Al-Rafidain J Med Sci. 2026;11(2):298-304.
DOI: https://doi.org/10.54133/ajms.v11i2.3361

AJMS

Online ISSN (3219-2789)

Research Article

Inter-Model Agreement of ChatGPT, Claude, and Grok for IBO-Based Lateral Cephalometric Measurements: An Exploratory Comparative Study

Sabina Saccomanno[1], Francesca Greco[2], Roberto Sacco[3], Lorenzo Silenzi Ederli[4]*, Lucia Memè[1]

[1]*Department of Life Science, Health, and Health Professions, Link University, Rome, Italy;* [2]*Department of Radiology, Nuovo San Giovanni Battista Hospital, Foligno, Italy;* [3]*Department of Clinical Dentistry, School of Dentistry, UiT The Arctic University of Norway, Tromsø, Norway;* [4]*Department of Life, Health & Environmental Sciences, Postgraduate School of Orthodontics, University of L'Aquila, L'Aquila, Italy*

Received: 12 August 2026; Revised: 8 October 2026; Accepted: 14 October 2026

Abstract

Background: General-purpose multimodal artificial intelligence (AI) systems can analyze radiographic images, but their numerical consistency for quantitative orthodontic measurements is unknown. ***Objective***: To compare the inter-model agreement of ChatGPT, Claude, and Grok when estimating standardized Italian Board of Orthodontics (IBO) cephalometric measurements on the same lateral cephalograms. ***Methods***: Twenty pre-existing lateral cephalograms from 20 patients aged 10-60 years were analyzed by ChatGPT, Claude, and Grok with the same standardized prompt. Twelve IBO parameters were requested. Three linear measurements were excluded because metric calibration was not validated, leaving nine angular parameters. Model-specific descriptive statistics, pairwise mean absolute differences (MAD), root mean square error, and two-way absolute-agreement single-measure intraclass correlation coefficients (ICC) were calculated. Friedman tests with Holm correction assessed systematic differences between models. ***Results***: Complete-case sample size ranged from 16 to 20 across angular parameters. ICCs were poor for all nine measurements (0.03-0.26). Pooled pairwise MADs were 2.14 degrees for ChatGPT-Claude, 2.71 degrees for ChatGPT-Grok, and 2.59 degrees for Claude-Grok. No angular parameter showed a statistically significant systematic difference among models after Holm correction. ***Conclusions***: ChatGPT, Claude, and Grok produced broadly similar group-level cephalometric estimates but poor case-level numerical agreement. General-purpose multimodal AI systems should not be considered interchangeable cephalometric measurement tools. Validation against an expert human reference standard is required before clinical use.

Keywords: Artificial intelligence; Cephalometry; Large language models; Multimodal artificial intelligence; Orthodontics; Reliability.

Key Messages

General-purpose multimodal AI systems can generate structured cephalometric measurements, but similar average values do not imply case-level interchangeability. In this study, ChatGPT, Claude, and Grok showed poor inter-model agreement across all nine analyzable angular parameters. Expert-supervised validation remains necessary before these systems are used for clinical cephalometric measurement.

اتفاق بين النماذج بين ChatGPT وClaude وGrok لقياسات الرأس العرضية المعتمدة على IBO: دراسة مقارنة استكشافية

الخلاصة

الخلفية: يمكن لأنظمة الذكاء الاصطناعي متعددة الوسائط العامة تحليل الصور الشعاعية، لكن اتساقها العددي للقياسات الكمية لتقويم الأسنان غير معروف. **الهدف**: مقارنة الاتفاق بين نماذج ChatGPT وكلود وغروك عند تقدير قياسات الرأس القياسية المعيارية لمجلس تقويم الأسنان الإيطالي (IBO) على نفس المقاطع الجانبية. **الطرائق**: تم تحليل عشرين رقاقة جانبيا موجودا مسبقا من 20 مريضا تتراوح أعمارهم بين 10-60 عاما بواسطة ChatGPT وClaude وGrok باستخدام نفس الطلب المعياري. تم طلب اثني عشر معاملة IBO. تم استبعاد ثلاثة قياسات خطية لأن المعايرة المترية لم تتحقق من صحتها، مما ترك تسعة معلمات زاوية. تم حساب إحصاءات وصفية خاصة بالنموذج، والفروق المطلقة المتوسطة الزوجية (MAD)، والخطأ المتوسط التربيعي الجذري، ومعاملات الارتباط داخل الفئة ذات التوافق المطلق ذو التوافق المطلق ثنائي الاتجاه. قيمت اختبارات فريدمان مع تصحيح هولم الفروق المنهجية بين النماذج. **النتائج**: تراوح حجم العينة في الحالة الكاملة بين 16 و20 عبر المعلمات الزاوية. كانت مؤشرات ICC ضعيفة لجميع القياسات التسعة (0.03-0.26). كانت معدلات التفاعل المزدوج المجمعة 2.14 درجة ل ChatGPT-Claude، و2.71 درجة ل ChatGPT-Grok، و2.59 درجة ل Claude-Grok. لم يظهر أي معلمة زاوية وجود فرق منهجي ذو دلالة إحصائية بين النماذج بعد تصحيح هولم. **الاستنتاجات**: قدم ChatGPT وكلود، وغروك تقديرات رأسية متشابهة إلى حد كبير على مستوى المجموعة ولكن في التوافق العددي ضعيف على مستوى الحالة. لا ينبغي اعتبار أنظمة الذكاء الاصطناعي متعددة الوسائط العامة أدوات قياس رأسية قابلة للتبادل. يتطلب الاستخدام السريري التحقق من المعيار البشري للخبير المرجعي البشري.

**Corresponding author*: Lorenzo S. Ederli. Department of Life, Health & Environmental Sciences, Postgraduate School of Orthodontics, University of L'Aquila, L'Aquila, Italy; Email: lorenzo.silenziederli@student.univaq.it

Article citation: Saccomanno S, Greco F, Sacco R, Ederli LS, Memè L. Inter-Model Agreement of ChatGPT, Claude, and Grok for IBO-Based Lateral Cephalometric Measurements: An Exploratory Comparative Study. *Al-Rafidain J Med Sci*. 2026;11(2):298-304. doi: https://doi.org/10.54133/ajms.v11i2.3361

© 2026 The Author(s). Published by Al-Rafidain University. This is an open access journal issued under the CC BY-NC-SA 4.0 license (https://creativecommons.org/licenses/by-nc-sa/4.0/).

INTRODUCTION

Cephalometric analysis remains an essential quantitative component of orthodontic diagnosis, treatment planning, growth assessment, and evaluation of treatment outcomes. Its reliability depends on reproducible localization of anatomical landmarks and consistent construction of skeletal and dentoalveolar measurements. Both analogue and digital methods are affected by observer-related and landmark-related variability, although contemporary digital tracing can provide efficient and reproducible measurements [1,2]. Artificial intelligence has increasingly been incorporated into dentistry for image interpretation, prediction, and automated measurement [3,4]. In orthodontics, automated cephalometric analysis is among the most extensively investigated applications. Reviews have described the rapid development of convolutional neural network-based landmark detection and increasingly accurate automated

workflows, while also emphasizing heterogeneity in datasets, reference standards, landmarks, and definitions of clinically acceptable errors [5-10]. Most of this literature concerns purpose-built computer-vision systems trained specifically to localize cephalometric landmarks. Experimental deep-learning models have achieved promising performance in comparison with human examiners [11-14]. Similarly, studies of automated or web-based cephalometric platforms have frequently reported satisfactory average agreement with conventional or computer-assisted tracing but have also identified measurement-specific discrepancies, particularly for dental and incisor-related variables [15-20]. External and commercial evaluations have reinforced this mixed picture. Human-versus-AI comparisons have found that performance differs across applications, landmarks, and cephalometric variables [21-25]. Recent systematic reviews caution that apparently small group-level differences or common 2-mm/2-degree thresholds should not be interpreted as proof of interchangeability, and they continue to recommend expert supervision [26,27]. General-purpose multimodal large language models (MLLMs) represent a different technological category. Rather than executing a deterministic landmark-localization pipeline, they integrate visual input with natural-language reasoning across broad domains. Medical imaging research has shown that these systems can interpret radiological images, but performance is strongly model- and task-dependent and remains limited by hallucinations, localization errors, and uncertain clinical generalizability [28-32]. Their ability to return plausible numerical values may therefore exceed their actual measurement reliability. To our knowledge, the numerical interchangeability of contemporary general-purpose multimodal AI systems for standardized lateral cephalometric measurements has not been directly evaluated using the same images and prompt. This exploratory study compared ChatGPT, Claude, and Grok on 20 lateral cephalograms using an IBO-based protocol. The primary objective was to quantify inter-model agreement for angular cephalometric measurements. Because no expert human reference tracing was available, the study was explicitly designed as an agreement study rather than an accuracy study.

METHODS

Study design and image sample

This exploratory retrospective methodological study included 20 pre-existing lateral cephalometric radiographs from 20 different patients, one radiograph per patient, aged 10-60 years. The radiographs had been acquired for clinical purposes between 2023 and 2026 from a private clinical archive. No radiograph was obtained specifically for research, and no additional radiographic exposure occurred. The sample was a non-random convenience sample. Without applying predefined clinical or radiographic selection criteria, the first 20 records encountered according to alphabetical archive order were included. Sex distribution was not available in the archived dataset and was not used in the analysis. This sampling procedure was chosen for feasibility and was not intended to provide a population-representative sample. The same 20 image files were submitted to each AI system. The images were used in the form retrieved from the archive, without investigator-applied cropping, brightness or contrast adjustment, resizing, enhancement, or other preprocessing. No cephalometric landmarks, lines, or manual tracings had been superimposed on the radiographs. Each image was mapped to a study case during data extraction to allow case-by-case statistical comparison among the three AI systems.

AI systems and standardized prompting

The evaluated systems were ChatGPT (GPT-5.6 Sol, OpenAI), Claude Sonnet 5 (Anthropic), and Grok 4.6 (xAI). All analyses were performed on 28 August 2026. Model names were recorded as displayed or identified for the systems used on the study date. Each platform received the same set of 20 radiographs and the same standardized Italian-language prompt. The prompt instructed the model to treat each radiograph as an independent case, identify predefined cephalometric landmarks and reference planes, calculate only the requested measurements, avoid using normative values to force or correct outputs, refrain from diagnosis and treatment advice, and report a measurement as non-evaluable rather than inventing a value when the necessary anatomy or metric calibration was insufficient. The complete standardized prompt is provided as Supplementary Material S1. For each platform, the 20 radiographs were uploaded together within a single conversation session. The model was instructed to analyze them as distinct cases. The first numerical output generated for each radiograph was retained as the definitive study output. No measurement was regenerated, manually corrected, or modified after comparison with another model. Outputs from one AI system were not shown to either of the other systems. No new conversation was initiated between individual radiographs; therefore, all cases remained within the same active conversational context.

IBO-based cephalometric measurements

The measurement framework was derived from the 2024-2025 Italian Board of Orthodontics (IBO) guidelines [33]. Requested landmarks and reference structures included Sella, Nasion, Point A, Pogonion, Anterior nasal spine, Posterior nasal spine, Gonion, Gnathion, and the long axes of the maxillary and mandibular incisors. Twelve parameters were requested: S.N/A, S.N/Pg, A.N/Pg, S.N/ANS.PNS, S.N/Go.Gn, ANS.PNS/Go.Gn, +1/ANS.PNS, -1/Go.Gn, -1/A.Pg, overjet, overbite, and the interincisal angle (Table 1). Angular values were requested to one decimal place. Linear values were to be expressed in millimeters only when a reliable metric calibration could be established. Across the three AI datasets, the linear variables -1/A.Pg, overjet, and overbite were not included in the agreement analysis

because metric calibration was not consistently validated from the raster images. The final quantitative comparison therefore comprised nine angular measurements.

Table 1: IBO-based cephalometric parameters requested from each AI system

No.	Domain	Parameter	Unit	Analysis
1	Sagittal skeletal	S.N/A	degree	Included
2	Sagittal skeletal	S.N/Pg	degree	Included
3	Sagittal skeletal	A.N/Pg	degree	Included
4	Vertical skeletal	S.N/ANS.PNS	degree	Included
5	Vertical skeletal	S.N/Go.Gn	degree	Included
6	Vertical skeletal	ANS.PNS/Go.Gn	degree	Included
7	Dentobasal	+1/ANS.PNS	degree	Included
8	Dentobasal	-1/Go.Gn	degree	Included
9	Dentobasal	-1/A.Pg	mm	Excluded: calibration NV
10	Dental	Overjet	mm	Excluded: calibration NV
11	Dental	Overbite	mm	Excluded: calibration NV
12	Dental	Interincisal angle	degree	Included

NV: not evaluable.

Ethical and data-protection considerations

No formal ethics committee approval or exemption was obtained for this retrospective methodological analysis. The radiographs were pre-existing clinical records, and no additional examination, intervention, treatment, or patient contact was performed for research purposes. The investigators report that verbal consent for research use of clinical radiographic data had been obtained from the patients. Patient identifiers were not used as analytical variables. Some archival source files retained identifying information in filenames or image overlays at the time of AI processing; these elements were not used for cephalometric measurement or case matching. No patient radiographs or identifying information are reproduced in this manuscript, and all materials prepared for publication are de-identified.

Statistical analysis

For each angular parameter, model-specific means, standard deviations, and evaluable sample sizes were calculated. Pairwise numerical agreement was summarized using mean signed difference, mean absolute difference (MAD), and root mean square error (RMSE). Inter-model reliability was assessed on complete cases using a two-way absolute-agreement, single-measure intraclass correlation coefficient (ICC). Percentile bootstrap 95% confidence intervals were estimated from 5,000 case-level resamples. ICC values below 0.50 were interpreted as poor, 0.50-0.75 as moderate, 0.75-0.90 as good, and above 0.90 as excellent [34]. Systematic differences among the three models were tested separately for each angular parameter using the Friedman test on complete cases. Holm correction was applied across the nine parameter-level tests. Statistical tests were two-sided, and an adjusted P value below 0.05 was considered statistically significant. The sample size was fixed at 20 radiographs as an exploratory study, and no a priori power calculation was performed. Analyses were conducted in Python 3.13.5 using pandas, NumPy, SciPy, scikit-learn, and stats models.

RESULTS

Twelve IBO measurements were requested from each model. Three linear variables (-1/A.Pg, overjet, and overbite) were excluded from quantitative agreement analysis because metric calibration was not consistently validated. The primary analysis therefore comprised nine angular measurements across 20 cases and three AI systems. ChatGPT and Grok returned numerical angular values for all 180 potential determinations. Claude returned 164 of 180 angular values, with missing values concentrated in cases in which specific anatomical structures were judged non-evaluable. Consequently, complete-case sample size ranged from 16 to 20 depending on the parameter. Model-specific means and standard deviations are presented in Table 2.

Table 2: Descriptive angular cephalometric estimates by AI system

Parameter	ChatGPT (n)	Claude (n)	Grok (n)
S.N/A	81.9±1.1 (20)	82.3±1.2 (20)	82.1±2.0 (20)
S.N/Pg	78.1±1.3 (20)	78.5±0.7 (19)	79± 2.1 (20)
A.N/Pg	3.7±2.1 (20)	3.8±1.5 (19)	3.1±0.5 (20)
S.N/ANS.PNS	8.3±0.9 (20)	8.3±0.7 (19)	9.1±1.2 (20)
S.N/Go.Gn	33.1±2.9 (20)	32.5±2.1 (18)	34.8±2.8 (20)
ANS.PNS/Go.Gn	24.8±2.1 (20)	24.2±1.4 (18)	25.8±1.9 (20)
+1/ANS.PNS	110.9±4.2 (20)	110.7±1.9 (18)	110.6±4.3 (20)
-1/Go.Gn	94.8±2.6 (20)	94.5±1.4 (17)	94±3.6 (20)
Interincisal angle	129.5±6.6 (20)	124.5±3.5 (16)	129.9±6.7 (20)

Data are presented as mean±SD. All values are degrees.

Group-level means were generally similar for several skeletal variables. For example, the mean S.N/A was 81.9 degrees for ChatGPT, 82.3 degrees for Claude, and 82.1 degrees for Grok. In contrast, larger differences were apparent for selected dental measurements, particularly the interincisal angle, for which the corresponding means were 129.5, 124.5, and 129.9 degrees. Absolute-agreement ICC estimates were below 0.30 for every angular parameter, indicating poor case-level inter-model reliability

(Table 3). ICCs ranged from 0.03 for S.N/Pg to 0.26 for both +1/ANS.PNS and -1/Go.Gn. Thus, no parameter reached the threshold conventionally interpreted as moderate agreement. Across all evaluable paired angular values, pooled MAD was 2.14 degrees for ChatGPT-Claude, 2.71 degrees for ChatGPT-Grok, and 2.59 degrees for Claude-Grok. Parameter-level MADs were lowest for S.N/ANS.PNS and highest for the interincisal angle. The pattern indicates that relatively similar model means can coexist with substantial case-by-case numerical disagreement.

Table 3: Inter-model agreement for the nine analyzable angular parameters

Parameter	n	ICC (95% CI)	MAD C-G1	MAD C-G2	MAD G1-G2	Holm *p*
S.N/A	20	0.06 (-0.20 to 0.30)	1.27	1.67	1.95	1.0
S.N/Pg	19	0.03 (-0.20 to 0.25)	1.14	1.99	1.7	1.0
A.N/Pg	19	0.04 (-0.17 to 0.25)	1.72	1.65	1.33	1.0
S.N/ANS.PNS	19	0.17 (-0.08 to 0.45)	0.59	1.22	1.18	0.566
S.N/Go.Gn	18	0.18 (-0.14 to 0.46)	2.18	3.08	3.09	1.0
ANS.PNS/Go.Gn	18	0.17 (-0.15 to 0.45)	1.66	2.05	2.07	1.0
+1/ANS.PNS	18	0.26 (-0.08 to 0.54)	2.89	3.97	3.64	1.0
-1/Go.Gn	17	0.26 (0.06 to 0.41)	2.29	2.71	2.82	1.0
Interincisal angle	16	0.23 (-0.04 to 0.45)	6.21	5.68	6.19	0.199

C: ChatGPT; G1: Claude; G2: Grok; ICC: two-way absolute-agreement single-measure intraclass correlation coefficient; CI: bootstrap confidence interval; MAD: mean absolute difference.

Friedman testing did not identify a statistically significant systematic difference among the three models for any of the nine angular measurements after Holm correction. The smallest unadjusted P value occurred for the interincisal angle, but it did not remain significant after multiplicity correction (Holm-adjusted p=0.199). Figure 1 summarizes ICC estimates, and Figure 2 shows pairwise MADs across parameters.

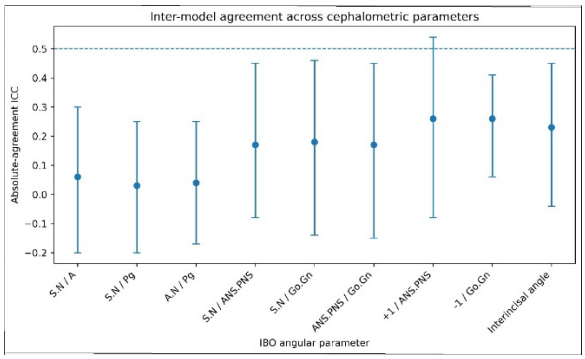


Figure 1: Absolute-agreement ICC for the nine analyzable IBO angular parameters. The dashed horizontal line indicates ICC = 0.50; all observed point estimates were below this level.

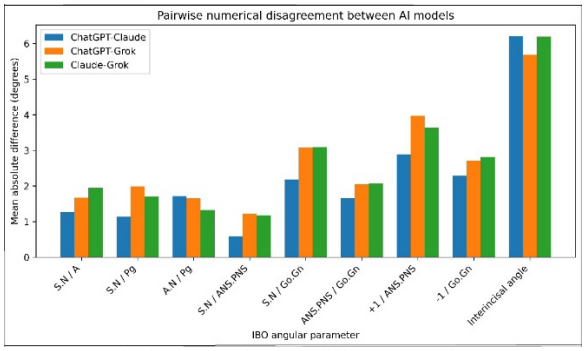


Figure 2: Pairwise mean absolute differences in degrees between ChatGPT, Claude and Grok for each analyzable angular parameter.

DISCUSSION

The principal finding of this study was that three general-purpose multimodal AI systems could all generate structured IBO-based angular cephalometric estimates, yet their case-level numerical agreement was consistently poor. ICC values remained below 0.30 for all nine analyzable parameters. This finding is important because group-level means were often close, and no systematic between-model difference remained significant after Holm correction. Similar averages therefore did not imply that the systems were interchangeable for an individual radiograph. Importantly, the present study evaluates agreement rather than accuracy. Agreement asks whether different models return similar numerical values for the same radiograph, whereas accuracy requires comparison with an external reference standard. Thus, the low ICCs show poor numerical interchangeability among the three models, but they do not prove that one model is more or less accurate than another. Conversely, close agreement between models would not by itself prove accuracy, because different systems could reproduce the same systematic localization error. Clinically, the present data cannot identify a preferred model or determine which output is closest to the true cephalometric value in the absence of expert reference tracing. This distinction between mean similarity and agreement is well recognized in measurement research and is especially relevant in cephalometry, where small differences in landmark position can propagate through angular constructions. Earlier work comparing analogue, digital, and automated tracing has shown that reproducibility depends on both the landmark and the measurement being derived [1,2,14]. Dedicated AI systems have often achieved favorable average performance, but several studies have identified broader limits of agreement for dental variables or measurements dependent on incisor inclination [15-20]. The present results are also consistent with comparative evaluations of commercial cephalometric AI. Indermun *et al*. found comparable overall precision between human-assisted and AI landmarking in a large South African dataset, while Silva *et al*. demonstrated application-specific differences in landmark accuracy and reproducibility [20,21]. Bor *et al*. observed significant measurement differences among multiple automated platforms [22]. Kunz *et al*. reported major provider-dependent variation relative to an expert gold standard, particularly for incisor-related parameters [23]. An AJMS study by Alhamdani *et al*. similarly documented differences between AI-driven and digital manual methods [24]. More recently, Azari-Mehr et al. showed that small mean differences could coexist with

poor ICC and wide limits of agreement for clinically important Steiner variables [25]. Together, these studies support the principle that aggregate similarity should not be equated with clinical interchangeability. Our study extends this literature from dedicated cephalometric software to general-purpose multimodal conversational AI. This is a meaningful distinction. Purpose-built systems typically identify explicit pixel coordinates and calculate measurements from a defined geometric pipeline, whereas MLLMs may infer anatomical relationships through a more general visual-language representation. Reviews of medical MLLMs describe rapid progress but emphasize limitations in grounding, transparency, hallucination control, and external validation [28,29]. Direct image-based comparisons have also shown substantial performance differences among general-purpose models in radiology [30,31]. In dentistry, a 2026 study of periapical radiographs found that MLLM performance became increasingly model-specific as anatomical localization demands increased [32]. This distinction is particularly relevant for quantitative cephalometry. Specialized CNN-based systems are trained to localize predefined landmarks and typically anchor those landmarks to explicit image coordinates, after which angles and distances are calculated through a fixed geometric pipeline [7-14]. General-purpose MLLMs are not primarily optimized for calibrated coordinate regression; instead, numerical answers may arise from broader visual-language representations without an exposed and reproducible set of landmark coordinates. Consequently, small variations in visual localization may propagate into different angular estimates while remaining difficult to audit. This lack of explicit spatial anchoring may partly explain why general-purpose MLLMs can generate plausible cephalometric values yet show weaker numerical consistency than purpose-built computer-vision systems [28,29]. The relatively larger disagreements observed for dentoalveolar measurements in the present sample are biologically and technically plausible. Superimposition, rotation, mixed dentition, or orthodontic hardware may obscure incisor axes, and small axis changes can significantly affect their angular measurements. Previous automated cephalometric studies have repeatedly identified dental landmarks and incisor-related variables as challenging [14,15,17,18,23,25]. General-purpose MLLMs do not currently provide a transparent coordinate-level tracing that allows an examiner to identify where a numerical discrepancy originated. The exclusion of the three linear measurements was methodologically appropriate. Pixel dimensions in heterogeneous raster exports cannot automatically be interpreted as millimeters. Previous validation studies of automated cephalometric systems generally rely on calibrated images or known geometric references when evaluating linear outcomes [15-19]. Reporting uncalibrated linear values would have created a false level of precision. For this reason, the present study is intentionally limited to angular outcomes. Another strength of the protocol was the use of the same images, the same prompt, and the first retained output for all three systems, with no post-hoc correction based on expected values or another model. This reduced investigator intervention after model output. However, the 20 images were processed in one conversation session per platform. Although the prompt explicitly instructed independent case analysis, context carry-over between images cannot be excluded. Because prior images and prior output remained available within the conversational context, later responses could theoretically have been influenced by earlier cases, response patterns, or inferred value ranges. The present design cannot determine whether such carry-over actually occurred. Future studies should therefore process each radiograph in an independent, stateless session, preferably with a randomized case order, and perform multiple independent runs for every model-image pair. This would help separate between-model disagreement from within-model stochastic variability and would permit direct estimation of intra-model repeatability.

Study Limitations

The study has several additional limitations. The sample was small and non-random, with broad age heterogeneity and no available sex distribution. Source images were heterogeneous raster exports rather than standardized calibrated DICOM files. Only one run per model and case was retained, so intra-model repeatability could not be estimated. The models did not output explicit landmark coordinates, preventing the direct calculation of landmark localization errors. Most importantly, no expert human reference tracing was available. The present findings, therefore, quantify inter-model agreement only and cannot determine which model is more accurate. Accordingly, the present inter-model comparison cannot be used to quantify model-specific error or bias, rank the systems by performance, or infer which numerical estimate is closest to the underlying anatomical reference. A definitive validation study should compare the same model outputs against independent tracings from at least two experienced orthodontists, with consensus adjudication of discrepant landmarks. Such a design would allow model-specific bias, MAE, RMSE, Bland-Altman limits of agreement, and clinically defined error thresholds to be assessed against a reference standard. Repeated model runs would additionally quantify stochastic intra-model variability. Until such evidence is available, general-purpose multimodal AI should be regarded as exploratory or assistive rather than a substitute for validated cephalometric software or expert tracing. Future reproducibility studies should incorporate repeated independent stateless runs for each model-case combination so that within-model ICCs and the dispersion of repeated numerical outputs can be quantified directly.

Conclusion

In this exploratory study of 20 lateral cephalograms, ChatGPT, Claude, and Grok produced structured IBO-based angular cephalometric measurements but showed poor inter-model agreement for every analyzed parameter (ICC 0.03-0.26). Similar group means did

not translate into case-level numerical interchangeability. These general-purpose multimodal AI systems should not currently be regarded as validated cephalometric measurement tools. Expert reference tracings, calibrated images, separate-session testing, and repeated AI runs are required before accuracy, and clinical reliability can be established.

Conflict of interests

The authors declared no conflict of interest.

Funding source

The authors did not receive any source of funds.

Data sharing statement

The data that supports the findings of this study are available from the corresponding author upon a reasonable request.

REFERENCES

1. Ongkosuwito EM, Katsaros C, van 't Hof MA, Bodegom JC, Kuijpers-Jagtman AM. The reproducibility of cephalometric measurements: a comparison of analogue and digital methods. *Eur J Orthod*. 2002;24(6):655-665. doi: 10.1093/ejo/24.6.655.
2. Narkhede S, Rao P, Sawant V, Sachdev SS, Arora S, Pawar AM, et al. Digital versus manual tracing in cephalometric analysis: A systematic review and meta-analysis. *J Pers Med*. 2024;14(6):566. doi: 10.3390/jpm14060566.
3. Schwendicke F, Samek W, Krois J. Artificial intelligence in dentistry: Chances and challenges. *J Dent Res*. 2020;99(7):769-774. doi: 10.1177/0022034520915714.
4. Khanagar SB, Al-Ehaideb A, Maganur PC, Vishwanathaiah S, Patil S, Baeshen HA, et al. Developments, application, and performance of artificial intelligence in dentistry - A systematic review. *J Dent Sci*. 2021;16(1):508-522. doi: 10.1016/j.jds.2020.06.019.
5. 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. doi: 10.1155/2022/1880113.
6. Kiełczykowski M, Kamiński K, Perkowski K, Zadurska M, Czochrowska E. Application of artificial intelligence (AI) in a cephalometric analysis: A narrative review. *Diagnostics (Basel)*. 2023;13(16):2640. doi: 10.3390/diagnostics13162640.
7. Schwendicke F, Chaurasia A, Arsiwala L, Lee JH, Elhennawy K, Jost-Brinkmann PG, et al. Deep learning for cephalometric landmark detection: systematic review and meta-analysis. *Clin Oral Investig*. 2021;25(7):4299-4309. doi: 10.1007/s00784-021-03990-w.
8. Mesquita GQTB, Vieira WA, Vidigal MTC, Travencolo BAN, Beaini TL, Spin-Neto R, et al. Artificial intelligence for detecting cephalometric landmarks: A systematic review and meta-analysis. *J Digit Imaging*. 2023;36(3):1158-1179. doi: 10.1007/s10278-022-00766-w.
9. Polizzi A, Leonardi R. Automatic cephalometric landmark identification with artificial intelligence: An umbrella review of systematic reviews. *J Dent*. 2024;146:105056. doi: 10.1016/j.jdent.2024.105056.
10. Hendrickx J, Gracea RS, Vanheers M, Winderickx N, Preda F, Shujaat S, et al. Can artificial intelligence-driven cephalometric analysis replace manual tracing? A systematic review and meta-analysis. *Eur J Orthod*. 2024;46(4):cjae029. doi:10.1093/ejo/cjae029.
11. Hwang HW, Park JH, Moon JH, Yu Y, Kim H, Her SB, et al. Automated identification of cephalometric landmarks: Part 2-Might it be better than human? *Angle Orthod*. 2020;90(1):69-76. doi: 10.2319/022019-129.1.
12. Hwang HW, Moon JH, Kim MG, Donatelli RE, Lee SJ. Evaluation of automated cephalometric analysis based on the latest deep learning method. *Angle Orthod*. 2021;91(3):329-335. doi: 10.2319/021220-100.1.
13. Kim YH, Lee C, Ha EG, Choi YJ, Han SS. A fully deep learning model for the automatic identification of cephalometric landmarks. *Imaging Sci Dent*. 2021;51(3):299-306. doi: 10.5624/isd.20210077.
14. Jeon S, Lee KC. Comparison of cephalometric measurements between conventional and automatic cephalometric analysis using convolutional neural network. *Prog Orthod*. 2021;22(1):14. doi: 10.1186/s40510-021-00358-4.
15. Meriç P, Naoumova J. Web-based fully automated cephalometric analysis: Comparisons between App-aided, computerized, and manual tracings. *Turk J Orthod*. 2020;33(3):142-149. doi: 10.5152/TurkJOrthod.2020.20062.
16. Mahto RK, Kafle D, Giri A, Luintel S, Karki A. Evaluation of fully automated cephalometric measurements obtained from web-based artificial intelligence driven platform. *BMC Oral Health*. 2022;22(1):132. doi: 10.1186/s12903-022-02170-w.
17. Çoban G, Öztürk T, Hashimli N, Yağci A. Comparison between cephalometric measurements using digital manual and web-based artificial intelligence cephalometric tracing software. *Dental Press J Orthod*. 2022;27(4):e222112. doi: 10.1590/2177-6709.27.4.e222112.oar.
18. Duran GS, Gökmen Ş, Topsakal KG, Görgülü S. Evaluation of the accuracy of fully automatic cephalometric analysis software with artificial intelligence algorithm. *Orthod Craniofac Res*. 2023;26(3):481-490. doi: 10.1111/ocr.12633.
19. Bao H, Zhang K, Yu C, Li H, Cao D, Shu H, et al. Evaluating the accuracy of automated cephalometric analysis based on artificial intelligence. *BMC Oral Health*. 2023;23(1):191. doi: 10.1186/s12903-023-02881-8.
20. Indermun S, Shaik S, Nyirenda C, Johannes K, Mulder R. Human examination and artificial intelligence in cephalometric landmark detection-is AI ready to take over? *Dentomaxillofac Radiol*. 2023;52(6):20220362. doi: 10.1259/dmfr.20220362.
21. Silva TP, Pinheiro MCR, Freitas DQ, Gaêta-Araujo H, Oliveira-Santos C. Assessment of accuracy and reproducibility of cephalometric identification performed by 2 artificial intelligence-driven tracing applications and human examiners. *Oral Surg Oral Med Oral Pathol Oral Radiol*. 2024;137(4):431-440. doi: 10.1016/j.oooo.2024.01.011.
22. Bor S, Çınarsoy Ciğerim S, Kotan S. Comparison of AI-assisted cephalometric analysis and orthodontist-performed digital tracing analysis. *Prog Orthod*. 2024;25(1):41. doi: 10.1186/s40510-024-00539-x.
23. Kunz F, Stellzig-Eisenhauer A, Widmaier LM, Zeman F, Boldt J. Assessment of the quality of different commercial providers using artificial intelligence for automated cephalometric analysis compared to human orthodontic experts. *J Orofac Orthop*. 2025;86(3):145-160. doi: 10.1007/s00056-023-00491-1.
24. Alhamdani NA, Khoshnaw BM, Omer ZQ. Assessing the accuracy and reliability of fully AI-driven cephalometric analysis in comparison to digital manual methods. *Al-Rafidain J Med Sci*. 2025;8(2):71-75. doi: 10.54133/ajms.v8i2.1839.
25. Azari-Mehr A, Bisbal-Puchades A, Marqués-Martínez L, Carmona-Santamaria M, García-Miralles E, Aura-Tormos JI, et al. Clinical reliability of AI-based cephalometric analysis using WebCeph: A comparative agreement study. *J Clin Med*. 2026;15(11):4155. doi: 10.3390/jcm15114155.
26. Ribas-Sabartés J, Sánchez-Molins M, d'Oliveira NG. The accuracy of algorithms used by artificial intelligence in cephalometric points detection: A systematic review. *Bioengineering (Basel)*. 2024;11(12):1286. doi: 10.3390/bioengineering11121286.
27. Lin YM, Auta A, Brown J, Cure R, Ireland R. The accuracy of artificial intelligence in identifying cephalometric landmarks: A scoping review. *Int Orthod*. 2026;24(4):101205. doi: 10.1016/j.ortho.2026.101205.
28. Nam Y, Kim DY, Kyung S, Seo J, Song JM, Kwon J, et al. Multimodal large language models in medical imaging: Current state and future directions. *Korean J Radiol*. 2025;26(10):900-923. doi: 10.3348/kjr.2025.0599.
29. Zhang A, Zhao E, Wang R, Zhang X, Wang J, Chen E. Multimodal large language models for medical image diagnosis: Challenges and opportunities. *J Biomed Inform*. 2025;169:104895. doi: 10.1016/j.jbi.2025.104895.
30. von der Stück MS, Vuskov R, Westfechtel S, Siepmann R, Kuhl C, Truhn D, et al. Visual large language models in radiology: A systematic multimodel evaluation of diagnostic

accuracy and hallucinations. *Life (Basel)*. 2026;16(1):66. doi: 10.3390/life16010066.
31. Sozer A, Sahin MC, Sozer B, Erol G, Tufek OY, Nernekli K, et al. Do LLMs have 'the eye' for MRI? Evaluating GPT-4o, Grok, and Gemini on brain MRI performance: First evaluation of Grok in medical imaging and a comparative analysis. *Diagnostics (Basel)*. 2025;15(11):1320. doi: 10.3390/diagnostics15111320.
32. Hatipoğlu Ö, Baytar M, Hatipoğlu FP. Diagnostic and localization performance of multimodal large language models in the interpretation of dental periapical radiographs. *BMC Oral Health*. 2026;26(1):1428. doi: 10.1186/s12903-026-08870-x.
33. Italian Board of Orthodontics. Linee guida della sessione d'esami I.B.O.: Certificato d'Eccellenza in Ortodonzia. 2024-2025.
34. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med*. 2016;15(2):155-163. doi: 10.1016/j.jcm.2016.02.012.