which shows that high performance in broad benchmarks does not necessarily translate to effective reasoning capabilities in scientific and clinical domains [25].

Our results suggest that the evaluative performance of large language models is stronger for anterior than for posterior relationships. These results are consistent with previous studies that have examined AI capabilities in orthodontics. For example, Hack et al. (2024) showed that AI models are superior at locating conspicuously visible attributes like front tooth position but are challenged by complex occlusal relationships that demand strong spatial references [1]. Similarly, Stetzel et al. (2024) found that deep learning models are successful at predicting aesthetic parts of the Index of Orthodontic Treatment Need (IOTN), which illustrate AI capabilities in visually-oriented analyses [26]. Notwithstanding, this outcome conflicts with that demonstrated in this study, particularly concerning sagittal and transverse parameter assessments, which are consistent with previous assessments that argued AI models are challenged when assessing parameters that depend on strong spatial references [24].

This study also demonstrated variations in agreement and accuracy across different large language models (Tables 4 and 5). These inter-model differences may be attributed to variations in underlying architectures and training paradigms. For example, Transformer-based LLMs, such as Vision Transformers (ViT), take a holistic approach to images, efficiently uncovering general irregularities in dental alignment, although they are challenged by making subtle distinctions in densely dependent orthodontic categories such as overjet and molar intercuspation [27]. Once again, this is consistent with previous reviews showing that CNNs trained using architecture-specific data excel at tasks requiring anatomical specificity compared to general AI systems such as LLMs [7]. The observed differences in outputs among the large language models appear to be attributable to variations in visual abstraction, depth of reasoning, and internal representational capacities, rather than true conceptual understanding [28]. The aggregated results are thus likely to represent the representational constraints presented by their non-specialized training data [29]. Notably, cross-model analysis in scientific artificial intelligence research has revealed high error correlations among leading large language models (LLMs), indicating shared inductive biases rather than independent failure modes [25].

In terms of healthcare informatics, the assessment of several LLMs together can be seen to represent the “hive mind.” While each one has different abilities and capacities, certain broad trends become apparent: strong abilities to detect anomalies in alignment and very limited abilities to detect complex spatial characteristics [1,30]. These findings point to the need to learn from all models together and also to recognize the limitations of AI and the importance of human input [29]. The observed convergence across models further suggests that aggregating general-purpose LLMs may offer limited benefit for inherently spatial clinical tasks, reinforcing the need for task-specific AI systems [25]. Future studies should concentrate on building task-focused orthodontic AI systems rather than on testing general-purpose LLMs. The next important area would be to develop appropriate deep learning models based on standardized intraoral photographs with multiple views, which would be appropriately labeled by specialists in the orthodontic field [24]. This would make available appropriate data for building a supervised learning model that would employ spatial reasoning at a finer scale with clinically interpretable outputs [26] that require joint effort on the part of orthodontists and IT personnel.

This study has several limitations. Image quality variability, reliance on a single orthodontist for reference annotation, and the relatively small sample size may have influenced the performance estimates. Additionally, some AI models generated unsolicited treatment

suggestions rather than purely diagnostic outputs, reflecting the limitations of output control. Notably, none of the models explicitly acknowledged diagnostic uncertainty or provided clinical disclaimers, which raises important considerations regarding ethical AI deployment in healthcare.

Funding

This research received no external funding

Conflict of interest

The authors declare no conflict of interest

Reference

1. Monill-González A, Rovira-Calatayud L, d'Oliveira NG, Ustrell Torrent JM. Artificial Intelligence in Orthodontics: Where Are We Now? A Scoping Review. Orthod Craniofac Res. 2021;24 Suppl 2:6-15. https://doi.org/10.1111/ocr.12517
2. Albalawi F, Alamoud K. Trends and Application of Artificial Intelligence Technology in Orthodontic Diagnosis and Treatment Planning—A Review. Applied Sciences. 2022;12(22):11864. https://doi.org/10.3390/app122211864
3. Subramanian AK, Chen Y, Almalki A, Sivamurthy G, Kafle D. Cephalometric Analysis in Orthodontics Using Artificial Intelligence - A Comprehensive Review. Biomed Res Int. 2022;2022:1880113. https://doi.org/10.1155/2022/1880113
4. Hack M, Drăgulin B, Hack L, ElSaafin M, Dumitrescu I, Stan D, Păcurar M. Comparative study on the results of orthodontic diagnostics by using algorithms generated by Artificial Intelligence and simple algorithms. Med Pharm Rep. 2024;97(2):215–21.
5. Liu J, Zhang C, Shan Z. Application of Artificial Intelligence in Orthodontics: Current State and Future Perspectives. Vol. 11, Healthcare (Switzerland). Multidisciplinary Digital Publishing Institute (MDPI); 2023;11(20):2760. https://doi.org/10.3390/healthcare11202760
6. Ryu J, Kim YH, Kim T, Jung S. Evaluation of Artificial Intelligence Model for Crowding Categorization and Extraction Diagnosis Using Intraoral Photographs. Sci Rep. 2023;13(5177):1-10 https://doi.org/10.1038/s41598-023-32514-7
7. Zhang R, Zhang L, Zhang D, Wang Y, Huang YS, Wang D, Li X. Development and Evaluation of a Deep Learning Model for Occlusion Classification in Intraoral Photographs. PeerJ. 2025;13:e20140. https://doi.org/10.7717/peerj.20140
8. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the Feasibility of ChatGPT in Healthcare: An Analysis of Multiple Clinical and Research Scenarios. J Med Syst. 2023;47(1):33. https://doi.org/10.1007/s10916-023-01925-4
9. Zheng J, Ding X, Pu JJ, Chung SM, H. Ai QY, Hung KF, Shan Z. Unlocking the Potentials of Large Language Models in Orthodontics: A Scoping Review. Bioengineering. 2024;11(11):1145. https://doi.org/10.3390/bioengineering11111145
10. Wu X, Huang Y, He Q. A large language model improves clinicians' diagnostic performance in complex critical illness cases. Crit Care. 2025 Dec 1;29(1); 29(1):230.https://doi.org/10.1186/s13054-025-05468-7
11. Hatia A, Doldo T, Parrini S, Chisci E, Cipriani L, Montagna L, Lagana G, Guenza G, Agosta E, Vinjolli F, Hoxha M, D'Amelio C, Favaretto N, Chisci G. Accuracy and Completeness of ChatGPT-Generated Information on Interceptive Orthodontics: A Multicenter Collaborative Study. J Clin Med. 2024 Feb 1;13(3). https://doi.org/10.3390/jcm13030735
12. Giannakopoulos K, Kavadella A, Salim AA, Stamatopoulos V, Kaklamanos EG. Evaluation of the Performance of Generative AI Large Language Models ChatGPT, Google Bard, and Microsoft Bing Chat in Supporting Evidence-Based Dentistry: Comparative Mixed Methods Study. J Med Internet Res. 2023; 25:e51580. https://doi.org/10.2196/51580
13. Jasim ES, Kadhom ZM, Al-Groosh D. Comparative evaluation of orthodontic treatment needs index and the dental aesthetic index to assess the need for orthodontic treatment from the participants' perspective: A cross-sectional study. J Orthod Sci. 2025 Sep 1;14(1). https://doi.org/10.4103/jos.jos_24_25

14. Roy P, Roy P, Koley S, Sheet S. Bolton's ratio variations in Angle's Class I, Class II and Class III malocclusions: An observational study. J Clin Exp Dent. 2025;17(3):e280–5. https://doi.org/10.4317/jced.62591
15. Stetzel L, Foucher F, Jang SJ, Wu T, Fields HW, Schumacher FL, Richmond S, Ko C. Artificial Intelligence for Predicting the Aesthetic Component of the Index of Orthodontic Treatment Need. Bioengineering. 2024 Sep 1;11(9). https://doi.org/10.3390/bioengineering11090861
16. Peter H. Brook, William C. Shaw. The development of an index of orthodontic treatment priority. Eur J Orthod. 1989 Aug 1;11(3):309–20. https://doi.org/10.1093/oxfordjournals.ejo.a035999
17. Nguyen MS, Nguyen MK, Saag M, Jagomägi T. The Need for Orthodontic Treatment Among Vietnamese School Children and Young Adults. Int J Dent. 2014; 2014:132301. https://doi.org/10.1155/2014/132301
18. Kozanecka A, Sarul M, Kawala B, Antoszewska-Smith J. Objectification of Orthodontic Treatment Needs: Does the Classification of Malocclusions or a History of Orthodontic Treatment Matter? Advances in Clinical and Experimental Medicine. 2016;25(6):1303-1312. https://doi.org/10.17219/acem/62828
19. Stanley Braun, William P. Hnat, Dana E. Fender, Harry L. Legan. The form of the human dental arch. Angle Orthod. 1998 Feb;68:29–36.
20. Huynh N, Zhang J, Pliska BT, Amin R, Narang I, Chadha NK, Cholette M, Kirk V, Montpetit A, Vézina K, Jacob S V, Laberge S, Hamoda MM, Almeida FR. Prevalence of Altered Craniofacial Morphology in Children With OSA. J Sleep Res. 2025; 34(5):e70060. https://doi.org/10.1111/jsr.70060
21. Shah NH, Entwistle DA, Pfeffer MA. Creation and Adoption of Large Language Models in Medicine. JAMA. 2023;
22. Kirschneck C, Kuhr K, Ohm C, Baudisch NF, Jordan AR. Comparison of Orthodontic Treatment Need and Malocclusion Prevalence According to KIG, ICON, and mIOTN in German 8- To 9-Year-Old Children of the Sixth German Oral Health Study (DMS 6). Journal of Orofacial Orthopedics / Fortschritte Der Kieferorthopädie. 2023;330(9):866-869. https://doi.org/10.1001/jama.2023.14217
23. Masood Y, Masood M, Binti Zainul NN, Abdul Araby NB, Hussain SF, Newton T. Impact of Malocclusion on Oral Health Related Quality of Life in Young People. Health Qual Life Outcomes. 2013;11:25. https://doi.org/10.1186/1477-7525-11-25
24. Vishwanathaiah S, Fageeh HN, Khanagar SB, Maganur PC. Artificial Intelligence Its Uses and Application in Pediatric Dentistry: A Review. Biomedicines. 2023;11(3):788. https://doi.org/10.3390/biomedicines11030788
25. Song Z, Lu J, Du Y, Yu B, Pruyn TM, Huang Y, et al. Evaluating Large Language Models in Scientific Discovery. 2025 Dec 17; https://doi.org/10.48550/arXiv.2512.15567
26. Stetzel L, Foucher F, Jang SJ, Wu TH, Fields H, Schumacher F, Richmond S, Ko CC. Artificial Intelligence for Predicting the Aesthetic Component of the Index of Orthodontic Treatment Need. Bioengineering. 2024 Sep 1;11(9). https://doi.org/10.3390/bioengineering11090861
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. 2020; https://doi.org/10.48550/arxiv.2010.11929
28. Azizi S. Applications of Artificial Intelligence in Diagnosis and Treatment Planning of Orthodontics: A Narrative Review. Saudi Dent J. 2025;37(7-9):70 https://doi.org/10.1007/s44445-025-00077-0
29. Ahmed AE, Aljohani WF, Abu Rukbah LK, Rajhi SA, Najmi NK, Zughlul MK, Alshammari AM, Alotaibi S, Almarhabi TH, Alamri M, Rebh SB. The Role of Artificial Intelligence in Stroke Imaging in Emergency Settings: A Systematic Review. Cureus. 2025;17(10) https://doi.org/10.7759/cureus.93941
30. Jiang L, Chai Y, Li M, Liu M, Fok R, Dziri N, Tsvetkov Y, Sap M, Albalak A, Choi Y. Artificial Hivemind: The Open-Ended Homogeneity of Language Models (and Beyond). 2025 Oct 27; https://doi.org/10.48550/arXiv.2510.22954

4. Bukti permintaan revisi

19 Maret 2026

UNIVERSITAS TRISAKTI

Joko Kusnoto <joko.k@trisakti.ac.id>

Manuscript No. [HIR-26-031] Letter of decision

1 message

Healthcare Informatics Research <em@editorialmanager.com> Thu, Mar 19, 2026 at 11:57 PM
Reply-To: Healthcare Informatics Research <hir@kosmi.org>
To: Joko Kusnoto <joko.k@trisakti.ac.id>

Ref.: Ms. No. HIR-26-031
Accuracy of Orthodontic Malocclusion Detection Using Multiple AI Models: A Comparative Study
Healthcare Informatics Research

Dear Dr Kusnoto,

Reviewers have now commented on your paper. You will see that they are advising that you revise your manuscript. If you are prepared to undertake the work required, I would be pleased to reconsider my decision.

For your guidance, reviewers' comments are appended below.

If you decide to revise the work, please submit a list of changes or a rebuttal against each point which is being raised when you submit the revised manuscript.

Your revision is due by Apr 02, 2026.

To submit a revision, go to https://www.editorialmanager.com/hir/ and log in as an Author. You will see a menu item call 'Submission Needing Revision'. You will find your submission record there.
You can download the manuscript files from 'File Inventory' and you should revise your submission using these files.

Yours sincerely,

Hyejung Chang, Ph.D.
Editor-in-Chief
Healthcare Informatics Research

Comments from the Editors and Reviewers:

Editor: Reviewer #1: [Major Comments]
1. Since the LLM outputs vary across parameters, it would be helpful to explain how these outputs were mapped into binary categories (normal/none/class 1 vs. abnormal/present/other classes). Clarifying whether the mapping was performed manually by an orthodontist or using regular expressions would improve the reader's understanding.
2. In the Methods - 8. Statistical Analysis section, please specify how many orthodontists performed the evaluation.
[Minor Comments]
1. The ROC curves in Figure 3 are difficult to distinguish. Providing a higher-resolution figure would enhance visual clarity.
2. In the right panel of Figure 4, it would be helpful to differentiate the curves by color, similar to Figure 3, to make them easier to interpret.
3. Please correct "Childrens" to "children".

Reviewer #2: I have read the submission in its current form from the beginning to the end. The authors present a comparative study evaluating the accuracy of four AI models in detecting malocclusion from intraoral photographs. I think the finding that these models perform better on visually salient features but struggle with spatial reasoning is interesting. However, I have the following major and minor points for you to consider before consideration for publication in Healthcare Informatics Research.

Major points for you to consider>

1. The study uses a single orthodontist's assessment as the reference standard; the cohen's kappas are calculated in reference to this man or woman. This poses a major limitation to the interpretation of the cohen's kappa values which makes "poor agreement" unclear if it is due to the limitations of the AI models themselves or from potential misjudgement on the part of the reference orthodontist. You already mentioned this in the limitations; you should add more to it.

2. I don't see an justification of the sample size. Images from just 50 children may not be sufficient to assess the powers of general purpose large language models to begin with. You should clarify how you justified using images from only 50 children. Since data augmentation is much simpler for 2D images compared to texts, maybe you should consider augmentation if you cannot properly justify the use of the small sample size.

3. The preprocessing section would benefit from additional information on how exactly the images were standardized across samples.

Minor ponts for you to consider>

1. Table 4 ists "overall interpretation" as "poor", "moderate", or "good". Please define the specific AUC or Kappa thresholds used to assign these qualitative labels to ensure objective interpretation. Also, consider adding that a single orthodontist was involved as reference.

2. The authors correctly identify that 2D images limit spatial interpretation. I think the discussion would be strengthened by mentioning if the AI models were fed with all five view simultaneously in a single prompt session or if they were prompted image by image.

Reviewer #3: MAJOR COMMENTS
1. INTERPRETATION OF DOMAIN-SPECIFIC PERFORMANCE
The manuscript concludes that current multimodal AI models show limited performance in orthodontic diagnosis. However, the results show different performance across orthodontic domains.
For example, in the alignment domain (anterior crowding and diastema), Gemini 3 Fast achieved a Cohen's κ of 0.63, which indicates moderate agreement. In contrast, most other domains showed much lower agreement.
This suggests that AI performance may depend on the type of orthodontic parameter, rather than being uniformly poor across all tasks.
Required
(a) Clarify whether the study conclusion applies to all orthodontic parameters or mainly to parameters that require more complex spatial interpretation (e.g., sagittal relationships, crossbite, arch morphology).
(b) Expand the Discussion to explain why the alignment domain showed relatively better performance compared to other domains.
(c) Revise the interpretation so that the conclusion reflects that model performance appears parameter-dependent, instead of suggesting that performance is uniformly limited across all orthodontic tasks.

2. PROMPTING STRATEGY AND STUDY LIMITATION
The manuscript concludes that the observed performance of multimodal LLMs indicates the need for task-specific orthodontic AI models. However, the study evaluates the models using only a single standardized prompting framework.
Required (no additional experiments required):
(a) In the Limitations section, explicitly state that the evaluation reflects model performance under one specific prompting configuration used in this study.
(b) Clarify that alternative prompting strategies were not evaluated, and therefore the results represent performance under the current prompting setup.

Revision_Due.ics
1K

5. Bukti konfirmasi submit revisi, respon kepada reviewer, dan artikel yang diresubmit

27 Maret 2026

UNIVERSITAS TRISAKTI

Joko Kusnoto <joko.k@trisakti.ac.id>

Submission Confirmation for HIR-26-031R1

1 message

Healthcare Informatics Research <em@editorialmanager.com> Fri, Mar 27, 2026 at 1:29 PM
Reply-To: Healthcare Informatics Research <hir@kosmi.org>
To: Joko Kusnoto <joko.k@trisakti.ac.id>

Ref.: Ms. No. HIR-26-031R1
Accuracy of Orthodontic Malocclusion Detection Using Multiple AI Models: A Comparative Study

Dear Dr Kusnoto,

Healthcare Informatics Research has received your revised submission.

You may check the status of your manuscript by logging onto Editorial Manager at (https://www.editorialmanager.com/hir/).

Kind regards,

_____

Dear Editor and Reviewers,

We sincerely appreciate the time and effort devoted to reviewing our manuscript and providing insightful comments. We have thoroughly revised the manuscript to address all concerns raised. The revisions have improved the clarity, methodological transparency, and interpretation of our findings. All changes have been incorporated and highlighted in the revised manuscript. A detailed point-by-point response is provided below.

Reviewer 1

No	Comments	Author's response
1	Since the LLM outputs vary across parameters, it would be helpful to explain how these outputs were mapped into binary categories (normal/none/class 1 vs. abnormal/present/other classes). Clarifying whether the mapping was performed manually by an orthodontist or using regular expressions would improve the reader's understanding.	Thank you for your valuable comment. We have revised the manuscript. Due to the variability of outputs generated by multimodal LLMs, all responses were reviewed and mapped into two categories by a specialist orthodontist: (1) normal, including "normal," "none," or Class I; and (2) abnormal, including "present," "abnormal," or any classification other than Class I. Page 6 For statistical analysis, outcomes were dichotomized into normal and abnormal categories to enable direct comparison of diagnostic agreement and discrimination [20]. The detailed mapping of multi-class orthodontic parameters into binary categories is presented in Table 1.
2	In the Methods - 8. Statistical Analysis section, please specify how many orthodontists performed the evaluation.	Thank you for this comment. We have clarified in the manuscript that the reference standard was established by a single specialist orthodontist, as described in the Statistical Analysis section. Page 9

		The Kappa test is used to assess the degree of agreement between an artificial intelligence model and the assessment of a single orthodontist
3	The ROC curves in Figure 3 are difficult to distinguish. Providing a higher-resolution figure would enhance visual clarity.	Thank you for your valuable comment. We have revised the manuscript. To improve visual clarity, we have provided a high-resolution version of the figure in the Supplementary Materials (Supplementary Figure S1). In the main manuscript, Figure 3 has been retained in a simplified format to maintain readability and comply with figure size limitations. Page 13 Figure 3. Integrated receiver operating characteristic (ROC) curves of multimodal AI models for orthodontic parameters (see Supplementary Figure S1 for high-resolution version)
4	In the right panel of Figure 4, it would be helpful to differentiate the curves by color, similar to Figure 3, to make them easier to interpret.	Thank you for your valuable comment. We have revised the manuscript. Page 16 Figure 4 has been revised using color reflected in the upper panel and accompanying data. This revision improves clarity and avoids duplication in the visual presentation.
5	Please correct "Childrens" to "children".	Thank you for your valuable comment. We have revised the manuscript to ensure consistent use of the term “children” throughout the text.

Reviewer 2

No	Comments	Author’s response
1	The study uses a single orthodontist's assessment as the	Thank you for your valuable comment. We have revised the manuscript.

	reference standard; the cohen's kappas are calculated in reference to this man or woman. This poses a major limitation to the interpretation of the cohen's kappa values which makes "poor agreement" unclear if it is due to the limitations of the AI models themselves or from potential misjudgement on the part of the reference orthodontist. You already mentioned this in the limitations; you should add more to it.	We agree that the use of a single orthodontist as the reference standard represents a key limitation and may affect the interpretation of Cohen's kappa values. Specifically, the observed level of agreement may reflect not only the performance of the AI models but also potential subjectivity in expert judgment. To address this concern, we have expanded the Limitations section to explicitly acknowledge this issue. While the reference examiner is a calibrated orthodontist with substantial clinical and academic experience, we recognize that reliance on a single evaluator may introduce bias and limit the robustness of the reference standard. Future studies should use multiple orthodontists and assess inter-rater reliability to establish a more objective and reproducible ground truth. Page 20 This study has several limitations, including variability in image quality, the use of a single orthodontist as the reference standard, and a relatively small sample size, which may limit generalizability. The reliance on a single evaluator may introduce subjectivity, and thus the observed Cohen's kappa values may reflect both AI performance and variability in expert judgment. Future studies should include multiple orthodontists and assess inter-rater reliability to establish a more robust reference standard.
2	I don't see an justification of the sample size. Images from just 50 children may not be sufficient to assess the powers of general purpose large language models to begin with. You should clarify how you justified using images from	Thank you for your valuable comment. We have revised the manuscript. We have added a detailed sample size justification in the revised manuscript. A total of 50 participants were included, each contributing five standardized intraoral images, resulting in 250 images for analysis. Moreover, if we calculated the prevalence of malocclusion in the world (p0=93%), then compared this to Indonesia (p1=52.8%) with the total of 250 images, then the power of study would certainly be 100%.

	only 50 children. Since data augmentation is much simpler for 2D images compared to texts, maybe you should consider augmentation if you cannot properly justify the use of the small sample size.	Page 4 2. Sample size This is an observational cross-sectional study aimed at evaluating the diagnostic accuracy of multimodal large language models (LLMs) in detecting orthodontic malocclusion from intraoral images. The study sample consisted of 50 pediatric participants, each with five intraoral images, which were obtained from various angles, including frontal, right lateral, left lateral, maxillary occlusal, and mandibular occlusal. The total sample size is 250, which is sufficient to cover all important aspects of orthodontic malocclusion, including alignment, sagittal, vertical, transverse, and arch morphology.
3	The preprocessing section would benefit from additional information on how exactly the images were standardized across samples.	Thank you for your valuable comment. We have revised the manuscript. The images were standardized by acquiring five predefined intraoral views for each subject and subsequently cropping them to focus specifically on the dentition, as illustrated in Figure 1. This process has been clearly described in the revised manuscript in Section 5 (Image Preparation and Preprocessing). Page 5 All images were anonymized prior to analysis by removing information that could identify patients. All images were standardized without digital enhancement, filtering, contrast adjustment, or color correction. However, images were cropped due to the presence of other objects such as hands and mirrors during image acquisition. Images were analyzed at their original resolution to preserve clinically relevant visual features, such as tooth arrangement, occlusal relationships, and arch morphology

4	Table 4 lists "overall interpretation" as "poor", "moderate", or "good". Please define the specific AUC or Kappa thresholds used to assign these qualitative labels to ensure objective interpretation. Also, consider adding that a single orthodontist was involved as reference.	Thank you for your valuable comment. We have revised the manuscript. We have clarified the interpretation criteria in the manuscript. Cohen's kappa values were calculated by comparing each AI model's output with the reference standard established by a specialist orthodontist, as detailed in the Statistical Analysis section. This framework allows agreement to be interpreted in relation to expert clinical assessment. Table 5 and 6 Footnote AUC values were interpreted based on established ROC guidelines: ≥0.90 (excellent), 0.80–0.89 (good), 0.70–0.79 (fair), 0.60–0.69 (poor), and <0.60 (fail). Page 9 and 10 9. Statistical analysis The Kappa test was used to assess the degree of agreement between an artificial intelligence model and the assessment of a single orthodontist. This method is widely recognized in dental research to quantify inter-rater reliability, particularly in studies examining diagnostic discrepancies [7,22] . By establishing a comparative framework, the Kappa test provides insight into how effectively AI models align with expert clinical assessments, thus enhancing the understanding of AI's potential in orthodontics [22] . Model performance was evaluated by dichotomizing diagnostic outputs into normal and abnormal categories and calculating the area under the receiver operating characteristic curve (AUC) to quantify each model's discriminative ability. AUC values were interpreted according to established ROC criteria: ≥0.90 (excellent), 0.80–0.89 (good), 0.70–0.79 (fair), 0.60–0.69 (poor), and <0.60 (fail) [23].

5	The authors correctly identify that 2D images limit spatial interpretation. I think the discussion would be strengthened by mentioning if the AI models were fed with all five view simultaneously in a single prompt session or if they were prompted image by image.	Thank you for your valuable comment. We agree with this point. We did not upload the images simultaneously to LLM in a single prompt, instead of prompted image by image. In our experimental setup, the analysis and evaluation process was conducted by dividing the input into separate image views to obtain parameter-specific results, as structured in Section 6 (Orthodontic Diagnostic Framework). Each intraoral image was evaluated independently by each multimodal model through a blind-testing procedure, without access to other image views or prior model outputs. This approach ensured that each assessment was based solely on the visual information contained within the corresponding image. We have clarified this procedure in the revised manuscript in Section 8 (Model Testing Workflow and Prompting Strategy). Page 8 The analysis and testing of the AI models were conducted in stages based on the "Orthodontic Diagnostic Framework" described in Section 6. Each intraoral image was evaluated independently by each multimodal model through a blind-testing procedure

Reviewer 3

No	Comments	Author's response
1	INTERPRETATION OF DOMAIN-SPECIFIC PERFORMANCE	Thank you for your valuable comment.

	The manuscript concludes that current multimodal AI models show limited performance in orthodontic diagnosis. However, the results show different performance across orthodontic domains. For example, in the alignment domain (anterior crowding and diastema), Gemini 3 Fast achieved a Cohen's κ of 0.63, which indicates moderate agreement. In contrast, most other domains showed much lower agreement. This suggests that AI performance may depend on the type of orthodontic parameter, rather than being uniformly poor across all tasks. Required (a) Clarify whether the study conclusion applies to all orthodontic parameters or mainly to parameters that require more complex spatial interpretation (e.g., sagittal	(a) We agree with the reviewer that the performance of multimodal AI models is not uniform across all orthodontic parameters. We have revised the Conclusion section to clarify that the observed limitations primarily apply to parameters requiring complex spatial interpretation, such as sagittal relationships, crossbite, and arch morphology, rather than to all orthodontic tasks. (b) We have expanded the Discussion section to explain the relatively better performance observed in the alignment domain. Specifically, we clarified that alignment-related features, such as crowding and diastema, are visually salient and can be interpreted from two-dimensional image patterns without requiring complex spatial reasoning. In contrast, other orthodontic parameters depend on fine-grained spatial relationships and three-dimensional interpretation, which are more challenging for current multimodal AI systems. (c) We have revised the overall interpretation throughout the manuscript to reflect that model performance is parameter-dependent rather than uniformly limited. The revised text now emphasizes that multimodal AI models demonstrate variable performance across orthodontic domains, with relatively better results in visually oriented tasks and substantially lower performance in spatially complex assessments. Page 17 and 18 Looking across these domains, a clear trend is present. Those parameters associated with anterior alignment, such as crowding and diastema, showed greater agreement (κ 0.00 - 0.63) and accuracy (0.56 - 0.85). The relatively higher performance observed in alignment parameters may be attributed to their visually salient and inherently two-dimensional characteristics, which can be directly inferred from intraoral photographs without requiring complex spatial interpretation. Features such as crowding and diastema manifest as explicit variations in tooth spacing and positioning, making them well-suited for pattern recognition.

	relationships, crossbite, arch morphology). (b) Expand the Discussion to explain why the alignment domain showed relatively better performance compared to other domains. (c) Revise the interpretation so that the conclusion reflects that model performance appears parameter-dependent, instead of suggesting that performance is uniformly limited across all orthodontic tasks.	Furthermore, alignment assessment relies on relatively stable and easily identifiable anatomical landmarks, reducing variability across evaluators and analytical approaches. These features are also less susceptible to perspective distortion caused by variations in camera positioning, in contrast to sagittal relationships and arch morphology, which depend on more complex inter-arch spatial relationships and three-dimensional interpretation. In contrast, recent studies have found that geometrically precise parameters, such as molar and canine relationships, as well as dental arch morphology, show suboptimal performance, often reflected by lower AUC values, due to their reliance on fine-grained spatial relationships and three-dimensional interpretation that are difficult to infer from two-dimensional images [24]. As for conclusion, current multimodal AI models demonstrate limited and parameter-dependent accuracy in detecting orthodontic malocclusions using intraoral photographs. These results emphasize the limitations of general purpose AI systems for orthodontic decision support and highlight the necessity of task-specific models trained on clinically annotated datasets.
2	PROMPTING STRATEGY AND STUDY LIMITATION The manuscript concludes that the observed performance of multimodal LLMs indicates the need for task-specific orthodontic AI models. However, the study evaluates the models using only a single standardized prompting framework.	Thank you for your valuable comment. (a) We have updated the Limitations section to explicitly state that all models were evaluated using a single standardized prompting framework, and that the reported results reflect model performance under this specific configuration. (b) We have further clarified that alternative prompting strategies were not evaluated, and that different prompt designs may lead to variations in performance. This statement has been added to ensure that the findings are interpreted within the context of the current prompting setup.

	Required (no additional experiments required): (a) In the Limitations section, explicitly state that the evaluation reflects model performance under one specific prompting configuration used in this study. (b) Clarify that alternative prompting strategies were not evaluated, and therefore the results represent performance under the current prompting setup.	Page 20 This study has several limitations. Variability in image quality, the use of a single orthodontist as the reference standard, and the relatively small sample size may have influenced performance estimates and limited generalizability. In addition, some models generated unsolicited treatment suggestions rather than strictly diagnostic outputs, highlighting limited output controllability. Furthermore, none of the models explicitly expressed diagnostic uncertainty or provided clinical disclaimers, raising important ethical considerations for the use of multimodal AI in healthcare settings. Moreover, all models were evaluated using a single standardized prompting framework. Therefore, the reported results reflect model performance under this specific prompting configuration only. Alternative prompting strategies were not evaluated, and different prompt designs may lead to variations in performance, however ensemble prompt may be used as an alternative. This limitation should be considered when interpreting the findings

Accuracy of Orthodontic Malocclusion Detection Using Multiple AI Models: A Comparative Study

Abstract

Objective: This study aimed to evaluate and compare the accuracy of multiple AI models (ChatGPT 5.2 Pro, Gemini 3 Fast, Claude 4.5 Sonnet, and Microsoft Copilot) in detecting orthodontic malocclusion features in standardized, multi-view, intraoral photographs. The reference standard was an orthodontist assessment. **Methods:** A cross-sectional observational study was conducted using five standardized intraoral photographs (frontal, right lateral, left lateral, maxillary occlusal, and mandibular occlusal) obtained from 50 children aged 9-12 years. The following eight malocclusion parameters were assessed: anterior crowding, diastema, overjet, overbite, molar relationship, canine relationship, crossbite, and dental-arch symmetrical. The diagnostic accuracy and agreement between each AI model and the orthodontist was evaluated using Cohen's kappa and AUC. **Results:** The agreement between the AI models and the orthodontist ranged from poor to moderate across all orthodontic domains, with Cohen's κ -0.15 to 0.63. Visually prominent alignment features, including anterior crowding and diastema, demonstrated comparatively higher agreement (κ 0.00-0.63) and discriminatory performance, with ROC-AUC values ranging from 0.56 to 0.85. In contrast, parameters requiring precise spatial interpretation such as sagittal relationships, overbite, crossbite, and arch morphology showed consistently low agreement (κ -0.15 to 0.38) and poor to near-random classification performance, with AUC values predominantly between 0.41 and 0.70, and in some cases approaching 0.50. **Conclusion:** Current multimodal AI models demonstrate limited and parameter-dependent accuracy in detecting orthodontic malocclusions using intraoral photographs. These results emphasize the limitations of general-purpose AI systems for orthodontic decision support and highlight the necessity of task-specific models trained on clinically annotated datasets.

Keywords: accuracy, artificial intelligence, detection, malocclusion, orthodontic

I. Introduction

Malocclusion is a prevalent oral health problem among children and adolescents, with prevalence rates reported to be as high as 30-40% in some populations [1]. Early detection is crucial for preventing functional impairment and the need for more complex orthodontic interventions later in life [1,2]. Traditional orthodontic screening methods largely rely on direct clinical examinations by trained specialists, which can limit accessibility to necessary dental care, especially in community settings like schools [2].

Recent advancements in artificial intelligence (AI) have shown promise in improving diagnostic accuracy within orthodontics. This includes the application of machine learning and deep learning techniques to various imaging modalities. For instance, convolutional neural

networks (CNN) have been effectively used in cephalometric analysis and panoramic image interpretation, yielding accurate automated orthodontic diagnoses [3]. However, there remains a gap in the utilization of intraoral photographs, which offer non-invasive, low-cost clinical records suitable for early screening and tele orthodontics [4,5]. The potential accuracy of intraoral photographs for diagnosing dental conditions has been recognized, with studies indicating strong diagnostic outcomes [6,7].

In recent years, large language models (LLMs), such as ChatGPT and similar AI frameworks, have surfaced, demonstrating the capability to interpret both textual and visual inputs. Studies evaluating LLM responses in orthodontics reveal moderate to high consistency but significant variability in accuracy compared to human orthodontic specialists [8] . While these models can generate fluent and structured answers, they often provide incomplete or misleading information, thus necessitating a cautious approach to clinical interpretation [9–11] . This emphasizes the importance of continuous evaluation and possible integration of LLMs in orthodontic practice, serving as an aid rather than a replacement for expert judgment.

Comparative studies indicate discrepancies between LLM-generated responses and those rendered by orthodontic experts, especially in malocclusion classification and treatment decision-making processes [9,12] . Despite the potential of LLMs for preliminary educational contributions, existing evidence suggests they cannot supplant the nuanced assessments made by experienced clinicians [1,3,11]. This aspect highlights the need for further investigations to rigorously validate these AI models' outputs against established clinical practices.

To evaluate the accuracy of malocclusion classification from clinical images via AI models, including advanced versions like (ChatGPT 5.2 Pro, Gemini 3 Fast, Claude 4.5 Sonnet, and Microsoft Copilot), standardized inputs are essential for comparison against expert evaluations. As of now, there is limited research focused on the performance of multimodal LLMs in analyzing standardized multi-view intraoral photographs against comprehensive orthodontic

diagnostic criteria. Future studies are necessary to explore this domain [7,9].

II. Method

1. Study design and population

This observational cross-sectional study was designed to evaluate the diagnostic performance of multimodal large language models (LLMs) in detecting orthodontic malocclusion features from standardized intraoral photographs. The study focused exclusively on image-based diagnostic interpretation without incorporating clinical examination, radiographic analysis, or treatment planning. The study was conducted using retrospective clinical photographs obtained from 50 children aged 9-12 years in Cimahi, West Java, Indonesia. No clinical intervention, treatment modification, or follow-up assessment was performed as part of this study.

2. Sample size

This is an observational cross-sectional study aimed at evaluating the diagnostic accuracy of multimodal large language models (LLMs) in detecting orthodontic malocclusion from intraoral images. The study sample consisted of 50 pediatric participants, each with five intraoral images, which were obtained from various angles, including frontal, right lateral, left lateral, maxillary occlusal, and mandibular occlusal. The total sample size is 250, which was sufficient to cover all important aspects of orthodontic malocclusion, including alignment, sagittal, vertical, transverse, and arch morphology.

3. Ethical approval and Consent to participate

All participants and their legal guardians provided written informed consent prior to the initiation of the study. Ethical approval for the study protocol was granted by the Health Researches Ethics Committee of the Faculty of Dentistry, Universitas Trisakti (No. 1006/S3/KEPK/FKG/9/2025).

4. Image collection

Five images were acquired for each participant, comprising frontal, left lateral, right lateral, maxillary, and mandibular occlusal views of the teeth (in Figure 1). All images were captured using an iPhone 15 (Apple Inc., Cupertino, CA, USA) with a resolution of 6048 × 4032 pixels (24 megapixels). A single investigator, who was calibrated for this study, acquired all images according to standardized intraoral photography protocols to ensure consistent angulation, lighting, and field of view.

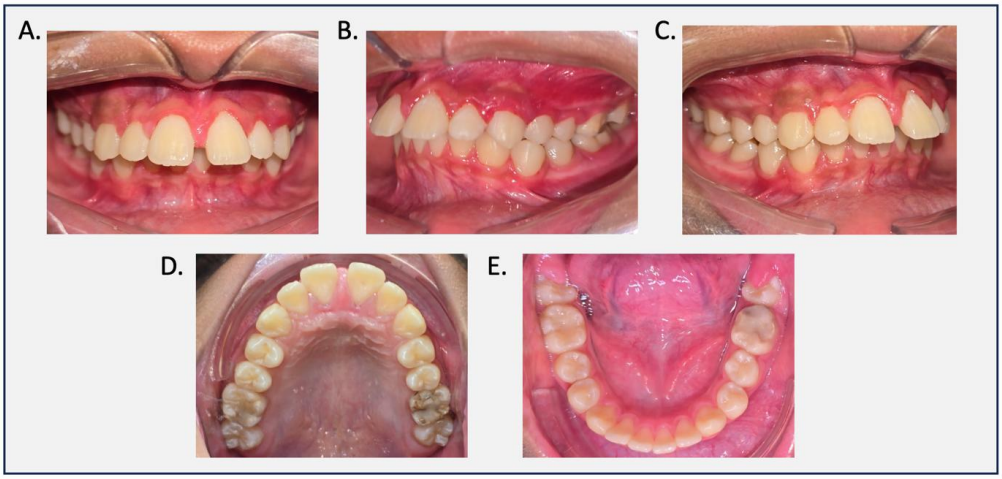


Figure 8. Sample of clinical images obtained from orthodontic patients. A. Frontal; B. Lateral left; C. Lateral right, D. Maxillary occlusal view, and E. Mandibular occlusal view

5. Image preparation and preprocessing

All images were anonymized prior to analysis by removing information that could identify patients. All images were standardized without digital enhancement, filtering, contrast adjustment, or color correction. However, images were cropped due to the presence of other objects such as hands and mirrors during image acquisition. Images were analyzed at their original resolution to preserve clinically relevant visual features, such as tooth arrangement, occlusal relationships, and arch morphology.

6. Orthodontic diagnostic framework

Orthodontic assessment within this study employed a structured diagnostic framework based on

established clinical standards, including Proffit's definitions of malocclusion, Angle's classification of sagittal relationships, and the Dental Health Component (DHC) of the Index of Orthodontic Treatment Need (IOTN) [13–15].

For analytical consistency and effective image-based evaluation, orthodontic parameters were systematically organized into five diagnostic domains that reflect varying levels of spatial complexity:

1. **Alignment Domain**: This domain included the assessment of anterior crowding and diastema in both the maxillary and mandibular arches. These features are commonly recognized indicators in orthodontic screening [16].
2. **Sagittal Relationship Domain**: This aspect comprised evaluating overjet, molar relationship, canine relationship, and overall Angle classification, focusing on anteroposterior dental relationships crucial for orthodontic diagnoses [17].
3. **Vertical Relationship Domain**: This domain emphasized overbite assessments categorized as normal, deep bite, or open bite, which denote significant vertical discrepancies in function [17].
4. **Transverse Relationship Domain**: In this domain, the presence and type of crossbite (anterior or posterior, unilateral or bilateral) were assessed, capturing transverse discrepancies often subtle in two-dimensional imaging [18].
5. **Arch Morphology Domain**: This domain evaluated dental arch symmetry through occlusal views, highlighting structural characteristics shaped by underlying skeletal patterns [19].

All parameters were systematically evaluated under this unified framework by both orthodontists and AI models, ensuring a structured approach. For statistical analysis, outcomes were dichotomized into normal and abnormal categories to enable direct comparison of diagnostic agreement and discrimination [20]. The detailed mapping of multi-class orthodontic parameters into binary categories is presented in Table 1.

Table 15. Mapping of multi-class orthodontic parameters into binary categories for statistical analysis

Domain	Parameter	Original Categories	Binary Category
Alignment	Maxillary crowding	None	Normal

		Mild, Moderate, Severe	Abnormal
	Mandibular crowding	None	Normal
		Mild, Moderate, Severe	Abnormal
	Diastema	None	Normal
		Presence (Maxilla or Mandible)	Abnormal
Sagittal	Overjet	Normal	Normal
		Increased, Edge-to-edge, Reverse	Abnormal
	Molar relationship (Right & Left)	Class I	Normal
		Class II, Class III	Abnormal
	Canine relationship (Right & Left)	Class I	Normal
		Class II, Class III	Abnormal
	Angle classification	Class I	Normal
		Class II, Class III	Abnormal
Vertical	Overbite	Normal	Normal
		Deep bite, Open bite	Abnormal
Transverse	Crossbite (Right & Left)	None	Normal
		Anterior, Posterior, Anterior+Posterior	Abnormal
Arch morphology	Arch symmetry	Symmetrical (Maxilla/Mandible)	Normal
		Asymmetrical	Abnormal

7. Generative AI Multimodal large language models

In this study, four publicly available multimodal large language models (LLMs) with image-text processing capabilities were evaluated. The models included ChatGPT 5.2 Pro (OpenAI), Claude 4.5 Sonnet (Anthropic), Gemini 3 Fast (Google), and Microsoft Copilot (Microsoft) [12,21] . These models were accessed through their official web-based interfaces, adhering to their default system configurations without any application programming interfaces (APIs), external tools, or model fine-tuning procedures [12]. This ensures that the evaluation reflects a true representation of how these LLMs would operate in clinical contexts.

8. Model testing workflow and prompting strategy

For model evaluation, a structured approach to model evaluation was used, as follows in Figure 2 and listed in Table 2. The analysis and testing of the AI models were conducted in stages based on the "Orthodontic Diagnostic Framework". Each intraoral image was evaluated independently by each multimodal model through a blind-testing procedure. This ensured that each model rating strictly depended on information contained within each intraoral image itself. Parameter-specific prompts were used for each category in orthodontics in accordance with an identical syntax and structure which aimed to capture the manner in which an orthodontist would logically reason. All prompts are in Indonesian language and have been identical for the whole study duration. Any outputs from each model are to be identified in predefined categories for orthodontics and reduced to two groups, normal or abnormal.

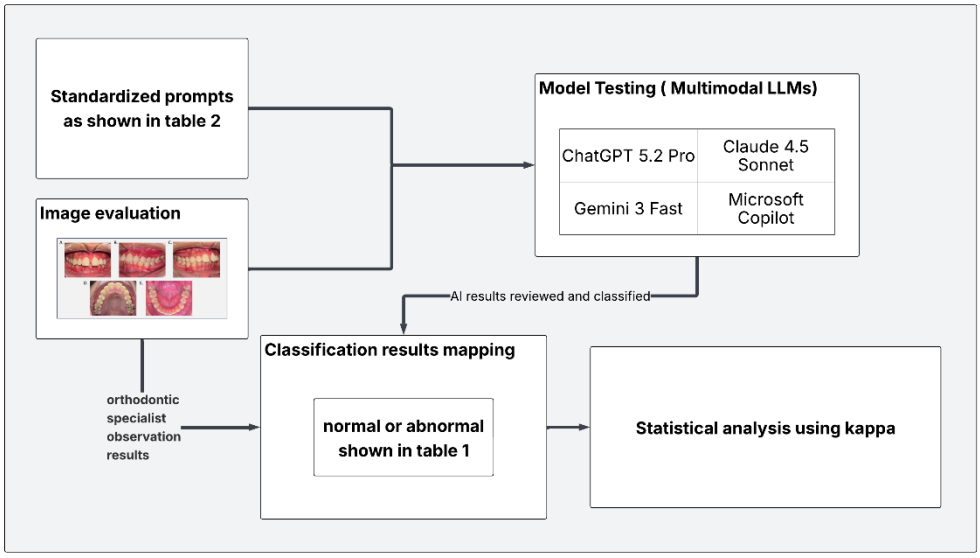


Figure 9. Schematic overview of multimodal LLM evaluation, result mapping, and agreement analysis

Table 16. Standardized prompts used for multimodal LLM evaluation by orthodontic domain

Orthodontic domain	Image view	Diagnostic focus	Standardized prompt
Alignment	Frontal intraoral	Anterior crowding	"Analyze this frontal intraoral photograph. Focus on anterior crowding. Identify whether crowding is present in the maxillary and/or mandibular anterior teeth, specify the

			affected teeth, classify severity as mild, moderate, or severe, and provide a diagnostic conclusion."
Alignment	Frontal intraoral	Diastema	"Analyze this frontal intraoral photograph. Focus on diastema. Determine whether diastema is present, identify its location (maxillary and/or mandibular), and provide a diagnostic conclusion."
Sagittal relationship	Lateral intraoral (right/left)	Molar relationship (M1)	"Analyze this lateral intraoral photograph. Focus on the relationship between the maxillary first molar and mandibular first molar. Classify the molar relationship as Class I, II, or III. Report findings separately for the right and left sides."
Sagittal relationship	Lateral intraoral (right/left)	Canine relationship (C)	"Analyze this lateral intraoral photograph. Focus on the relationship between the maxillary canine and mandibular canine. Classify the canine relationship as Class I, II, or III. Report findings separately for the right and left sides."
Sagittal relationship	Lateral intraoral	Overjet	"Analyze this lateral intraoral photograph. Focus on overjet. Classify the overjet as normal, increased, edge-to-edge, or reverse, and provide a diagnostic conclusion."
Vertical relationship	Frontal intraoral	Overbite	"Analyze this frontal intraoral photograph. Focus on overbite. Classify the vertical overlap as normal, deep bite, or open bite, and provide a diagnostic conclusion."
Transverse relationship	Frontal intraoral	Crossbite	"Analyze this frontal intraoral photograph. Focus on transverse relationships. Identify the presence of crossbite, specify whether it is unilateral or bilateral, and provide a diagnostic conclusion."
Arch morphology	Maxillary occlusal	Arch symmetry	"Analyze this maxillary occlusal photograph. Assess the arch is symmetrical or asymmetrical, and provide a diagnostic conclusion."
Arch morphology	Mandibular occlusal	Arch symmetry	"Analyze this mandibular occlusal photograph. Assess the arch is symmetrical or asymmetrical, and provide a diagnostic conclusion."

9. Statistical analysis

The Kappa test was used to assess the degree of agreement between an artificial intelligence model and the assessment of a single orthodontist. This method is widely recognized in dental research to quantify inter-rater reliability, particularly in studies examining diagnostic discrepancies [7]. By establishing a comparative framework, the Kappa test provides insight into how effectively AI models align with expert clinical assessments, thus enhancing the understanding of AI's potential in orthodontics [22]. Model performance was evaluated by dichotomizing diagnostic outputs into normal and abnormal categories and calculating the area under the receiver operating characteristic curve (AUC) to quantify each model's discriminative ability. AUC values were interpreted according to established ROC criteria: $\geq$ 0.90 (excellent), 0.80–0.89 (good), 0.70–0.79 (fair), 0.60–0.69 (poor), and <0.60 (fail) [23].

III. Results

A total of 50 children were included in this study. The participants' ages ranged from 9 to 12 years (Table 3). For each subject, five standardized intraoral photographs were obtained, comprising frontal, left lateral, right lateral, maxillary occlusal, and mandibular occlusal views. In total, 250 intraoral images were analyzed and used as input for the artificial intelligence models.

Table 17. subject demographic results

Variable	Value
Gender [n, %]	
Male	20 (40)
Female	30 (60)
Age (year) [mean, SD]	10.56 (0.73)

For all orthodontic fields, the data showed a high degree of clinical variability but represented a typical range of malocclusions (Table 4). For the alignment field of orthodontics, both arches showed a predominance of anterior crowding, mostly mild-to-moderate in nature. Diastemas were found more often in the mandible than the maxilla. Sagittal field examination showed a predominance of normal overjet and Angle Class I relationships, however a high proportion of individuals with Class II malocclusions and high overjet values was observed. Also in the sagittal field, a predominance of normal overbite was found with the deep bite as primary anomaly and open bite in rare occurrence. Crossbites are rare in the transverse field since the largest proportion of the population had no crossbites. When occurred, crossbites are mostly unilateral. The variability in arch type showed a preponderance of symmetry in the maxillary and mandibular arches. Taken together, the data represent a clinically variable but typical set appropriate for the comparative study of the performance of multi-model AI systems.

Table 18. Integrated distribution of orthodontic diagnostic characteristics across all domains

Domain	Parameter	n (%)
Alignment	Maxillary crowding	
	None	6 (12)
	Mild	23 (46)
	Moderate	13 (26)
	Severe	8 (16)
	Mandibular crowding	
	None	6 (12)
	Mild	16 (32)
	Moderate	21 (42)
	Severe	7 (14)
	Diastema	
	Maxilla	4 (8)
	Mandible	12 (24)
Sagittal relationship	Overjet	
	Normal	31 (62)
	Increased	15 (30)
	Edge-to-edge	2 (4)
	Reverse	2 (4)
	Molar relationship (right)	
	Class I	31 (62)
	Class II	18 (36)
	Class III	1 (2)
	Molar relationship (left)	
	Class I	30 (60)
	Class II	18 (36)
	Class III	2 (4)
	Canine relationship (right)	
	Class I	28 (56)
	Class II	22 (44)
	Canine relationship (left)	
	Class I	30 (60)
	Class II	19 (38)
	Class III	1 (2)
	Angle classification	
	Class I	27 (54)
	Class II	21 (42)
	Class III	2 (4)
Vertical relationship	Overbite	
	Normal	31 (62)
	Open bite	1 (2)
	Deep bite	18 (36)
Transverse relationship	Crossbite (left)	
	None	41 (82)
	Anterior	6 (12)
	Posterior	2 (4)
	Anterior + posterior	1 (2)
	Crossbite (right)	
	None	41 (82)
	Anterior	6 (12)

	Posterior	3 (6)
	Arch symmetry	
Arch morphology	Maxilla (symmetrical)	39 (78)
	Mandible (symmetrical)	40 (80)

After the descriptive analysis, Receiver Operating Characteristic (ROC) analysis was performed to evaluate the discriminatory powers of multimodal models for normal vs. abnormal instances of orthodontic manifestations. Figure 3 highlights that there are significant domain-specific variations in model performance based upon the ROC curves. Parameters related to alignment had the highest discriminatory powers with larger distances of their respective ROC curves from the reference diagonal, whereas sagittal relationship parameters had moderate discriminatory powers. However, vertical, transverse, and arch morphology had nearly random classification accuracies with clustering of their respective ROC curves close to the reference diagonal. Based on these results, it is concluded that multimodal models are efficient for visually more prominent orthodontic manifestations compared to complex spatial relationships.

As shown in Table 5, the degree of discrimination achieved by the multimodal AI models varied significantly across the orthodontics tasks. The alignment domain had the highest degree of discrimination, with Gemini having the highest area under the curve (AUC) value of 0.85, followed by ChatGPT with AUC values approaching 0.80, which illustrated high discrimination abilities related to visually observable aspects like crowding and diastema. Although the values were similar, they were marginally lower for Claude, while Copilot had the lowest discrimination ability.

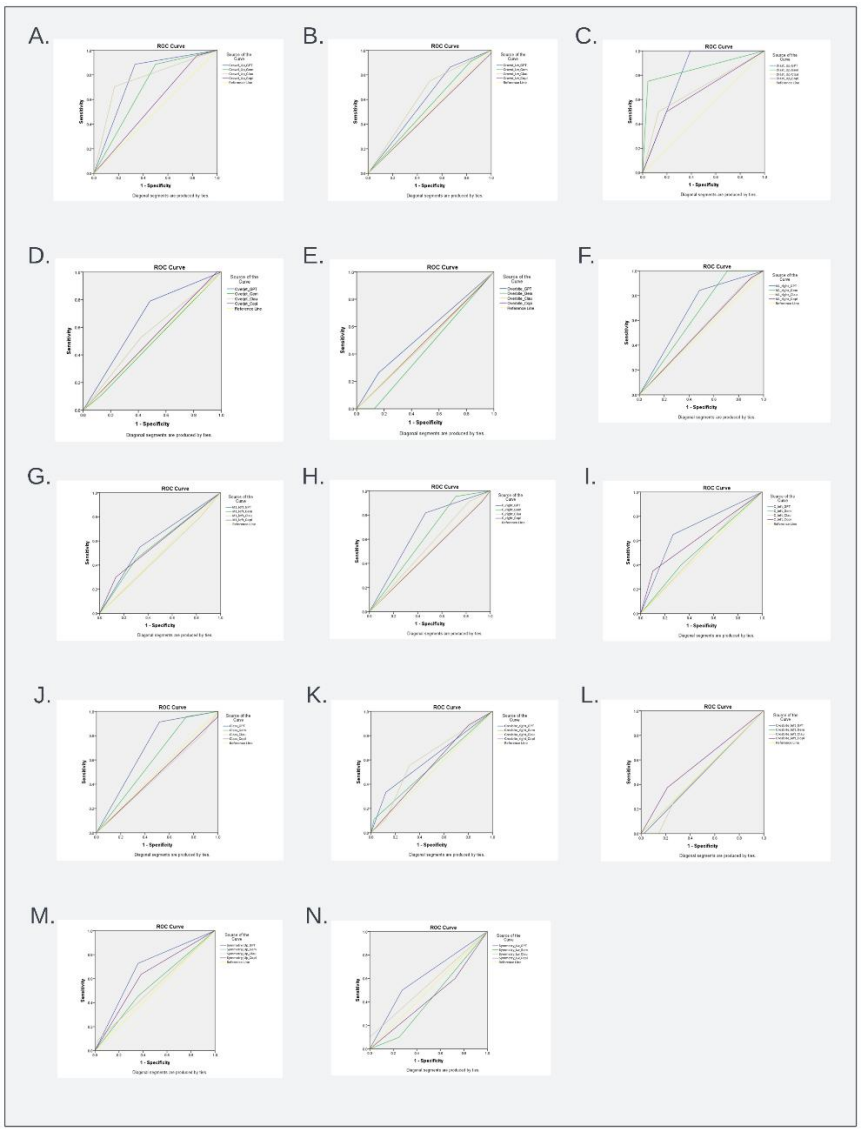


Figure 10. Integrated receiver operating characteristic (ROC) curves of multimodal AI models for orthodontic parameters (see Supplementary Figure S1 for high-resolution version). Panels A-O represent ROC curves for individual orthodontic parameters evaluated from standardized multi-view intraoral photographs. Panel A: maxillary crowding; B: mandibular crowding; C: maxillary diastema; D: overjet; E: overbite; F: right first molar relationship; G: left first molar relationship; H: right canine relationship; I: left canine relationship; J: Angle classification; K: right crossbite; L: left crossbite; M: maxillary arch symmetry; N: mandibular arch symmetry. Each panel compares the discriminatory performance of ChatGPT 5.2 Pro, Gemini 3 Fast, Claude 4.5 Sonnet, and Microsoft Copilot against the orthodontist reference standard

Table 19. Integrated ROC-AUC performance of multimodal AI models across orthodontic domains

Orthodontic domain	Parameters included	ChatGPT 5.2 Pro (AUC)	Gemini 3 Fast (AUC)	Claude 4.5 Sonnet (AUC)	Microsoft Copilot (AUC)	Overall interpretation
Alignment	Crowding (Max, Mand), Diastema (Max)	0.60-0.80	0.69-0.85	0.68-0.77	0.56-0.65	Poor - Good
Sagittal relationships	Overjet, M1 (R/L), Canine (R/L), Angle classification	0.65-0.70	0.53-0.65	0.47-0.55	0.48-0.52	Poor
Vertical relationship	Overbite	0.55	0.44	0.49	0.50	Fail
Transverse relationships	Crossbite (R/L)	0.49-0.61	0.50-0.54	0.48-0.60	0.54-0.58	Fail - Poor
Arch morphology	Symmetry (Max, Mand)	0.41-0.68	0.43-0.61	0.47-0.60	0.44-0.63	Fail - Poor

Note: Max maxillary; Mand mandibular; R right; L left; AUC area under curve;
AUC values were interpreted based on established ROC guidelines: ≥0.90 (excellent), 0.80‑0.89 (good), 0.70‑0.79 (fair), 0.60‑0.69 (poor), and <0.60 (fail).

The ability to discriminate reduced in the sagittal relationship domain, in which the highest AUC values of up to 0.70 were recorded for ChatGPT, representing a fair ability to discriminate, and in which the remaining models performed worse. Conversely, a fail ability to discriminate was observed in both vertical and transverse relationship domains for all models, reflected in AUC values close to the random classification threshold of 0.50. Lastly, the ability to discriminate in the arch morphology domain was found to be the worst, in which AUC values did not reach levels associated with fair discriminant ability, reflecting the continued limitations of multimodal models in accurately processing complex dental geometry features.

Data analysis of Cohen's κ statistics for both the multimodal models and orthodontist assessment showed poor to moderate agreement for all domains (Table 6). For alignment domains, there was generally poor agreement between models and orthodontist assessment, with moderate agreement for model Gemini and κ values of 0.63 at best for orthodontic alignment domains, demonstrating moderate agreement in orthodontic alignment domains by model Gemini compared to lower agreement but

similar to that of models ChatGPT and Claude, in addition to minimal agreement for model Copilot. While there was also poor agreement for models in sagittal relationship domains, ChatGPT demonstrated relatively better agreement with κ values of 0.38 at best in orthodontic sagittal relationship domains compared to models Gemini and Claude, in addition to minimal agreement for models Copilot and ChatGPT. Conversely, there was poor agreement in vertical domains for all models compared to agreement in transverse domains of similar κ values near zero or lower for both vertical and transverse domains, similar minimal agreement for domains of arch morphology regarding agreement between models and assessment by human experts. Overall, the combined ROC–AUC and κ findings suggest that while multimodal AI models can moderately discriminate visually salient alignment features, their agreement with orthodontist evaluation remains limited, particularly for parameters requiring precise spatial interpretation.

Table 20. Summary of agreement (Cohen's kappa) between AI models and orthodontist assessment by orthodontic domains

Orthodontic domain	Parameters included	ChatGPT 5.2 Pro (κ)	Gemini 3 Fast (κ)	Claude 4.5 Sonnet (κ)	Microsoft Copilot (κ)	Overall agreement
Alignment	Crowding (Max, Mand), Diastema (Max)	0.11-0.33	0.15-0.63	0.16-0.25	0.00-0.17	Poor-Moderate
Sagittal relationships	Overjet, M1 (R/L), Canine (R/L), Angle classification	0.20-0.38	0.08-0.21	−0.03-0.07	−0.04-0.23	Poor-Moderate
Vertical relationship	Overbite	0.13	−0.15	0.05	−0.01	Poor
Transverse relationships	Crossbite (R/L)	−0.02-0.13	0.00-0.05	0.09-0.11	0.03-0.07	Poor
Arch morphology	Symmetry (Max, Mand)	0.07-0.27	−0.15-0.15	0.07-0.15	−0.07-0.19	Poor

Note: Max maxillary; Mand mandibular; R right; L left; AUC area under curve;

Figure 4 illustrates a comparison between the discrimination abilities of the four AI models, which use radar charts. ROC analysis indicates significant differences in discrimination ability among the AI models, whereby ChatGPT has the highest discrimination ability, followed by Gemini, while the

abilities of Claude and Microsoft Copilot are limited, as shown by the ROC curves close to the reference line.

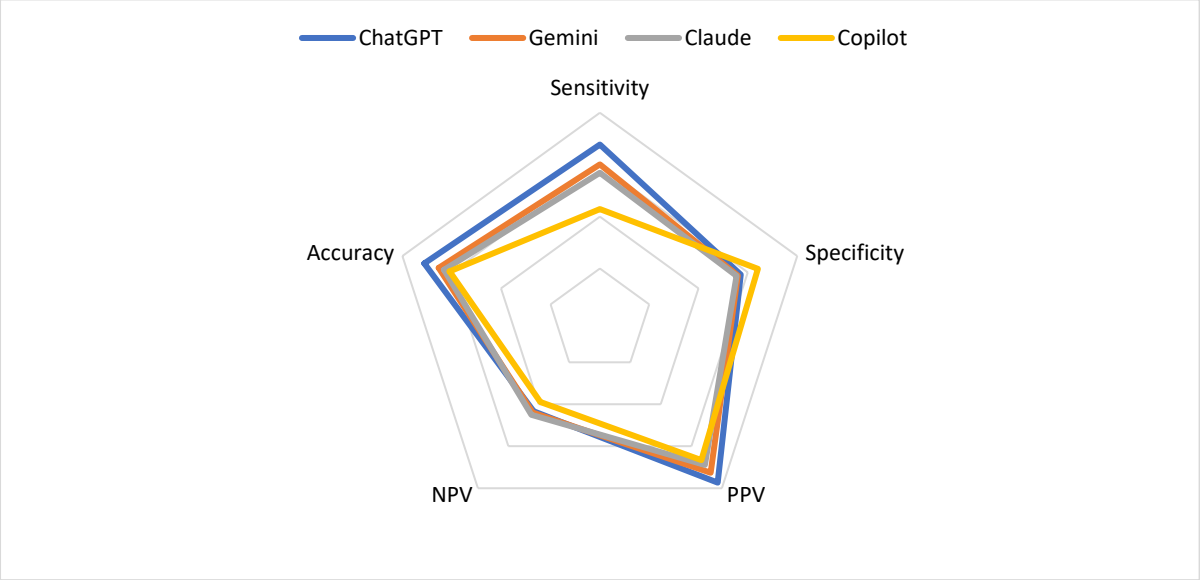


Figure 11. Comparative diagnostic performance of AI models based on classification metrics and ROC curve. Each panel compares the discriminatory performance of ChatGPT 5.2 Pro, Gemini 3 Fast, Claude 4.5 Sonnet, and Microsoft Copilot

From the radar chart, the performance of the model with the best balance for sensitivity, specificity, positive predictive value, negative predictive value, and accuracy is the ChatGPT model. The model with the best competing sensitivity and positive and negative predictive values with somewhat reduced balance with regards to the indices used is the Gemini model. Finally, the model with the least overall performance regarding sensitivity and accuracy with somewhat reduced specificity is Claude and Copilot.

Thus, multimodal AI models demonstrate domain-specific performance with high discrimination for visually salient alignment features and low discrimination and agreement values for parameters that need accurate spatial interpretation, thereby emphasizing domain-specific limitations of general-purpose AI in the analysis of dental images.

IV. Discussion

This study offers an exploratory analysis regarding the clinical readiness of the multiview intraoral photography collection and the diagnostic ability of multimodal large language models in assessing orthodontic malocclusion. This dataset we have describes a high level of clinical variability in terms of alignment, sagittal, vertical, transverse, and arch morphology, aspects that are crucial for effectively analyzing artificial intelligence (AI) tools [1]. Significantly, the data sample consisted of both male and female subjects of different ages, supplemented with multi-angle images, allowing for clear investigation of both simple and complex orthodontic details [6].

Looking across these domains, a clear trend is present. Those parameters associated with anterior alignment, such as crowding and diastema, showed greater agreement (κ 0.00 - 0.63) and accuracy (0.56 - 0.85). The relatively higher performance observed in alignment parameters may be attributed to their visually salient and inherently two-dimensional characteristics, which can be directly inferred from intraoral photographs without requiring complex spatial interpretation. Features such as crowding and diastema manifest as explicit variations in tooth spacing and positioning, making them well-suited for pattern recognition.

Furthermore, alignment assessment relies on relatively stable and easily identifiable anatomical landmarks, reducing variability across evaluators and analytical approaches. These features are also less susceptible to perspective distortion caused by variations in camera positioning, in contrast to sagittal relationships and arch morphology, which depend on more complex inter-arch spatial relationships and three-dimensional interpretation. In contrast, recent studies have found that geometrically precise parameters, such as molar and canine relationships, as well as dental arch morphology, show suboptimal performance, often reflected by lower AUC values, due to their reliance on fine-grained spatial relationships and three-dimensional interpretation that are difficult to infer from two-dimensional images [24]. The

implications of these findings support the hypothesis that general multi-modal learning models and LLMs can be useful in preliminary orthodontic analysis, but there are quite striking variations in some areas of orthodontic expertise [1]. This is supported by a recent assessment of LLMs using large-scale benchmarks, which shows that high performance in broad benchmarks does not necessarily translate to effective reasoning capabilities in scientific and clinical domains [25].

Our results suggest that the evaluative performance of large language models is stronger for anterior than for posterior relationships. These results are consistent with previous studies that have examined AI capabilities in orthodontics. For example, Hack et al. showed that AI models are superior at locating conspicuously visible attributes like front tooth position but are challenged by complex occlusal relationships that demand strong spatial references [4]. Similarly, Stetzel et al. found that deep learning models are successful at predicting aesthetic parts of the Index of Orthodontic Treatment Need (IOTN), which illustrate AI capabilities in visually oriented analyses [15]. Notwithstanding, this outcome conflicts with that demonstrated in this study, particularly concerning sagittal and transverse parameter assessments, which are consistent with previous assessments that argued AI models are challenged when assessing parameters that depend on strong spatial references [24].

This study also demonstrated variations in agreement and accuracy across different large language models (Tables 4 and 5). These inter-model differences may be attributed to variations in underlying architectures and training paradigms. For example, Transformer-based LLMs, such as Vision Transformers (ViT), take a holistic approach to images, efficiently uncovering general irregularities in dental alignment, although they are challenged by making subtle distinctions in densely dependent orthodontic categories such as overjet and molar intercuspation [26]. Once again, this is consistent with previous reviews showing that CNNs trained using architecture-specific data excel at tasks requiring anatomical specificity

compared to general AI systems such as LLMs [7]. The observed differences in outputs among the large language models appear to be attributable to variations in visual abstraction, depth of reasoning, and internal representational capacities, rather than true conceptual understanding [27]. The aggregated results are thus likely to represent the representational constraints presented by their non-specialized training data [28]. Notably, cross-model analysis in scientific artificial intelligence research has revealed high error correlations among leading large language models (LLMs), indicating shared inductive biases rather than independent failure modes [25].

In terms of healthcare informatics, the assessment of several LLMs together can be seen to represent the "hive mind." While each one has different abilities and capacities, certain broad trends become apparent: strong abilities to detect anomalies in alignment and very limited abilities to detect complex spatial characteristics [1,29]. These findings point to the need to learn from all models together and also to recognize the limitations of AI and the importance of human input [28]. The observed convergence across models further suggests that aggregating general-purpose LLMs may offer limited benefit for inherently spatial clinical tasks, reinforcing the need for task-specific AI systems [25]. Future studies should concentrate on building task-focused orthodontic AI systems rather than on testing general-purpose LLMs. The next important area would be to develop appropriate deep learning models based on standardized intraoral photographs with multiple views, which would be appropriately labeled by specialists in the orthodontic field [24]. This would make available appropriate data for building a supervised learning model that would employ spatial reasoning at a finer scale with clinically interpretable outputs [15] that require joint effort on the part of orthodontists and IT personnel.

This study has several limitations, including variability in image quality, the use of a single orthodontist as the reference standard, and a relatively small sample size, which may limit

generalizability. The reliance on a single evaluator may introduce subjectivity, and thus the observed Cohen's kappa values may reflect both AI performance and variability in expert judgment. Future studies should include multiple orthodontists and assess inter-rater reliability to establish a more robust reference standard. In addition, some models generated unsolicited treatment suggestions rather than strictly diagnostic outputs, highlighting limited output controllability. Furthermore, none of the models explicitly expressed diagnostic uncertainty or provided clinical disclaimers, raising important ethical considerations for the use of multimodal AI in healthcare settings. Moreover, all models were evaluated using a single standardized prompting framework. Therefore, the reported results reflect model performance under this specific prompting configuration only. Alternative prompting strategies were not evaluated, and different prompt designs may lead to variations in performance, however ensemble prompt may be used as an alternative. This limitation should be considered when interpreting the findings.

As for conclusion, current multimodal AI models demonstrate limited and parameter-dependent accuracy in detecting orthodontic malocclusions using intraoral photographs. These results emphasize the limitations of general-purpose AI systems for orthodontic decision support and highlight the necessity of task-specific models trained on clinically annotated datasets.

Funding

This research received no external funding

Conflict of interest

The authors declare no conflict of interest

Reference

1. Monill-González A, Rovira-Calatayud L, d'Oliveira NG, Ustrell Torrent JM. Artificial Intelligence in Orthodontics: Where Are We Now? A Scoping Review. Orthod Craniofac Res. 2021. doi:10.1111/ocr.12517
2. Albalawi F, Alamoud K. Trends and Application of Artificial Intelligence Technology in Orthodontic Diagnosis and Treatment Planning—A Review. Applied Sciences. 2022. doi:10.3390/app122211864
3. Subramanian AK, Chen Y, Almalki A, Sivamurthy G, Kafle D. Cephalometric Analysis in Orthodontics Using Artificial Intelligence - A Comprehensive Review. Biomed Res Int. 2022. doi:10.1155/2022/1880113
4. Hack M, Drăgulin B, Hack L, ElSaafin M, Dumitrescu I, Stan D, Păcurar M. Comparative study on the results of orthodontic diagnostics by using algorithms generated by Artificial Intelligence and simple algorithms. Med Pharm Rep. 2024;97(2):215–21. doi:10.15386/mpr-2702
5. Liu J, Zhang C, Shan Z. Application of Artificial Intelligence in Orthodontics: Current State and Future Perspectives. Healthcare (Switzerland). Multidisciplinary Digital Publishing Institute (MDPI); 2023. doi:10.3390/healthcare11202760
6. Ryu J, Kim YH, Kim T, Jung S. Evaluation of Artificial Intelligence Model for Crowding Categorization and Extraction Diagnosis Using Intraoral Photographs. Sci Rep. 2023. doi:10.1038/s41598-023-32514-7
7. Zhang R, Zhang L, Zhang D, Wang Y, Huang YS, Wang D, Li X. Development and Evaluation of a Deep Learning Model for Occlusion Classification in Intraoral Photographs. PeerJ. 2025. doi:10.7717/peerj.20140
8. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the Feasibility of ChatGPT in Healthcare: An Analysis of Multiple Clinical and Research Scenarios. J Med Syst. 2023. doi:10.1007/s10916-023-01925-4
9. Zheng J, Ding X, Pu JJ, Chung SM, H. Ai QY, Hung KF, Shan Z. Unlocking the Potentials of Large Language Models in Orthodontics: A Scoping Review. Bioengineering. 2024. doi:10.3390/bioengineering11111145

10. Wu X, Huang Y, He Q. A large language model improves clinicians' diagnostic performance in complex critical illness cases. Crit Care. 2025 Dec 1;29(1). doi:10.1186/s13054-025-05468-7 PubMed PMID: 40481529.

11. Hatia A, Doldo T, Parrini S, Chisci E, Cipriani L, Montagna L, Lagana G, Guenza G, Agosta E, Vinjolli F, Hoxha M, D'Amelio C, Favaretto N, Chisci G. Accuracy and Completeness of ChatGPT-Generated Information on Interceptive Orthodontics: A Multicenter Collaborative Study. J Clin Med. 2024 Feb 1;13(3). doi:10.3390/jcm13030735

12. Giannakopoulos K, Kavadella A, Salim AA, Stamatopoulos V, Kaklamanos EG. Evaluation of the Performance of Generative AI Large Language Models ChatGPT, Google Bard, and Microsoft Bing Chat in Supporting Evidence-Based Dentistry: Comparative Mixed Methods Study. J Med Internet Res. 2023. doi:10.2196/51580

13. Jasim ES, Kadhom ZM, Al-Groosh D. Comparative evaluation of orthodontic treatment needs index and the dental aesthetic index to assess the need for orthodontic treatment from the participants' perspective: A cross-sectional study. J Orthod Sci. 2025 Sep 1;14(1). doi:10.4103/jos.jos_24_25

14. Roy P, Roy P, Koley S, Sheet S. Bolton's ratio variations in Angle's Class I, Class II and Class III malocclusions: An observational study. J Clin Exp Dent. 2025;17(3):e280–5. doi:10.4317/jced.62591

15. Stetzel L, Foucher F, Jang SJ, Wu T, Fields HW, Schumacher FL, Richmond S, Ko C. Artificial Intelligence for Predicting the Aesthetic Component of the Index of Orthodontic Treatment Need. Bioengineering. 2024. doi:10.3390/bioengineering11090861

16. Peter H. Brook, William C. Shaw. The development of an index of orthodontic treatment priority. Eur J Orthod. 1989 Aug 1;11(3):309–20. doi:10.1093/oxfordjournals.ejo.a035999

17. Nguyen MS, Nguyen MK, Saag M, Jagomägi T. The Need for Orthodontic Treatment Among Vietnamese School Children and Young Adults. Int J Dent. 2014. doi:10.1155/2014/132301

18. Kozanecka A, Sarul M, Kawala B, Antoszewska-Smith J. Objectification of Orthodontic Treatment Needs: Does the Classification of Malocclusions or a History of Orthodontic Treatment Matter? Advances in Clinical and Experimental Medicine. 2016. doi:10.17219/acem/62828

19. Stanley Braun, William P. Hnat, Dana E. Fender, Harry L. Legan. The form of the human dental arch. Angle Orthod. 1998 Feb;68:29–36.

20. Huynh N, Zhang J, Pliska BT, Amin R, Narang I, Chadha NK, Cholette M, Kirk V, Montpetit A, Vézina K, Jacob S V, Laberge S, Hamoda MM, Almeida FR. Prevalence of Altered Craniofacial Morphology in Children With OSA. J Sleep Res. 2025. doi:10.1111/jsr.70060

21. Shah NH, Entwistle DA, Pfeffer MA. Creation and Adoption of Large Language Models in Medicine. JAMA. 2023. doi:10.1001/jama.2023.14217

22. Masood Y, Masood M, Binti Zainul NN, Abdul Araby NB, Hussain SF, Newton T. Impact of Malocclusion on Oral Health Related Quality of Life in Young People. Health Qual Life Outcomes. 2013. doi:10.1186/1477-7525-11-25

23. Nahm FS. Receiver operating characteristic curve: overview and practical use for clinicians. Korean J Anesthesiol. 2022 Feb 1;75(1):25–36. doi:10.4097/kja.21209 PubMed PMID: 35124947.

24. Vishwanathaiah S, Fageeh HN, Khanagar SB, Maganur PC. Artificial Intelligence Its Uses and Application in Pediatric Dentistry: A Review. Biomedicines. 2023. doi:10.3390/biomedicines11030788

25. Song Z, Lu J, Du Y, Yu B, Pruyn TM, Huang Y, Guo K, Luo X, Qu Y, Qu Y, Wang Y, Wang H, Guo J, Gan J, Shojaee P, Luo D, Bran AM, Li G, Zhao Q, Luo SXL, Zhang Y, Zou X, Zhao W, Zhang YF, Zhang W, Zheng S, Zhang S, Khan ST, Rajabi-Kochi M, Paradi-Maropakis S, Baltoiu T, Xie F, Chen T, Huang K, Luo W, Fang M, Yang X, Cheng L, He J, Hassoun S, Zhang X, Wang W, Reddy CK, Zhang C, Zheng Z, Wang M, Cong L, Gomes CP, Hsieh CY, Nandy A, Schwaller P, Kulik HJ, Jia H, Sun H, Moosavi SM, Duan C. Evaluating Large Language Models in Scientific Discovery [Internet]. 2025 Dec 17. Available from: http://arxiv.org/abs/2512.15567

26. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. 2020. doi:10.48550/arxiv.2010.11929

27. Azizi S. Applications of Artificial Intelligence in Diagnosis and Treatment Planning of Orthodontics: A Narrative Review. Saudi Dent J. 2025. doi:10.1007/s44445-025-00077-0

28. Ahmed AE, Aljohani WF, Abu Rukbah LK, Rajhi SA, Najmi NK, Zughlul MK, Alshammari AM, Alotaibi S, Almarhabi TH, Alamri M, Rebh SB. The Role of Artificial Intelligence in Stroke Imaging in Emergency Settings: A Systematic Review. Cureus. 2025. doi:10.7759/cureus.93941

29. Jiang L, Chai Y, Li M, Liu M, Fok R, Dziri N, Tsvetkov Y, Sap M, Albalak A, Choi Y. Artificial Hivemind: The Open-Ended Homogeneity of Language Models (and Beyond) [Internet]. 2025 Oct 27. Available from: http://arxiv.org/abs/2510.22954

6. Bukti konfirmasi artikel accepted

21 April 2026

UNIVERSITAS TRISAKTI

Joko Kusnoto <joko.k@trisakti.ac.id>

Accepted Manuscript

HIR <hir@kosmi.org> Tue, Apr 21, 2026 at 1:15 PM
Reply-To: HIR <hir@kosmi.org>
To: Joko Kusnoto <joko.k@trisakti.ac.id>

Dear Author,

Thank you for your message.
Your manuscript is currently undergoing proofreading.
We will proceed with the proofreading of this version, and once that is completed, we will send it to you for your revie
w later this week.

Please feel free to contact us if you need any further assistance.

Kind regards,
Editorial Office
Healthcare Informatics Research

==========

Healthcare Informatics Research
The Korean Society of Medical Informatics
Tel : +82-2-733-7637
E-mail : hir@kosmi.org

==========

7. Bukti permintaan proofreading

24 April 2026

UNIVERSITAS TRISAKTI

Joko Kusnoto <joko.k@trisakti.ac.id>

[HIR-26-031] Author Proof Review Request

4 messages

HIR <hir@kosmi.org> Fri, Apr 24, 2026 at 7:46 AM
Reply-To: HIR <hir@kosmi.org>
To: joko.k@trisakti.ac.id

대용량 첨부파일 1개 - 다운로드 기한 : 2026-05-08 / 최대 500회 가능

HIR-26-031_sup.pdf (35.99 MB)

Dear Dr. Kusnoto:

We are attaching the edited manuscript file for your review.
Please examine it carefully, especially **the editor's notes,** and let us know which parts you would like to revise.

Please note that this proof stage is not intended for extensive corrections, additions, or deletions.
It is recommended that editing be limited to correcting typographical errors, incorrect dates, grammatical errors, and updating information about references which were in press.

We have also done English proofreading of the main text. Kindly check for any unintended changes in meaning.
For your reference, we are attaching the file 'HIR-26-031_EngProof.docx' — you do **not need to return** this file

We would appreciate receiving your response **within 24 hours (by April 25)**.
★ Please download and open the PDF file 'HIR-26-031(+s).pdf', 'HIR-26-031_sup.pdf' and respond to th e editor's comments in the annotations.
★ Please do not edit the file directly.

If you have any questions, please don't hesitate to contact us.
Thank you very much.

The Editorial Office
Healthcare Informatics Research

==========

Healthcare Informatics Research
The Korean Society of Medical Informatics
Tel : +82-2-733-7637
E-mail : hir@kosmi.org
==========

2 attachments

HIR-26-031(+s).pdf
905K

HIR-26-031_EngProof.docx
2573K

Original Article

Healthc Inform Res. 2026 April;32(2):1-13.
https://doi.org/10.4258/hir.2026.32.2.1
pISSN 2093-3681 • eISSN 2093-369X

HIR
Healthcare Informatics Research

Accuracy of Orthodontic Malocclusion Detection Using Multiple AI Models: A Comparative Study

Hillda Herawati[1], Joko Kusnoto[1], Indrayadi Gunardi[2], Anggit Wirasto[3], Tri Erri Astoeti[4]

[1]Faculty of Dentistry, Universitas Trisakti, Jakarta, Indonesia
[2]Department of Oral Medicine, Faculty of Dentistry, Universitas Trisakti, Jakarta, Indonesia
[3]Informatics Study Program, Faculty of Science and Technology, Universitas Harapan Bangsa, Purwokerto, Indonesia
[4]Department of Dental Public Health and Preventive Dentistry, Faculty of Dentistry, Universitas Trisakti, Jakarta, Indonesia

Objectives: This study aimed to evaluate and compare the accuracy of multiple artificial intelligence (AI) models (ChatGPT 5.2 Pro, Gemini 3 Fast, Claude 4.5 Sonnet, and Microsoft Copilot) in detecting orthodontic malocclusion features in standardized multiview intraoral photographs. The reference standard was assessment by an orthodontist. **Methods:** A cross-sectional observational study was conducted using five standardized intraoral photographs (frontal, right lateral, left lateral, maxillary occlusal, and mandibular occlusal) obtained from 50 children aged 9–12 years. The following eight malocclusion parameters were assessed: anterior crowding, diastema, overjet, overbite, molar relationship, canine relationship, crossbite, and dental arch symmetry. Diagnostic accuracy and agreement between each AI model and the orthodontist were evaluated using Cohen's kappa (κ) and the area under the receiver operating characteristic curve (AUC). **Results:** Agreement between the AI models and the orthodontist ranged from poor to moderate across all orthodontic domains, with Cohen's κ values ranging from -0.15 to 0.63. Visually prominent alignment features, including anterior crowding and diastema, demonstrated comparatively higher agreement (κ, 0.00–0.63) and discriminatory performance, with AUC values ranging from 0.56 to 0.85. In contrast, parameters requiring precise spatial interpretation, such as sagittal relationships, overbite, crossbite, and arch morphology, showed consistently low agreement (κ, -0.15 to 0.38) and poor to near-random classification performance, with AUC values predominantly ranging from 0.41 to 0.70 and, in some cases, approaching 0.50. **Conclusions:** Current multimodal AI models demonstrate limited, parameter-dependent accuracy in detecting orthodontic malocclusions from intraoral photographs. These findings emphasize the limitations of general-purpose AI systems for orthodontic decision support and highlight the need for task-specific models trained on clinically annotated datasets.

Keywords: Accuracy, Artificial Intelligence, Detection, Malocclusion, Orthodontic

Submitted:
Revised:
Accepted:

Corresponding Author
Joko Kusnoto
Department of Orthodontics, Faculty of Dentistry, Universitas Trisakti, Jl. Kyai Tapa no. 260, Jakarta 11440, Indonesia. Tel: +62-21-5672731, E-mail: joko.k@trisakti.ac.id (https://orcid.org/0009-0009-8955-4459)

I. Introduction

Malocclusion is a prevalent oral health problem among children and adolescents, with reported prevalence rates as high as 30%-40% in some populations [1].Early detection is essential to prevent functional impairment and reduce the need for more complex orthodontic interventions later in [1,2]. Traditional orthodontic screening methods rely largely on direct clinical examinations by trained specialists, which may limit access to needed dental care, particularly in community settings such as schools [2].

Recent advances in artificial intelligence (AI) have shown promise for improving diagnostic accuracy in orthodontics, including through the application of machine learning and deep learning to various imaging modalities. For example, convolutional neural networks (CNNs) have been used effectively in cephalometric analysis and panoramic image interpretation, yielding accurate automated orthodontic diagnoses [3]. However, the use of intraoral photographs remains relatively underexplored, despite their value as noninvasive, low-cost clinical records suitable for early screening and teleorthodontics [4,5]. The diagnostic potential of photographs has nevertheless been recognized, and previous studies have reported strong diagnostic performance for dental conditions assessed from these images [6,7].

In recent years, large language models (LLMs), including ChatGPT and similar AI frameworks, have emerged with the ability to interpret both textual and visual inputs. Studies evaluating LLM responses in orthodontics have reported moderate to high consistency but substantial variability in accuracy relative to assessments by orthodonticspecialists [8]. Although these models can generate fluent and well-structured responses, they may also provide incomplete or misleading information, underscoring the need for caution-clinical interpretation [9-11]. These observations support the view that LLMs may serve as adjunctive tools in orthodontic practice rather than replacements for expert judgment.

Comparative studies have identified discrepancies between LLM-generated responses and assessments by orthodontic experts, particularly in malocclusion classification and treatment decision-making [9,12]. Although these models may have value for preliminary educational purposes, existing evidence suggests that they cannot replace the nuanced assessments of experienced clinicians [1,3,11]. This limitation highlights the need for further research to rigorously validate AI-generated outputs against established clinical standards.

To evaluate the accuracy of malocclusion classification from clinical images by AI models, including advanced versions such as ChatGPT 5.2 Pro, Gemini 3 Fast, Claude 4.5 Sonnet, and Microsoft Copilot, standardized inputs are required for comparison with expert evaluations. At present, research on the performance of multimodal LLMs in analyzing standardized multiview intraoral photographs against comprehensive orthodontic diagnostic criteria remains limited [7,9]. Further studies are therefore needed in thisarea.

II. Methods

1. Study Design and Population

This cross-sectional observational study was designed to evaluate the diagnostic performance of multimodal LLMs in detecting orthodontic malocclusion features from standardized intraoral photographs. The study focused exclusively on image-based diagnostic interpretation and did not incorporate clinical examination, radiographic analysis, or treatment planning. The study used retrospective clinical photographs obtained from 50 children aged 9–12 years in Cimahi, West Java, Indonesia. No clinical intervention, treatment modification, or follow-up assessment was performed as part of the study.

2. Sample Size

This observational cross-sectional study aimed to evaluate the diagnostic accuracy of multimodal LLMs in detecting orthodontic malocclusion from intraoral images. The study sample consisted of 50 pediatric participants, each contributing five intraoral images obtained from different views, including frontal, right lateral, left lateral, maxillary occlusal, and mandibular occlusal views. The total sample size was 250 images, which was considered sufficient to cover key aspects of orthodontic malocclusion, including alignment, sagittal, vertical, transverse, and arch morphology.

3. Ethical Approval and Consent to Participate

Written informed consent was obtained from all participants and their legal guardians before study initiation. Ethical approval for the study protocol was granted by the Health Research Ethics Committee of the Faculty of Dentistry, Universitas Trisakti (No. 1006/S3/KEPK/FKG/9/2025).

4. Image Collection

Five images were acquired for each participant, comprising frontal, left lateral, right lateral, maxillary occlusal, and mandibular occlusal views (Figure 1). All images were captured using an iPhone 15 (Apple Inc., Cupertino, CA, USA) at a resolution of 6048 × 4032 pixels (24 megapixels). All images were obtained by a single investigator who was calibrated for the study, following standardized intraoral photography protocols to ensure consistent angulation, lighting, and field of view.

5. Image Preparation and Pre-processing

All images were anonymized before analysis by removing patient-identifying information. The images were standardized

https://doi.org/10.4258/hir.2026.32.2.1

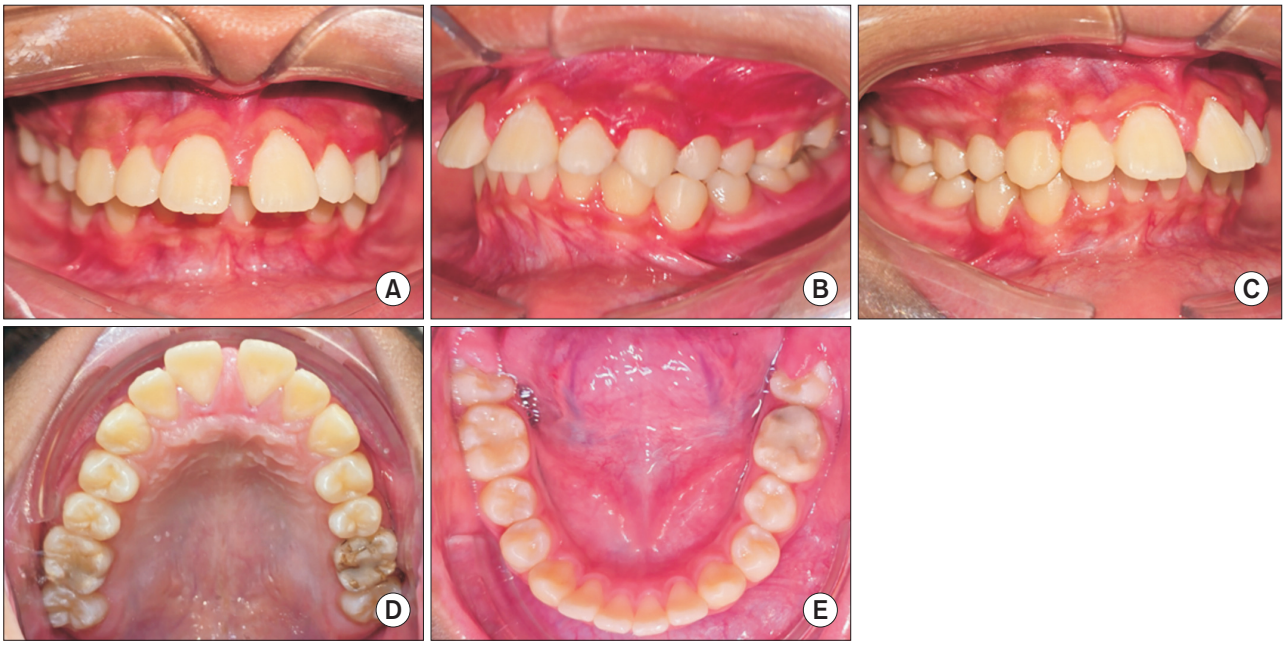


Figure 1. Sample clinical images obtained from orthodontic patients: (A) frontal view, (B) left lateral view, (C) right lateral view, (D) maxillary occlusal view, and (E) mandibular occlusal view.

without digital enhancement, filtering, contrast adjustment, or color correction. However, the images were cropped to remove extraneous objects, such as hands and mirrors, that were visible during image acquisition. Images were analyzed at their original resolution to preserve clinically relevant visual features, such as tooth arrangement, occlusal relationships, and arch morphology.

6. Orthodontic Diagnostic Framework

Orthodontic assessment in this study followed a structured diagnostic framework based on established clinical standards, including Proffit's definitionsof malocclusion, Angle's classification of sagittal relationships, and the Dental Health Component of the Index of Orthodontic Treatment Need (IOTN) [13-15].

For analytical consistency and effective image-based evaluation, orthodontic parameters were organized into five diagnostic domains reflecting different levels of spatial complexity:

1) Alignment domain: This domain included assessment of anterior crowding and diastema in both the maxillary and mandibular arches. These features are commonly recognized indicators in orthodontic screening [16].
2) Sagittal relationship domain: This domain comprised assessment of overjet, molar relationship, canine relationship, and overall Angle classification, focusing on anteroposterior dental relationships that are central to orthodontic diagnosis [17].
3) Vertical relationship domain: This domain emphasized overbite assessment, categorized as normal, deep bite, or open bite, representing clinically important vertical discrepancies [17].
4) Transverse relationship domain: This domain assessed the presence and type of crossbite (anterior or posterior, unilateral or bilateral), capturing transverse discrepancies that are often subtle on two-dimensional imaging [18].
5) Arch morphology domain: This domain evaluated dental arch symmetry using occlusal views, highlighting structural characteristics shaped by underlying skeletal patterns [19].

All parameters were evaluated systematically within this unified framework by both the orthodontist and the AI models, thereby ensuring a structured assessment approach. For statistical analysis, outcomes were dichotomized as normal or abnormal to enable direct comparison of diagnostic agreement and discrimination [20]. The detailed mapping of multilevel orthodontic parameters to binary categories is presented in Table 1.

7. Generative AI Multimodal Large Language Models

In this study, four publicly available multimodal LLMs with image-text processing capabilities were evaluated: ChatGPT 5.2 Pro (OpenAI), Claude 4.5 Sonnet (Anthropic), Gemini 3 Fast (Google), and Microsoft Copilot (Microsoft[12,21]. These models were accessed through their official web-based

Table 1. Mapping of multi-class orthodontic parameters into binary categories for statistical analysis

Domain	Parameter	Original category	Binary category
Alignment	Maxillary crowding	None	Normal
		Mild, Moderate, Severe	Abnormal
	Mandibular crowding	None	Normal
		Mild, Moderate, Severe	Abnormal
	Diastema	None	Normal
		Presence (maxilla or mandible)	Abnormal
Sagittal	Overjet	Normal	Normal
		Increased, Edge-to-edge, Reverse	Abnormal
	Molar relationship (right & left)	Class I	Normal
		Class II, Class III	Abnormal
	Canine relationship (right & left)	Class I	Normal
		Class II, Class III	Abnormal
	Angle classification	Class I	Normal
		Class II, Class III	Abnormal
Vertical	Overbite	Normal	Normal
		Deep bite, Open bite	Abnormal
Transverse	Crossbite (right & left)	None	Normal
		Anterior, Posterior, Anterior + Posterior	Abnormal
Arch morphology	Arch symmetry	Symmetrical (maxilla/mandible)	Normal
		Asymmetrical	Abnormal

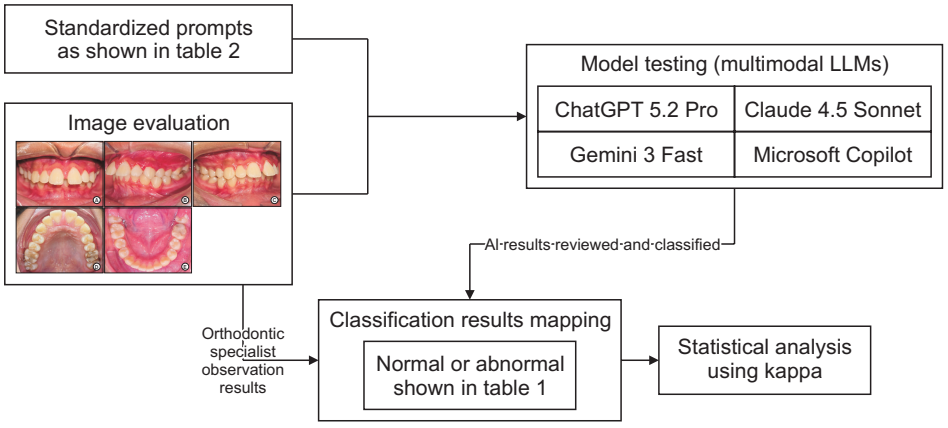


Figure 2. Schematic overview of multimodal large language model (LLM) evaluation, result mapping, and agreement analysis.

interfaces using their default system configurations, without application programming interfaces, external tools, or model fine-tuning [12]. This approach was intended to reflect how these multimodal LLMs would operate in real-world clinical use.

8. Model Testing Workflow and Prompting Strategy

For model evaluation, a structured workflow was used, as illustrated in Figure 2 and detailed in Table 2. Analysis and testing of the AI models were conducted in stages according to the orthodontic diagnostic framework. Each intraoral image was evaluated independently by each multimodal model in a blinded testing procedure. This approach ensured that each model's rating depended solely on information contained in the corresponding intraoral image. Parameter-specific prompts were used for each orthodontic category with identical syntax and structure, with the aim of approximating the logical reasoning process used by an orthodontist.

https://doi.org/10.4258/hir.2026.32.2.1

Table 2. Standardized prompts used for multimodal LLM evaluation by orthodontic domain

Orthodontic domain	Image view	Diagnostic focus	Standardized prompt
Alignment	Frontal intraoral	Anterior crowding	"Analyze this frontal intraoral photograph. Focus on anterior crowding. Identify whether crowding is present in the maxillary and/or mandibular anterior teeth, specify the affected teeth, classify severity as mild, moderate, or severe, and provide a diagnostic conclusion."
Alignment	Frontal intraoral	Diastema	"Analyze this frontal intraoral photograph. Focus on diastema. Determine whether diastema is present, identify its location (maxillary and/or mandibular), and provide a diagnostic conclusion."
Sagittal relationship	Lateral intraoral (right/left)	Molar relationship (M1)	"Analyze this lateral intraoral photograph. Focus on the relationship between the maxillary first molar and mandibular first molar. Classify the molar relationship as Class I, II, or III. Report findings separately for the right and left sides."
Sagittal relationship	Lateral intraoral (right/left)	Canine relationship (C)	"Analyze this lateral intraoral photograph. Focus on the relationship between the maxillary canine and mandibular canine. Classify the canine relationship as Class I, II, or III. Report findings separately for the right and left sides."
Sagittal relationship	Lateral intraoral	Overjet	"Analyze this lateral intraoral photograph. Focus on overjet. Classify the overjet as normal, increased, edge-to-edge, or reverse, and provide a diagnostic conclusion."
Vertical relationship	Frontal intraoral	Overbite	"Analyze this frontal intraoral photograph. Focus on overbite. Classify the vertical overlap as normal, deep bite, or open bite, and provide a diagnostic conclusion."
Transverse relationship	Frontal intraoral	Crossbite	"Analyze this frontal intraoral photograph. Focus on transverse relationships. Identify the presence of crossbite, specify whether it is unilateral or bilateral, and provide a diagnostic conclusion."
Arch morphology	Maxillary occlusal	Arch symmetry	"Analyze this maxillary occlusal photograph. Assess the arch is symmetrical or asymmetrical, and provide a diagnostic conclusion."
Arch morphology	Mandibular occlusal	Arch symmetry	"Analyze this mandibular occlusal photograph. Assess the arch is symmetrical or asymmetrical, and provide a diagnostic conclusion."

LLM: large language model.

All prompts were written in Indonesian and were kept identical throughout the study. Outputs from each model were classified into predefined orthodontic categories and then reduced to two groups: normal and abnormal.

9. Statistical Analysis

Cohen's kappa (κ) was used to assess agreement between each AI model and the single orthodontist's assessment. This method is widely used in dental research to quantify inter-rater reliability, particularly in studies examining diagnostic discrepancies[7]. By providing a comparative framework, Cohen's κ helps indicate how closely AI model outputs align with expert clinical assessments [22]. Model performance was evaluated by dichotomizing diagnostic outputs into normal and abnormal categories and calculating the area under the receiver operating characteristic curve (AUC) to quantify discriminative ability. AUC values were interpreted according to established receiver operating characteristic criteria as follows: >0.90, excellent; 0.80–0.89, good; 0.70–0.79, fair; 0.60–0.69, poor; and <0.60, fail [23].

III. Results

A total of 50 children were included in this study. Participants were 9–12 years old (Table 3). For each participant, five standardized intraoral photographs were obtained, com-

Table 3. Demographic characteristics of the subjects

Variable	Value
Sex	
Male	20 (40)
Female	30 (60)
Age (yr)	10.56 ± 0.73

Values are presented as number (%) or mean ± standard deviation.

Table 4. Integrated distribution of orthodontic diagnostic characteristics across all domains

Parameter	n (%)
Alignment domain	
Maxillary crowding	
None	6 (12)
Mild	23 (46)
Moderate	13 (26)
Severe	8 (16)
Mandibular crowding	
None	6 (12)
Mild	16 (32)
Moderate	21 (42)
Severe	7 (14)
Diastema	
Maxilla	4 (8)
Mandible	12 (24)
Sagittal relationship domain	
Overjet	
Normal	31 (62)
Increased	15 (30)
Edge-to-edge	2 (4)
Reverse	2 (4)
Molar relationship (right)	
Class I	31 (62)
Class II	18 (36)
Class III	1 (2)
Molar relationship (left)	
Class I	30 (60)
Class II	18 (36)
Class III	2 (4)
Canine relationship (right)	
Class I	28 (56)
Class II	22 (44)

Continued on the next column.

prising frontal, left lateral, right lateral, maxillary occlusal, and mandibular occlusal views. In total, 250 intraoral images were analyzed and used as input for the AI models.

Across all orthodontic domains, the dataset showed substantial clinical variability while representing a typical range of malocclusions (Table 4). In the alignment domain, both arches showed a predominance of anterior crowding, most of which was mild to moderate in severity. Diastemas were observed more frequently in the mandible than in the maxilla. In the sagittal domain, normal overjet and Angle Class I relationships predominated; however, a high proportion of participants also exhibited Class II malocclusion and increased overjet. In the vertical domain, normal overbite was most common, with deep bite representing the primary anomaly and open bite occurring rarely. In the transverse

Table 4. Continued

Parameter	n (%)
Canine relationship (left)	
Class I	30 (60)
Class II	19 (38)
Class III	1 (2)
Angle classification	
Class I	27 (54)
Class II	21 (42)
Class III	2 (4)
Vertical relationship domain	
Overbite	
Normal	31 (62)
Open bite	1 (2)
Deep bite	18 (36)
Transverse relationship domain	
Crossbite (left)	
None	41 (82)
Anterior	6 (12)
Posterior	2 (4)
Anterior + posterior	1 (2)
Crossbite (right)	
None	41 (82)
Anterior	6 (12)
Posterior	3 (6)
Arch morphology domain	
Arch symmetry	
Maxilla (symmetrical)	39 (78)
Mandible (symmetrical)	40 (80)

https://doi.org/10.4258/hir.2026.32.2.1

domain, crossbite was uncommon, and most participants had no crossbite; when present, crossbite was usually unilateral. Symmetry predominated in both the maxillary and mandibular arches. Overall, the dataset represented a clinically variable yet typical sample suitable for comparative evaluation of multimodel AI system performance.

Following the descriptive analysis, receiver operating characteristic (ROC) analysis was performed to evaluate the discriminatory performance of the multimodal models in classifying normal versus abnormal orthodontic findings. Figure 3 shows marked domain-specific variation in model performance across the ROC curves. Parameters related to alignment demonstrated the greatest discriminatory ability, with ROC curves positioned farther from the reference diagonal, whereas sagittal relationship parameters showed only moderate discriminatory ability. In contrast, the verti-

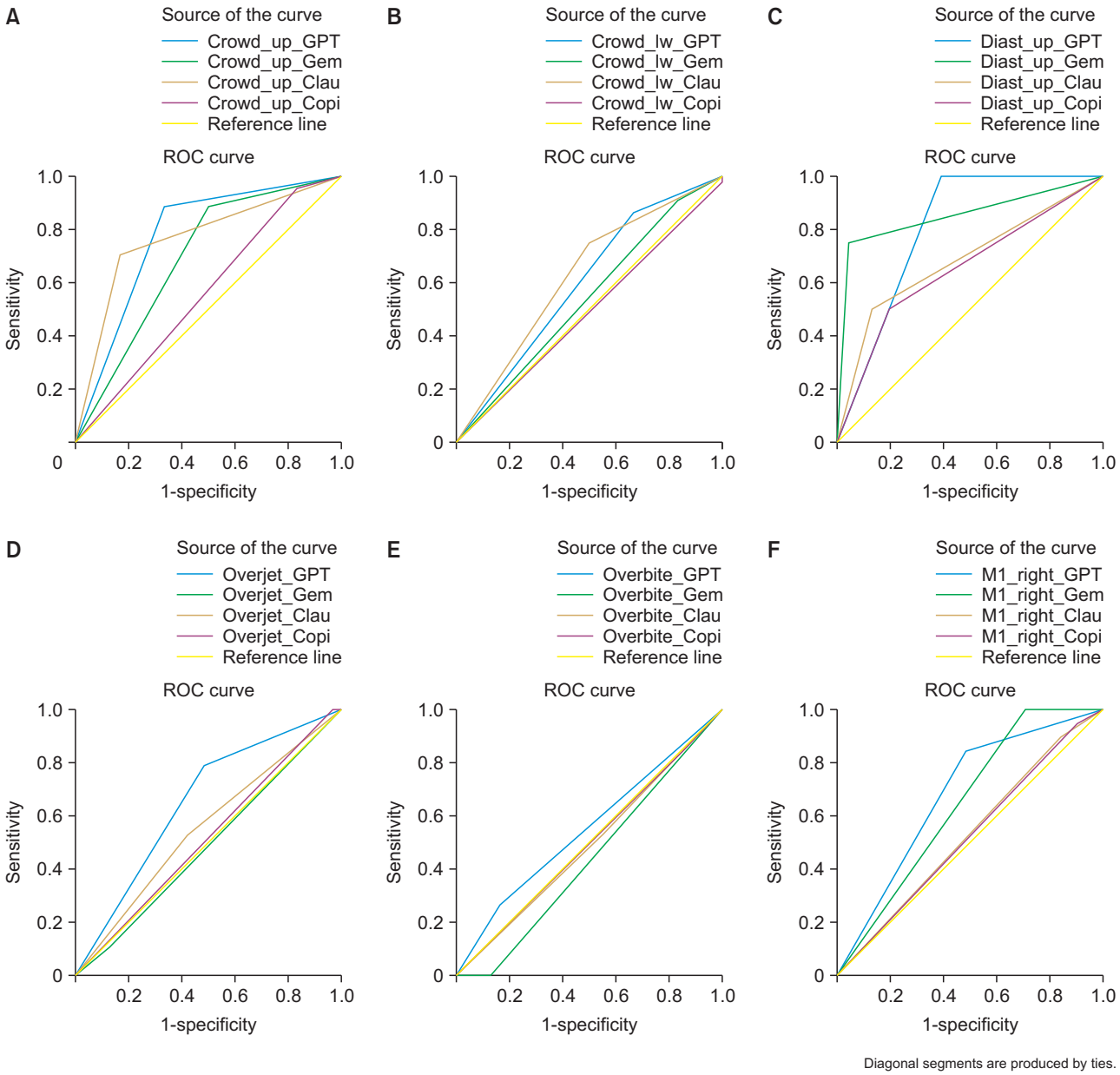
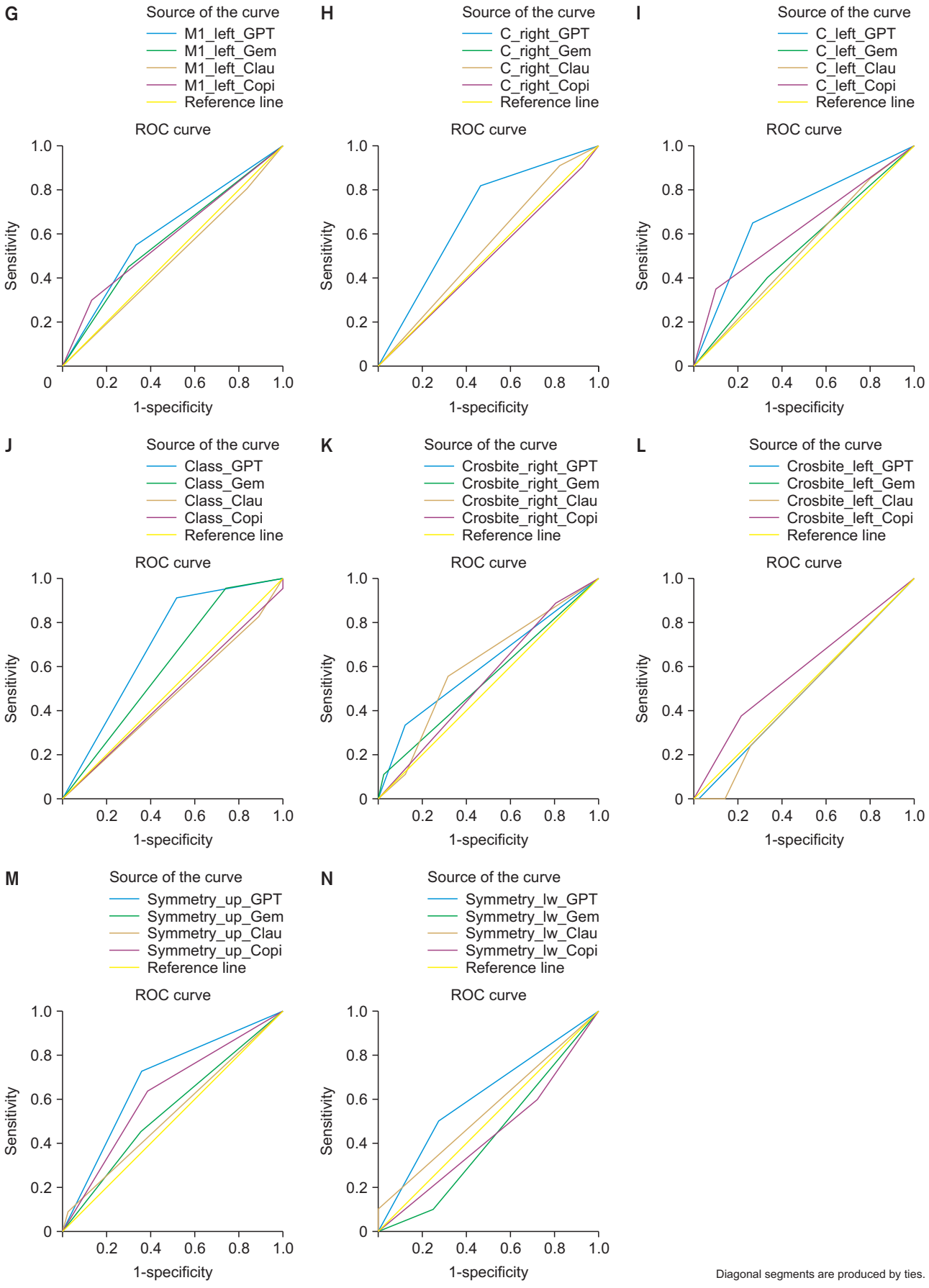


Figure 3. Integrated receiver operating characteristic (ROC) curves of multimodal AI models for orthodontic parameters (see Supplementary Figure S1 for a high-resolution version). Panels (A)–(N) represent ROC curves for individual orthodontic parameters evaluated from standardized multiview intraoral photographs; (A) maxillary crowding, (B) mandibular crowding, (C) maxillary diastema, (D) overjet, (E) overbite, (F) right first molar relationship, (G) left first molar relationship, (H) right canine relationship, (I) left canine relationship, (J) Angle classification, (K) right crossbite, (L) left crossbite, (M) maxillary arch symmetry, and (N) mandibular arch symmetry. Each panel compares the discriminatory performance of ChatGPT 5.2 Pro, Gemini 3 Fast, Claude 4.5 Sonnet, and Microsoft Copilot against the orthodontist reference standard.



Diagonal segments are produced by ties.

Figure 3. Continued

https://doi.org/10.4258/hir.2026.32.2.1

cal, transverse, and arch morphology domains showed near-random classification performance, with their ROC curves clustering close to the reference diagonal. These findings indicate that multimodal models perform better for visually prominent orthodontic features than for complex spatial relationships.

As shown in Table 5, the discriminative performance of the multimodal AI models varied substantially across orthodontic tasks. The alignment domain showed the highest discriminative performance, with Gemini achieving the highest AUC value of 0.85, followed by ChatGPT, with AUC values approaching 0.80, indicating relatively strong discrimination for visually observable features such as crowding and diastema. Claude showed similar but slightly lower values, whereas Copilot demonstrated the lowest discriminative performance.

Discriminative performance declined in the sagittal relationship domain, in which the highest AUC values, up to 0.70, were observed for ChatGPT, representing fair discrimination, whereas the other models performed less well. In contrast, all models showed poor discriminative performance in the vertical and transverse relationship domains, with AUC values close to the random-classification threshold of 0.50. Discriminative performance was lowest in the arch morphology domain, in which AUC values did not reach the threshold for fair discrimination, reflecting the continued limitations of multimodal models in processing complex dental geometric features accurately.

Analysis of Cohen's κ statistics for agreement between the multimodal models and the orthodontist showed poor to moderate agreement across all domains (Table 6). In the alignment domain, agreement was generally poor, although Gemini achieved moderate agreement, with κ values up to 0.63. ChatGPT and Claude showed lower agreement, whereas Copilot showed only minimal agreement. Agreement in the sagittal relationship domain was also poor overall, although ChatGPT performed relatively better, with κ values up to 0.38; Gemini and Claude performed less well, and Copilot showed minimal agreement. In the vertical, transverse, and arch morphology domains, agreement was consistently poor, with κ values near zero or below zero across models. Overall, the combined ROC-AUC and κ findings suggest that, although multimodal AI models can moderately discriminate visually salient alignment features, agreement with orthodontist evaluation remains limited, particularly for parameters requiring precise spatial interpretation.

Figure 4 presents a radar-chart comparison of the discrimination abilities of the four AI models. ROC analysis indicates notable differences in discriminative performance among the models, with ChatGPT showing the highest overall performance, followed by Gemini, whereas Claude and Microsoft Copilot showed more limited performance, as reflected by ROC curves that lay closer to the reference line.

According to the radar chart, ChatGPT showed the best overall balance across sensitivity, specificity, positive predictive value, negative predictive value, and accuracy. Gemini showed comparatively strong sensitivity and positive and negative predictive values, although its overall balance across performance indices was somewhat lower. Claude and Copilot showed the weakest overall performance, particularly

Table 5. Integrated ROC-AUC performance of multimodal AI models across orthodontic domains

Orthodontic domain	Parameters included	AUC				Overall interpretation
		ChatGPT 5.2 Pro	Gemini 3 Fast	Claude 4.5 Sonnet	Microsoft Copilot	
Alignment	Crowding (maxillary, mandibular), Diastema (maxillary)	0.60–0.80	0.69–0.85	0.68–0.77	0.56–0.65	Poor–Good
Sagittal relationships	Overjet, M1 (R/L), Canine (R/L), Angle classification	0.65–0.70	0.53–0.65	0.47–0.55	0.48–0.52	Poor
Vertical relationship	Overbite	0.55	0.44	0.49	0.50	Fail
Transverse relationships	Crossbite (R/L)	0.49–0.61	0.50–0.54	0.48–0.60	0.54–0.58	Fail–Poor
Arch morphology	Symmetry (maxillary, mandibular)	0.41–0.68	0.43–0.61	0.47–0.60	0.44–0.63	Fail–Poor

ROC: receiver operating characteristic, AUC: area under the curve; AI: artificial intelligence, R: right, L: left.
AUC values were interpreted based on established ROC guidelines: ≥0.90 (excellent), 0.80–0.89 (good), 0.70–0.79 (fair), 0.60–0.69 (poor), and <0.60 (fail).

Table 6. Summary of agreement (Cohen's kappa) between AI models and orthodontist assessment by orthodontic domains

Orthodontic domain	Parameters included	Cohen's kappa				Overall agreement
		ChatGPT 5.2 Pro	Gemini 3 Fast	Claude 4.5 Sonnet	Microsoft Copilot	
Alignment	Crowding (maxillary, mandibular), Diastema (maxillary)	0.11–0.33	0.15–0.63	0.16–0.25	0.00–0.17	Poor–Moderate
Sagittal relationships	Overjet, M1 (R/L), Canine (R/L), Angle classification	0.20–0.38	0.08–0.21	-0.03–0.07	-0.04–0.23	Poor–Moderate
Vertical relationship	Overbite	0.13	-0.15	0.05	-0.01	Poor
Transverse relationships	Crossbite (R/L)	-0.02–0.13	0.00–0.05	0.09–0.11	0.03–0.07	Poor
Arch morphology	Symmetry (maxillary, mandibular)	0.07–0.27	-0.15–0.15	0.07–0.15	-0.07–0.19	Poor

AI: artificial intelligence, R: right, L: left.

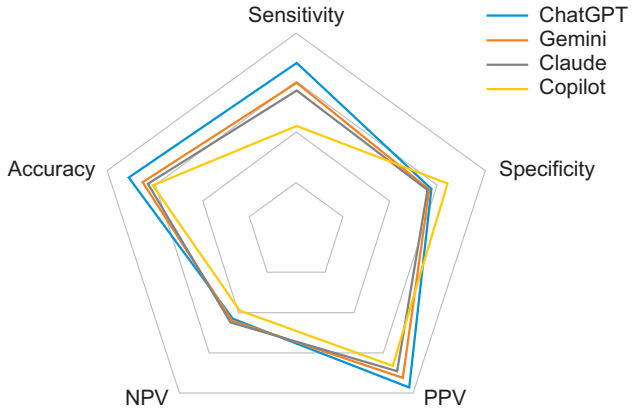


Figure 4. Comparative diagnostic performance of AI models based on classification metrics and receiver operating characteristic (ROC) curve. Each panel compares the discriminatory performance of ChatGPT 5.2 Pro, Gemini 3 Fast, Claude 4.5 Sonnet, and Microsoft Copilot. AI: artificial intelligence, NPV: negative predictive value, PPV: positive predictive value.

with respect to sensitivity and accuracy, with somewhat reduced specificity as well.

Taken together, these findings indicate that multimodal AI models show domain-specific performance, with higher discrimination for visually salient alignment features and lower discrimination and agreement for parameters requiring accurate spatial interpretation. These results emphasize the domain-specific limitations of general-purpose AI systems in the analysis of dental images.

IV. Discussion

This study provides an exploratory analysis of the clinical applicability of a multiview intraoral photograph dataset and the diagnostic performance of multimodal large language models in assessing orthodontic malocclusion. The dataset showed substantial clinical variability across alignment, sagittal, vertical, transverse, and arch morphology domains, all of which are important for evaluating AI tools [1].Notably, the sample included both male and female participants across the target age range and was supplemented with multi-angle images, enabling assessment of both simple and complex orthodontic features [6].

Across these domains, a clear pattern emerged. Parameters related to anterior alignment, such as crowding and diastema, showed higher agreement (κ, 0.00–0.63) and AUC values (0.56–0.85). The relatively stronger performance observed for alignment parameters may be attributable to their visually salient, inherently two-dimensional characteristics, which can be inferred directly from intraoral photographs without requiring complex spatial interpretation. Features such as crowding and diastema appear as explicit variations in tooth spacing and position, making them well suited to pattern recognition. In addition, alignment assessment relies on relatively stable and easily identifiable anatomical landmarks, which may reduce variability across evaluators and analytical approaches. These features are also less susceptible to perspective distortion caused by variation in camera positioning than sagittal relationships and arch morphology, both of which depend on more complex interarch spatial relationships and three-dimensional interpretation.

In contrast, recent studies have found that geometrically precise parameters, such as molar and canine relationships and dental arch morphology, show suboptimal performance, often reflected by lower AUC values, because they depend

https://doi.org/10.4258/hir.2026.32.2.1

on fine-grained spatial relationships and three-dimensional interpretation that are difficult to infer from two-dimensional images [24]. The implications of these findings support the hypothesis that general multimodal learning models and LLMs may have value in preliminary orthodontic analysis, butthat their performance varies markedly across areas of orthodontic assessment [1]. This interpretation is further supported by recent large-scale benchmark evaluations showing that strong performance on broad benchmarks does not necessarily translate into effective reasoning in scientific and clinical domains [25].

Our results suggest that the evaluative performance of large language models is stronger for anterior than for posterior relationships. These findings are consistent with previous studies examining AI capabilities in orthodontics. For example, Hack et al. [4] reported that AI models perform better in identifying conspicuously visible features, such as anterior tooth position, than in assessing complex occlusal relationships that require robust spatial reference frameworks. Similarly, Stetzel et al. [15] found that deep learning models perform well in predicting aesthetically oriented components of the IOTN, illustrating the strengths of AI in visuallyoriented analyses. However, these findings contrast with those of the present study with respect to sagittal and transverse parameter assessment, while still remaining broadly consistent with prior reports that AI models are challenged when evaluating parameters that depend on strong spatial references [24].

This study also demonstrated variation in agreement and accuracy across the evaluated large language models (Tables 4 and 5). These intermodel differences may reflect differences in underlying architectures and training paradigms. For example, Transformer-based models, including Vision Transformers, process images holistically and may therefore detect general irregularities in dental alignment efficiently, although they remain limited in making subtle distinctions among densely interdependent orthodontic categories such as overjet and molarintercuspation [26]. These observations are consistent with previous reviews showing that CNNs trained on task-specific datasets outperform general AI systems, including LLMs, on tasksrequiring high anatomical specificity [7]. The observed differences in outputs among the evaluated large language models may therefore reflect variation in visual abstraction, depth of reasoning, and internal representational capacity than true conceptual understanding [27]. Overall, the aggregated findings likely reflect the representational constraints imposed by nonspecialized training data [28]. Notably, cross-model analyses in scientific AI research have shown high error correlations among leading large language models, suggesting shared inductive biases rather than independent failure modes [25].

From a healthcare informatics perspective, evaluating several LLMs together may be viewed as a form of collective model behavior. Although each model has distinct strengths and limitations, broad patterns remain evident, including stronger performance in detecting alignment anomalies and much weaker performance in identifying complexspatial characteristics [1,29]. These findings highlight the importance of learning from multiple models collectively while also recognizing the limitations of AI and the continuing need for human expertise [28]. The observed convergence across models further suggests that aggregating general-purpose LLMs may offer limited benefit for inherently spatial clinical tasksreinforcing the need for task-specific AI systems [25]. Future studies should therefore focus on developing task-specific orthodontic AI systems rather than continuing to test general-purpose LLMs alone. An important next step will be the development of appropriately labeled, standardized multiview intraoral photograph datasets curated by orthodonticspecialists [24]. Such datasets could support supervised learning models capable of finer-scale spatial reasoning and clinically interpretable outputs [15], although this will requirecollaboration between orthodontists and information technology specialists.

This study has several limitations, including variability in image quality, use of a single orthodontist as the reference standard, and a relatively small sample size, all of which may limit generalizability. Reliance on a single evaluator may introduce subjectivity; therefore, the observed Cohen's κ values may reflect both AI performance and variability in expert judgment. Future studies should include multiple orthodontists and assess inter-rater reliability to establish a more robust reference standard. In addition, some models generated unsolicited treatment suggestions rather than strictly diagnostic outputs, highlighting limited output controllability. Furthermore, none of the models explicitly expressed diagnostic uncertainty or provided clinical disclaimers, raising important ethical considerations regarding the use of multimodal AI in healthcare settings. Moreover, all models were evaluated using a single standardized prompting framework. Therefore, the reported results reflect performance under this specific prompting configuration only. Alternative prompting strategies were not evaluated, and different prompt designs may lead to variation in performance; however, ensemble prompting may be considered as an alternative. This limitation should be considered when interpreting the findings. In conclusion, current multimodal

AI models demonstrate limited, parameter-dependent accuracy in detecting orthodontic malocclusions from intraoral photographs. These findings emphasize the limitations of general-purpose AI systems for orthodontic decision support and highlight the need for task-specific models trained on clinically annotated datasets.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

ORCID

Hillda Herawati (https://orcid.org/0009-0001-4222-9159)
Joko Kusnoto (https://orcid.org/0009-0009-8955-4459)
Indrayadi Gunardi (https://orcid.org/0000-0002-5525-5559)
Anggit Wirasto (https://orcid.org/0000-0002-1707-5979)
Tri Erri Astoeti (https://orcid.org/0000-0002-9341-504X)

Supplementary Materials

Supplementary materials can be found via https://doi.org/10.4258/hir.2026.32.2.000.

References

1. Monill-Gonzalez A, Rovira-Calatayud L, d'Oliveira NG, Ustrell-Torrent JM. Artificial intelligence in orthodontics: where are we now? A scoping review. Orthod Craniofac Res 2021;24 Suppl 2:6-15. https://doi.org/10.1111/ocr.12517
2. Albalawi F, Abalkhail KA. Trends and application of artificial intelligence technology in orthodontic diagnosis and treatment planning: a review. Appl Sci 2022;12(22):11864. https://doi.org/10.3390/app122211864
3. Subramanian AK, Chen Y, Almalki A, Sivamurthy G, Kafle D. Cephalometric analysis in orthodontics using artificial intelligence: a comprehensive review. Biomed Res Int 2022;2022:1880113. https://doi.org/10.1155/2022/1880113
4. Hack M, Dragulin B, Hack L, ElSaafin M, Dumitrescu I, Stan D, et al. Comparative study on the results of orthodontic diagnostics by using algorithms generated by artificial intelligence and simple algorithms. Med Pharm Rep 2024;97(2):215-21. https://doi.org/10.15386/mpr-2702
5. Liu J, Zhang C, Shan Z. Application of artificial intelligence in orthodontics: current state and future perspectives. Healthcare (Basel) 2023;11(20):2760. https://doi.org/10.3390/healthcare11202760
6. Ryu J, Kim YH, Kim TW, Jung SK. Evaluation of artificial intelligence model for crowding categorization and extraction diagnosis using intraoral photographs. Sci Rep 2023;13(1):5177. https://doi.org/10.1038/s41598-023-32514-7
7. Zhang R, Zhang L, Zhang D, Wang Y, Huang Y, Wang D, et al. Development and evaluation of a deep learning model for occlusion classification in intraoral photographs. PeerJ 2025;13:e20140. https://doi.org/10.7717/peerj.20140
8. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. J Med Syst 2023;47(1):33. https://doi.org/10.1007/s10916-023-01925-4
9. Zheng J, Ding X, Pu JJ, Chung SM, Ai QY, Hung KF, et al. Unlocking the potentials of large language models in orthodontics: a scoping review. Bioengineering (Basel) 2024;11(11):1145. https://doi.org/10.3390/bioengineering11111145
10. Wu X, Huang Y, He Q. A large language model improves clinicians' diagnostic performance in complex critical illness cases. Crit Care 2025;29(1):230. https://doi.org/10.1186/s13054-025-05468-7
11. Hatia A, Doldo T, Parrini S, Chisci E, Cipriani L, Montagna L, et al. Accuracy and completeness of ChatGPT-generated information on interceptive orthodontics: a multicenter collaborative study. J Clin Med 2024;13(3):735. https://doi.org/10.3390/jcm13030735
12. Giannakopoulos K, Kavadella A, Aaqel Salim A, Stamatopoulos V, Kaklamanos EG. Evaluation of the performance of generative AI large language models ChatGPT, Google Bard, and Microsoft Bing Chat in supporting evidence-based dentistry: comparative mixed methods study. J Med Internet Res 2023;25:e51580. https://doi.org/10.2196/51580
13. Jasim ES, Kadhom ZM, Al-Groosh D. Comparative evaluation of orthodontic treatment needs index and the dental aesthetic index to assess the need for orthodontic treatment from the participants' perspective: a cross-sectional study. J Orthod Sci 2025;14:30. https://doi.org/10.4103/jos.jos_24_25
14. Roy P, Roy P, Koley S, Sheet S. Bolton's ratio variations in Angle's Class I, Class II and Class III malocclusions: an observational study. J Clin Exp Dent 2025;17(3):e280-5.

https://doi.org/10.4317/jced.62591
15. Stetzel L, Foucher F, Jang SJ, Wu TH, Fields H, Schumacher F, et al. Artificial intelligence for predicting the aesthetic component of the index of orthodontic treatment need. Bioengineering (Basel) 2024;11(9):861. https://doi.org/10.3390/bioengineering11090861
16. Brook PH, Shaw WC. The development of an index of orthodontic treatment priority. Eur J Orthod 1989; 11(3):309-20. https://doi.org/10.1093/oxfordjournals.ejo.a035999
17. Nguyen SM, Nguyen MK, Saag M, Jagomagi T. The need for orthodontic treatment among vietnamese school children and young adults. Int J Dent 2014;2014:132301. https://doi.org/10.1155/2014/132301
18. Kozanecka A, Sarul M, Kawala B, Antoszewska-Smith J. Objectification of orthodontic treatment needs: does the classification of malocclusions or a history of orthodontic treatment matter? Adv Clin Exp Med 2016; 25(6):1303-12. https://doi.org/10.17219/acem/62828
19. Braun S, Hnat WP, Fender DE, Legan HL. The form of the human dental arch. Angle Orthod 1998;68(1):29-36. https://doi.org/10.1043/0003-3219(1998)068<0029:TFOTHD>2.3.CO;2
20. Huynh N, Zhang J, Pliska B, Amin R, Narang I, Chadha N, et al. Prevalence of altered craniofacial morphology in children with OSA. J Sleep Res 2025;34(5):e70060. https://doi.org/10.1111/jsr.70060
21. Shah NH, Entwistle D, Pfeffer MA. Creation and adoption of large language models in medicine. JAMA 2023; 330(9):866-9. https://doi.org/10.1001/jama.2023.14217
22. Masood Y, Masood M, Zainul NN, Araby NB, Hussain SF, Newton T. Impact of malocclusion on oral health related quality of life in young people. Health Qual Life Outcomes 2013;11(1):25. https://doi.org/10.1186/1477-7525-11-25
23. Nahm FS. Receiver operating characteristic curve: overview and practical use for clinicians. Korean J Anesthesiol 2022;75(1):25-36. https://doi.org/10.4097/kja.21209
24. Vishwanathaiah S, Fageeh HN, Khanagar SB, Maganur PC. Artificial intelligence its uses and application in pediatric dentistry: a review. Biomedicines 2023;11(3):788. https://doi.org/10.3390/biomedicines11030788
25. Song Z, Lu J, Du Y, Yu B, Pruyn TM, Huang Y, et al. Evaluating large language models in scientific discovery [Internet]. Ithaca (NY): arXiv.org; 2025 [cited 2026 Mar 1]. Available from: https://arxiv.org/abs/2512.15567.
26. 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 [Internet]. Ithaca (NY): arXiv.org; 2020 [cited 2026 Mar 1]. Available from: https://arxiv.org/abs/2010.11929v1.
27. Azizi S, Hatampoor S, Tahamtan S. Applications of artificial intelligence in diagnosis and treatment planning of orthodontics: a narrative review. Saudi Dent J 2025; 37(7-9):70. https://doi.org/10.1007/s44445-025-00077-0
28. Ahmed AE, Aljohani WF, Abu Rukbah LK, Rajhi SA, Najmi NK, Zughlul MK, et al. The role of artificial intelligence in stroke imaging in emergency settings: a systematic review. Cureus 2025;17(10):e93941. https://doi.org/10.7759/cureus.93941
29. Jiang L, Chai Y, Li M, Liu M, Fok R, Dziri N, et al. Artificial hivemind: The open-ended homogeneity of language models (and beyond) [Internet]. Ithaca (NY): arXiv.org; 2025 [cited 2026 Mar 1]. Available from: https://arxiv.org/abs/2510.22954.

8. Bukti penerimaan proofreading serta artikel yang telah dilakukan proofreading

24 April 2026

UNIVERSITAS TRISAKTI

Joko Kusnoto <joko.k@trisakti.ac.id>

[HIR-26-031] Author Proof Review Request

4 messages

HIR <hir@kosmi.org> Fri, Apr 24, 2026 at 7:46 AM
Reply-To: HIR <hir@kosmi.org>
To: joko.k@trisakti.ac.id

대용량 첨부파일 1개 - 다운로드 기한 : 2026-05-08 / 최대 500회 가능

HIR-26-031_sup.pdf (35.99 MB)

Dear Dr. Kusnoto:

We are attaching the edited manuscript file for your review.
Please examine it carefully, especially **the editor's notes,** and let us know which parts you would like to revise.

Please note that this proof stage is not intended for extensive corrections, additions, or deletions.
It is recommended that editing be limited to correcting typographical errors, incorrect dates, grammatical errors, and updating information about references which were in press.

We have also done English proofreading of the main text. Kindly check for any unintended changes in meaning.
For your reference, we are attaching the file 'HIR-26-031_EngProof.docx' — you do **not need to return** this file

We would appreciate receiving your response **within 24 hours (by April 25)**.
★ Please download and open the PDF file 'HIR-26-031(+s).pdf', 'HIR-26-031_sup.pdf' and respond to th e editor's comments in the annotations.
★ Please do not edit the file directly.

If you have any questions, please don't hesitate to contact us.
Thank you very much.

The Editorial Office
Healthcare Informatics Research

==========

Healthcare Informatics Research
The Korean Society of Medical Informatics
Tel : +82-2-733-7637
E-mail : hir@kosmi.org
==========

2 attachments

HIR-26-031(+s).pdf
905K

HIR-26-031_EngProof.docx
2573K

Joko Kusnoto <joko.k@trisakti.ac.id> Fri, Apr 24, 2026 at 2:01 PM
To: HIR <hir@kosmi.org>

Dear Editorial Office Healthcare Informatics Research

We would like to submit the revision of the HIR-26-031 proofreading document (attached PDF file HIR-26-031(+s) REVISION). Please make the revision on:
- Authors affiliation
- Keywords
- Proofreading typing error (highlighted in purple with revision comment on seven locations)

Thank you very much for the kind opportunity to publish our manuscript at HIR and look forward to the further instructions.

Sincerely,
Joko Kusnoto

[Quoted text hidden]

HIR-26-031(+s) REVISION.pdf
853K

HIR <hir@kosmi.org> Fri, Apr 24, 2026 at 2:41 PM
Reply-To: HIR <hir@kosmi.org>
To: Joko Kusnoto <joko.k@trisakti.ac.id>

Dear Dr. Kusnoto,

Thank you for your message and for your careful review of the proofreading.
We have checked the revised file you submitted.

Regarding the authors' affiliations, please note that positions and academic degrees are not included in the byline.
Therefore, the first author's affiliation will appear as:
Faculty of Dentistry, Universitas Trisakti, Jakarta, Indonesia.

Thank you again for your thorough review.

Kind regards,
Editorial Office
Healthcare Informatics Research