Iran J Orthod. 2026 December; 21(2):e1467. doi:10.32592/ijorth.19.1.1467.
Published online 2026 December 1 Research Article
Original Article **Open Access**

Agreement between AI-based and Human Assessment of Smile Attractiveness in Orthodontically Treated Class II Patients: A Method-Comparison Study

Muskan Chanduka [1], Shweta Nagesh [2]*

[1] Resident, Department of Orthodontics, Saveetha Dental College and Hospitals, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, India
[2] Assistant Professor, Department of Orthodontics, Saveetha Dental College and Hospitals, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, India

***Corresponding author:** Shweta Nagesh

Address: Assistant Professor, Department of Orthodontics, Saveetha Dental College and Hospitals, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, India,

Email: shwe14ta@gmail.com

Received: 2026 August 05; **Revised:** 2026 October 05; **Accepted:** 2026 October 30

Abstract

Aim: To evaluate the agreement and reliability between generic Artificial Intelligence (AI) tools and human panels (orthodontists and lay observers) in rating smile attractiveness before and after Class II functional orthodontic treatment, and to compare their sensitivity to treatment-induced esthetic changes.
Methods: Standardized frontal smile photographs of 30 skeletal Class II patients (aged 11–14 years) taken before (T1) and after (T2) twin-block and fixed appliance therapy were cropped and converted to grayscale. Sixty images were independently assessed on a 10-point Likert scale by ChatGPT (GPT-4o), Face Rate, 30 calibrated orthodontists, and 30 lay observers. Inter-rater, intra-rater, and test–retest reliability was evaluated using the intraclass correlation coefficient (ICC). Pairwise agreement was assessed using Bland–Altman analysis (predefined acceptable limits: ±2.0 points) and Lin's concordance correlation coefficient (CCC). Between-modality differences and within-modality pre–post changes were analyzed using Friedman and Wilcoxon signed-rank tests, respectively, with Bonferroni correction. The modality × time interaction was evaluated using generalized estimating equations (GEE).
Results: Pairwise individual-level agreement across all 12 comparisons was poor, with CCC values ranging from −0.155 to 0.041. Bland–Altman limits of agreement exceeded the ±2.0 point threshold in 11 of 12 comparisons. At T1, both AI tools assigned significantly higher attractiveness scores compared with orthodontists (mean difference: ChatGPT +2.99, Face Rate +2.83; $P<0.001$). At T2, while mean AI scores became comparable to orthodontists' scores, Lin's concordance remained negligible (CCC = 0.029 and −0.155). Orthodontists were the only rater group that detected a statistically significant post-treatment improvement (median change: +3.17 points, $P<0.001$), whereas changes detected by ChatGPT (+0.25), Face Rate (+0.11), and lay observers (+0.10) were not significant (all $P>0.10$). GEE modeling confirmed a significant modality × time interaction ($P<0.001$). AI tools showed high test–retest stability (ChatGPT ICC = 0.923; Face Rate: identical scores).
Conclusion: While generic AI tools captured general scoring trends, their limited individual-level agreement and lower sensitivity to subtle treatment changes indicate that they should be interpreted with caution. Therefore, current generic AI applications are best viewed as supplementary aids rather than standalone tools in orthodontic esthetic assessment.

Keywords: Artificial Intelligence, Orthodontics, Esthetics, Smile.

Copyright © 2026, Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/) which permits copy and redistribute the material just in noncommercial usages, provided the original work is properly cited.

Background

Facial and smile esthetics are fundamental to orthodontic practice because treatment success and patient satisfaction depend not only on dental and occlusal corrections but also on facial harmony and psychosocial well-being (1,2). In patients undergoing functional appliance therapy for Class II malocclusion, mandibular repositioning and associated soft tissue changes are expected to improve the facial profile and smile esthetics (3,4). Reliable esthetic assessment is essential for evaluating and appraising treatment outcomes. Such assessments have traditionally relied on rating scales administered to clinicians or laypeople. However, substantial perceptual differences exist between evaluator groups (5,6). Orthodontists tend to judge smile or facial attractiveness against diagnostic norms, such as smile arc and micro- and macro-aesthetics, whereas laypeople judge based on overall appearance and social acceptability and do not recognize subtle dentofacial disharmonies (6,7). As a result, professional and non-professional assessors often rate the same smile differently (8). Smile perception is also influenced by dental proportions, smile parameters, gingival display, and soft tissue factors (9,10). Because no universal definition of beauty exists, relying on a single group of evaluators may not fully represent treatment outcomes.

Artificial intelligence (AI) has emerged as a transformative force in orthodontics, with literature documenting its expanding role in diagnosis, treatment planning, and clinical decision-making, including generative AI platforms such as ChatGPT that are being integrated into clinical and educational workflows (11–13). Several AI-based applications now generate facial and smile attractiveness scores, and some of these scores show moderate to strong associations with human judgments (14,15). Previous studies have used AI to assess facial attractiveness after orthognathic surgery (16) and dental alignment (17). However, these studies used purpose-trained models, and evidence remains limited on whether commercially available generic AI tools can reliably assess smile attractiveness in an orthodontic context. Furthermore, no study has systematically evaluated the agreement between such tools and human raters using established method-comparison frameworks. Therefore, this study's primary objective was to evaluate agreement between two commercially available AI tools and human raters (orthodontists and lay observers) in assessing smile attractiveness, using pre- and post-orthodontic treatment photographs to capture a range of esthetic presentations. The secondary objective was to compare each assessment modality's sensitivity to treatment-related esthetic change.

Methods

Study Design and Setting

This observational study was conducted in the Department of Orthodontics, Saveetha Dental College and Hospitals, Chennai, India. The records of patients who completed treatment between 2021 and 2025 were used, and the images were rated between January and May 2026. The study was approved by the Institutional Ethics Committee of Saveetha Dental College and Hospitals (Reference number: SRB/SDC/ORTHO-2306/25/422) and was conducted in accordance with the Declaration of Helsinki. Written informed consent was obtained from the parents or legal guardians of all participants for the use of their de-identified images, including analysis by AI-based tools. The study was reported in accordance with the Guidelines for Reporting Reliability and Agreement Studies (GRRAS) (18).

All the patient identifiers were removed before analysis, and the images were cropped and converted to grayscale to reduce the possibility of identification; only these de-identified images were uploaded to the AI tools. ChatGPT sessions were not linked to patient records, and images were not shared or stored beyond the assessment process. The ethical protocol was guided by the ACCEPT-AI framework for pediatric data use in AI research (19).

Participants

Records were identified from the departmental database of patients with skeletal Class II malocclusion who had been treated with removable functional appliance therapy followed by fixed appliance therapy between 2021 and 2025. The inclusion criteria were: (i) age 11–14 years at the start of treatment; (ii) skeletal Class II malocclusion treated with a twin-block functional appliance followed by fixed appliance therapy; and (iii) standardized frontal smile photographs available before and after treatment. Exclusion criteria included craniofacial syndromes, facial asymmetry, developmental disorders, previous orthodontic treatment, extraction orthodontic treatment following functional appliance therapy, and incomplete or non-standardized photographs. Consecutive sampling was used, and all records meeting the eligibility criteria were included. Of 50 records screened, 20 were excluded (10 with incomplete photographic records and 10 with non-

standardized photographs), and 30 patients' records were included and analyzed. All the patients were treated with a twin-block appliance for 9–10 months, followed by fixed appliance therapy for alignment and finishing. A sensitivity power analysis (G*Power, version 3.1.9.2) showed that, with 30 patients, α=0.05 and power=0.80, the study could detect intra-subject pre–post differences with an effect size of dz≥0.54 (Wilcoxon's signed-rank test).

Image Acquisition and Standardization

Standardized frontal smile photographs were obtained at the pre-treatment stage (T1) and after completion of the full course of orthodontic treatment, comprising functional appliance therapy followed by fixed appliance treatment (T2), under controlled lighting, natural head posture, and consistent camera-to-subject distance. Images were cropped to include only the region inferior to the eyes and converted to grayscale to minimize bias from periocular features, skin tone, and complexion, and to enhance de-identification, ensuring assessment was focused on dental display, lip morphology, and smile dynamics (20). The same 60 processed images (30 at T1 and 30 at T2) were used for all four assessment modalities. An example of the standardized image used for assessment is presented in Figure 1.

Assessment Modalities and Outcomes

Smile attractiveness was independently evaluated with four modalities: two AI tools (ChatGPT and Face Rate) and two manual scoring methods (orthodontists and lay observers). ChatGPT (GPT-4o, OpenAI, San Francisco, California, USA) was accessed through its web interface. Each image was uploaded in a new chat session together with a standardized prompt instructing the model to rate smile attractiveness on a scale of 1 to 10, considering tooth display, alignment, smile arc, and lip dynamics, and to return a single overall score. Default model settings were used. Face Rate (Renaissance Apps LLC, USA) is a commercially available mobile application that identifies anatomical landmarks and geometric relationships to generate an attractiveness score on a 10-point scale. Each image was uploaded individually through the application interface, and the score displayed was recorded. Both tools assessed all 60 images.

Thirty calibrated orthodontists with over 3 years' experience and 30 lay observers (adults with at least graduate-level education and no dental training) were recruited from the institute. Before scoring, both raters and lay observers were calibrated in separate sessions regarding the scoring procedure and the tool. Afterward, both groups independently assessed smile esthetics. The size of each panel was based on previous smile attractiveness studies with comparable panel sizes (5,21). All raters evaluated the same 60 photographs (30 pre- and 30 post-treatment) using a 10-point Likert rating scale (1 = least attractive, 10 = most attractive). Both orthodontist and lay observer panels assessed photographs under blinded conditions; pre- and post-treatment images were presented in randomized order without identification of treatment phase. To assess intra-rater reliability, each rater re-scored 20 images (10 at T1 and 10 at T2) after 2 weeks. To assess the stability of the AI tools, a subset of 10 images was evaluated three times by each tool at 2-week intervals, with a new session opened for each evaluation.

The outcome was the smile attractiveness score (1–10). For each human rater group, the image score was the mean of the 30 raters' scores; for the AI tools, it was the tool's score.

Statistical Analysis

Statistical analysis was performed with Python (version 3.10) using the SciPy (Version 1.15; NumFOCUS, United States) and Statsmodels (version 0.15) libraries. All tests were two-tailed, and the significance level was set at $P<0.05$. Normality of the scores of each modality at each time point was assessed with the Shapiro–Wilk test. Inter-rater reliability within each human rater group was assessed with the intraclass correlation coefficient (ICC), and intra-rater reliability was assessed with the ICC calculated from the 20 re-scored images. Test–retest reliability of the AI tools was assessed with the ICC.

Agreement between each pair of modalities (six pairs) at each time point was evaluated with Bland–Altman analysis, reporting the mean difference (bias) with its 95% CI and the 95% limits of agreement (LoA; bias ± 1.96 SD of the differences) with their 95% CIs. Differences greater than ±2 points on the 10-point scale were defined a priori as clinically unacceptable. Lin's concordance correlation coefficient (CCC) with its 95% CI was calculated for each pair, together with its precision (Pearson's r) and accuracy (bias correction factor, Cb) components.

Differences between the four modalities at each time point were analyzed with Friedman's test, followed by pairwise Wilcoxon's signed-rank tests with Bonferroni correction (six comparisons per time point). Pre–post changes within each modality were analyzed with the Wilcoxon signed-rank test

with Bonferroni correction (four comparisons). To test whether the magnitude of change differed between modalities (modality × time interaction), Friedman's test was applied to the change scores (T2 − T1). Because each patient contributed eight correlated observations (four modalities at two time points), the interaction was also modeled with generalized estimating equations (GEE), with modality, time, and their interaction as factors, patient as the cluster, an exchangeable working correlation structure, and robust standard errors.

Results

Based on the Shapiro-Wilk test, several distributions departed significantly from normality, and the scores were derived from an ordinal scale; non-parametric tests were used for hypothesis testing. Scores were summarized as mean, standard deviation (SD), and 95% confidence interval (CI), and as median and interquartile range (IQR).

Intra-rater reliability was excellent for both human rater groups, with ICC values of 0.92 for orthodontists and 0.89 for lay observers. ICC for inter-rater reliability was moderate to good for human groups: orthodontists 0.76 (T1) and 0.78 (T2), lay observers 0.73 (T1) and 0.69 (T2). Test–retest reliability was assessed by evaluating a subset of 10 photographs (at T1 and T2) three times with each AI tool at 2-week intervals, opening a new session each time. Face Rate produced identical scores across all three trials for every patient, indicating perfect test–retest agreement; no inter-trial variance was available for ICC estimation. ChatGPT demonstrated excellent test–retest reliability (ICC=0.923).

Table 1 shows descriptive statistics. At T1, ChatGPT (7.25) and Face Rate (7.09) had the highest mean scores, lay observers (5.79) were intermediate, and orthodontists (4.26) had the lowest. At T2, the mean scores were 7.70 for ChatGPT, 7.38 for orthodontists, 7.18 for Face Rate, and 5.95 for lay observers.

Table 2 and Figures 2-4 show the Bland–Altman results. At T1, both AI tools showed a large positive bias relative to orthodontists. At T2, the bias relative to orthodontists was 0.33 points for ChatGPT and −0.20 points for Face Rate. Relative to lay observers, both AI tools showed a positive bias at T1 and at T2. The bias between ChatGPT and Face Rate was 0.16 points at T1 and 0.53 points at T2, with wide LoA at both time points. Orthodontist scores were 1.53 points lower than lay observer scores at T1 and 1.43 points higher at T2. The LoA were narrowest for Face Rate versus orthodontists at T2 (−1.05 to 0.65) and widest for ChatGPT versus orthodontists at T1 (0.55 to 5.43). Only the Face Rate–orthodontist comparison at T2 had LoA within the a priori limit of ±2 points.

Table 1: Smile attractiveness scores by assessment modality and time point

Modality	Time	Mean (SD)	95% CI	Median (IQR)
ChatGPT	T1	7.25 (1.31)	6.76–7.74	7.55 (6.33–8.10)
	T2	7.70 (0.97)	7.34–8.06	8.00 (7.35–8.38)
Face Rate	T1	7.09 (0.37)	6.96–7.23	7.07 (6.76–7.34)
	T2	7.18 (0.32)	7.06–7.30	7.15 (7.03–7.44)
Orthodontists	T1	4.26 (0.26)	4.17–4.36	4.27 (4.10–4.46)
	T2	7.38 (0.24)	7.29–7.47	7.42 (7.21–7.55)
Lay Observers	T1	5.79 (0.49)	5.61–5.97	5.68 (5.41–6.25)
	T2	5.95 (0.39)	5.81–6.10	5.92 (5.80–6.24)

SD = standard deviation; CI = confidence interval; IQR = interquartile range (25th–75th percentile); T1 = pre-treatment; T2 = post-treatment. Scores are on a 10-point scale.

Table 2: Bland–Altman analysis of agreement between assessment modalities

Time	Comparison	Bias (95% CI)	SD of differences	Lower LoA (95% CI)	Upper LoA (95% CI)
T1	ChatGPT vs. Orthodontists	2.99 (2.52 to 3.45)	1.25	0.55 (−0.26 to 1.35)	5.43 (4.63 to 6.24)
	Face Rate vs. Orthodontists	2.83 (2.64 to 3.02)	0.51	1.84 (1.51 to 2.17)	3.83 (3.50 to 4.16)
	ChatGPT vs. Lay Observers	1.46 (0.96 to 1.96)	1.34	−1.16 (−2.02 to −0.29)	4.08 (3.21 to 4.94)

Table 2: Bland–Altman analysis of agreement between assessment modalities

Time	Comparison	Bias (95% CI)	SD of differences	Lower LoA (95% CI)	Upper LoA (95% CI)
T2	Face Rate vs. Lay Observers	1.30 (1.08 to 1.53)	0.61	0.10 (−0.29 to 0.50)	2.51 (2.11 to 2.90)
	ChatGPT vs. Face Rate	0.16 (−0.37 to 0.68)	1.40	−2.59 (−3.49 to −1.68)	2.90 (1.99 to 3.80)
	Orthodontists vs. Lay Observers	−1.53 (−1.70 to −1.36)	0.45	−2.42 (−2.71 to −2.12)	−0.64 (−0.93 to −0.35)
	ChatGPT vs. Orthodontists	0.33 (−0.04 to 0.69)	0.98	−1.60 (−2.23 to −0.96)	2.25 (1.61 to 2.88)
	Face Rate vs. Orthodontists	−0.20 (−0.36 to −0.04)	0.44	−1.05 (−1.34 to −0.77)	0.65 (0.37 to 0.93)
	ChatGPT vs. Lay Observers	1.75 (1.35 to 2.15)	1.07	−0.35 (−1.04 to 0.35)	3.85 (3.16 to 4.54)
	Face Rate vs. Lay Observers	1.23 (1.04 to 1.41)	0.49	0.27 (−0.05 to 0.58)	2.18 (1.87 to 2.50)
	ChatGPT vs. Face Rate	0.53 (0.14 to 0.91)	1.03	−1.49 (−2.15 to −0.82)	2.54 (1.88 to 3.21)
	Orthodontists vs. Lay Observers	1.43 (1.24 to 1.61)	0.49	0.46 (0.14 to 0.78)	2.39 (2.07 to 2.71)

Bias = mean difference (first − second modality) in points on the 10-point scale; SD = standard deviation; LoA = 95% limits of agreement (bias ± 1.96 SD); CI = confidence interval; T1 = pre-treatment; T2 = post-treatment. A priori acceptable limits: ±2 points.

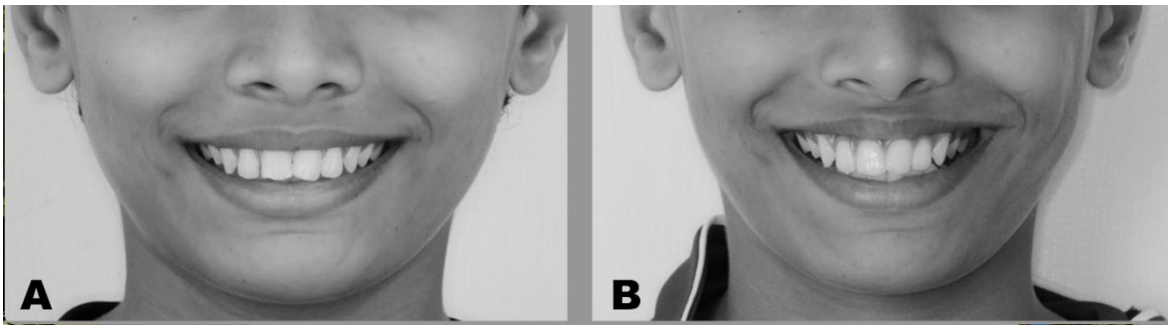


Figure 1: Standardized image used for assessment of smile attractiveness.

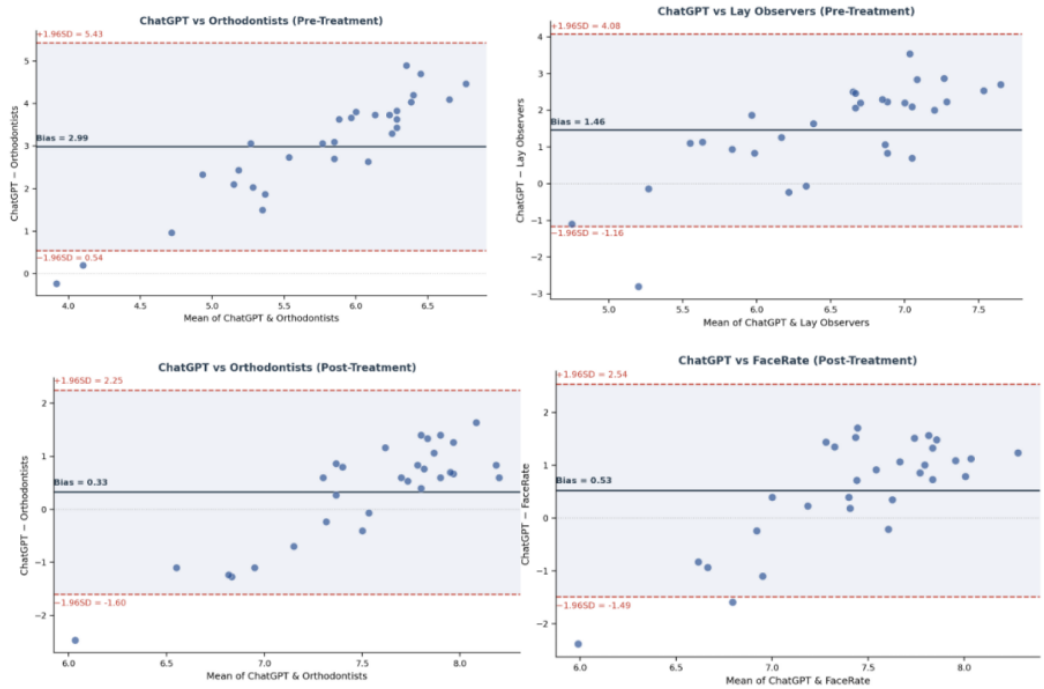


Figure 2: Bland–Altman plots comparing ChatGPT with Orthodontists and Lay Observers at pre-treatment (T1) and post-treatment (T2).

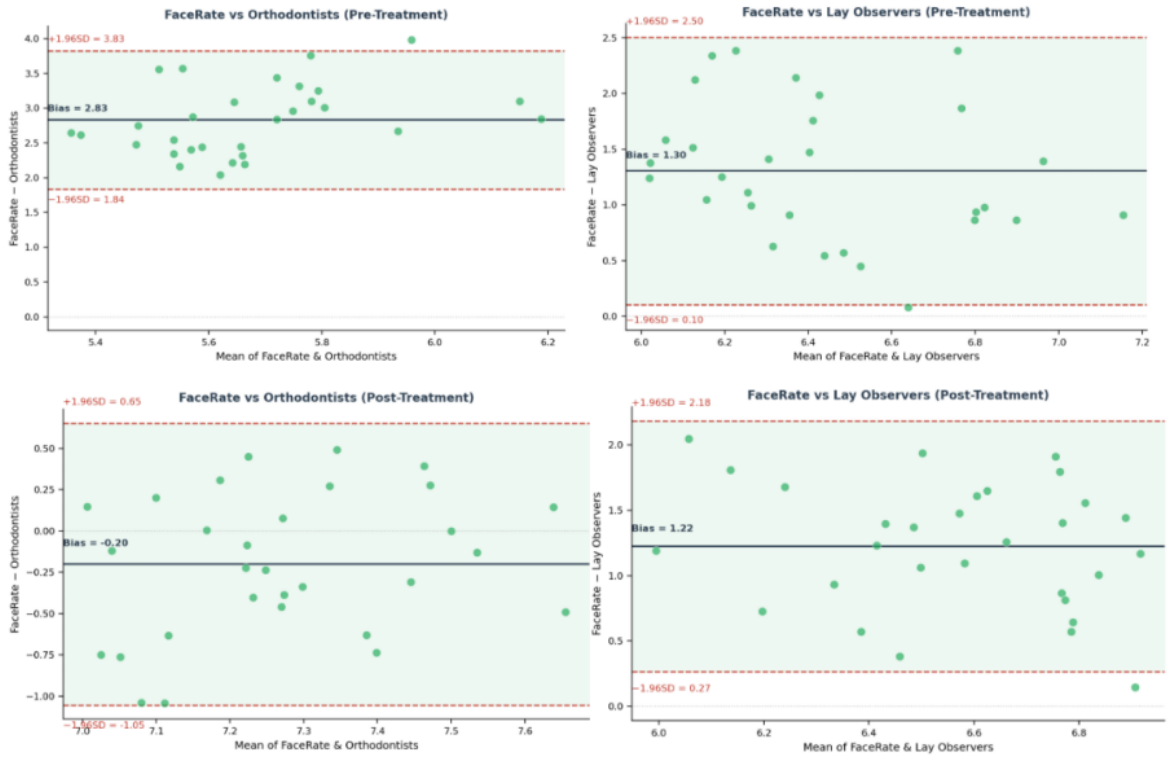


Figure 3: Bland–Altman plots comparing Face Rate with Orthodontists and Lay Observers at pre-treatment (T1) and post-treatment (T2).

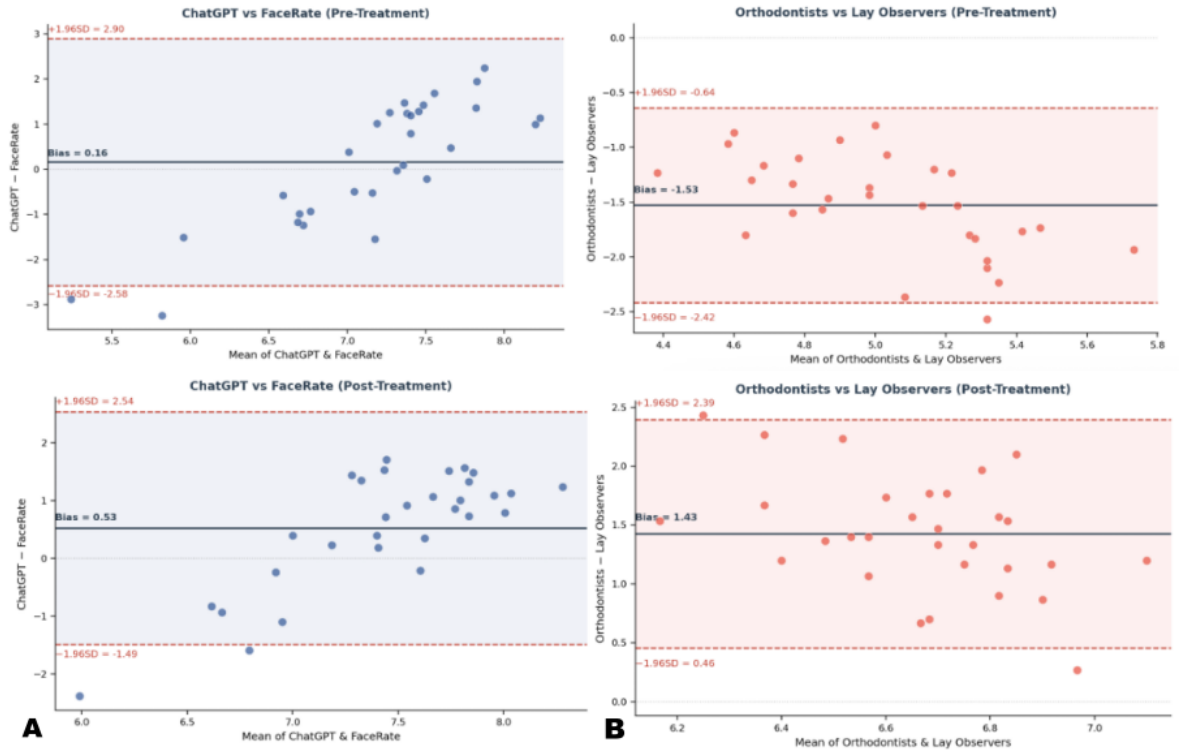


Figure 4: Bland–Altman plots comparing ChatGPT versus Face Rate, and Orthodontists versus Lay Observers at pre-treatment (T1) and post-treatment (T2).

The CCC was poor for all 12 comparisons, ranging from −0.155 to 0.041 (Table 3). The precision component (Pearson's r) ranged from −0.272 to 0.392, and the accuracy component (Cb) ranged from 0.023 to 0.766. For Face Rate versus orthodontists at T2, the narrowest LoA comparison yielded a CCC of −0.155 (95% CI: −0.416 to 0.130).

The scores differed significantly among the modalities at T1 (P<0.001) and T2 (P<0.001) (Table 4). At T1, all pairwise differences were significant (P<0.001) except between ChatGPT and Face Rate (P=1.000). At T2, the differences between ChatGPT and orthodontists (P=0.311), Face Rate and orthodontists (P=0.184), and ChatGPT and Face Rate (P=0.059) were not significant. In contrast, lay observer scores were significantly lower than those of all other modalities (P<0.001).

Table 3: Lin's concordance correlation coefficient (CCC) between assessment modalities

Time	Comparison	CCC (95% CI)	Pearson r	Cb
T1	ChatGPT vs. Orthodontists	0.021 (−0.003 to 0.044)	0.335	0.061
	Face Rate vs. Orthodontists	−0.006 (−0.015 to 0.003)	−0.272	0.023
	ChatGPT vs. Lay Observers	0.041 (−0.073 to 0.154)	0.132	0.307
	Face Rate vs. Lay Observers	0.000 (−0.063 to 0.063)	0.000	0.170
	ChatGPT vs. Face Rate	−0.054 (−0.241 to 0.136)	−0.104	0.519
	Orthodontists vs. Lay Observers	0.036 (0.000 to 0.073)	0.392	0.092
T2	ChatGPT vs. Orthodontists	0.029 (−0.128 to 0.184)	0.067	0.426
	Face Rate vs. Orthodontists	−0.155 (−0.416 to 0.130)	−0.202	0.766
	ChatGPT vs. Lay Observers	−0.013 (−0.079 to 0.053)	−0.072	0.179
	Face Rate vs. Lay Observers	0.008 (−0.043 to 0.059)	0.060	0.137
	ChatGPT vs. Face Rate	−0.017 (−0.186 to 0.152)	−0.038	0.461
	Orthodontists vs. Lay Observers	−0.013 (−0.044 to 0.018)	−0.155	0.083

CCC = Lin's concordance correlation coefficient; CI = confidence interval; Pearson r = precision component; Cb = bias correction factor (accuracy component; 1 = no systematic difference); T1 = pre-treatment; T2 = post-treatment.

Table 4: Comparison of smile attractiveness scores between assessment modalities (Friedman test and pairwise Wilcoxon signed-rank tests)

Time	Comparison	Median difference	Test statistic	P	Effect size
T1	Friedman test (all modalities)	—	$\chi^2(3) = 69.92$	<0.001*	W = 0.78
	ChatGPT vs. Orthodontists	3.20	Z = 4.74	<0.001*	r = 0.87
	Face Rate vs. Orthodontists	2.79	Z = 4.78	<0.001*	r = 0.87
	ChatGPT vs. Lay Observers	1.93	Z = 3.91	<0.001*	r = 0.71
	Face Rate vs. Lay Observers	1.25	Z = 4.78	<0.001*	r = 0.87
	ChatGPT vs. Face Rate	0.43	Z = 0.95	1.000	r = 0.17
	Orthodontists vs. Lay Observers	−1.50	Z = 4.78	<0.001*	r = 0.87
T2	Friedman test (all modalities)	—	$\chi^2(3) = 55.11$	<0.001*	W = 0.61
	ChatGPT vs. Orthodontists	0.60	Z = 1.94	0.311	r = 0.36
	Face Rate vs. Orthodontists	−0.18	Z = 2.16	0.184	r = 0.39
	ChatGPT vs. Lay Observers	2.05	Z = 4.59	<0.001*	r = 0.84
	Face Rate vs. Lay Observers	1.24	Z = 4.78	<0.001*	r = 0.87
	ChatGPT vs. Face Rate	0.82	Z = 2.58	0.059	r = 0.47
	Orthodontists vs. Lay Observers	1.40	Z = 4.78	<0.001*	r = 0.87

Median difference = median of paired differences (first − second modality) in points; P = Bonferroni-corrected P-value for pairwise comparisons (six per time point) and uncorrected p-value for the Friedman test; W = Kendall's W; r = effect size (Z/√N); T1 = pre-treatment; T2 = post-treatment. *Statistically significant

Orthodontist scores increased significantly from T1 to T2 (P<0.001) (Table 5). The changes in ChatGPT (P=0.182), Face Rate (P=0.692), and lay observer scores (P=1.000) were not significant after Bonferroni correction. The magnitude of change differed significantly among the modalities (P<0.001): the change in orthodontist scores was greater than that in each of the other modalities (all P<0.001), whereas the changes in ChatGPT, Face Rate, and lay observer scores did not differ from one another (all P≥0.666). GEE confirmed a significant modality × time interaction (P<0.001), with a significant estimated change only for orthodontists.

Discussion

The study's primary finding was that neither ChatGPT nor Face Rate agreed with orthodontists or lay observers at the individual patient level, as indicated by CCC values close to zero and LoA exceeding the predefined limit in 11 of 12 comparisons. Agreement also depended on the time point: both AI tools rated pre-treatment smiles about 3 points higher than orthodontists, whereas their mean post-treatment scores were close to orthodontists'. In addition, orthodontists were the only raters who detected a significant change in smile attractiveness after treatment. This finding aligns with Paduano et al. (3), who reported that orthodontists judged post-treatment silhouettes of Class II patients treated with functional appliances significantly more attractive and noted that this sensitivity also applies to frontal smile images.

ChatGPT detected a small, non-significant increase in scores after treatment. Kramer (14) reported medium-to-large associations between ChatGPT's attractiveness ratings and human judgments. However, that study used photographs of whole faces, whereas the present study used cropped grayscale smile images, which could explain the conflicting results. Also, ChatGPT is a general-purpose model not trained to assess orthodontic outcomes, so its ratings may reflect overall impressions rather than specific dental and soft tissue changes produced by treatment. Its excellent test–retest reliability indicates that poor agreement was not due to random variation between sessions. Face Rate did not detect a significant change, and its scores varied little between patients. It also showed excellent test–retest reliability, but its narrow score range suggests limited ability to discriminate between smiles.

These findings differ from studies that used purpose-trained AI models. Patcas et al. (16) used a convolutional neural network to show improvements in facial attractiveness after orthognathic surgery, and Obwegeser et al. (17) used an AI model to quantify the influence of dental esthetics on facial attractiveness. Those models were trained specifically for facial assessment and applied to full-face images, whereas the tools currently used are generic and broadly classify

overall esthetics. The present study results also align with a recent scoping review, which concluded that AI accuracy and reliability in orthodontics have not yet consistently surpassed those of expert clinicians (22).

Table 5: Pre–post change in smile attractiveness scores within each assessment modality

Modality	Median change (IQR)	Wilcoxon Z	P	r	GEE change (95% CI)	GEE P
ChatGPT	0.25 (−0.10 to 1.05)	2.00	0.182	0.37	0.45 (0.05 to 0.86)	0.113
Face Rate	0.11 (−0.16 to 0.34)	1.36	0.692	0.25	0.08 (−0.04 to 0.21)	0.747
Orthodontists	3.17 (2.97 to 3.30)	4.78	<0.001*	0.87	3.12 (3.02 to 3.22)	<0.001*
Lay Observers	0.10 (−0.29 to 0.47)	1.15	1.000	0.21	0.16 (−0.06 to 0.39)	0.629

Change = T2 − T1 in points on the 10-point scale; IQR = interquartile range; P = Bonferroni-corrected p-value (four comparisons); r = effect size (Z/√N); GEE = generalized estimating equations (exchangeable working correlation, robust standard errors); CI = confidence interval. Friedman test on change scores: $\chi^2(3) = 53.40$, $P<0.001$, $W = 0.59$. GEE modality × time interaction: Wald $\chi^2(3) = 1385.74$, $P<0.001$. *Statistically significant

The post-treatment similarity between the mean AI scores and orthodontist scores must be interpreted with caution. Similar mean scores do not mean the two methods rank patients in the same order, and the CCC values show they do not. Moreover, this similarity results mainly from a large increase in orthodontists' scores, whereas AI scores changed little. The apparent post-treatment agreement therefore reflects changes in orthodontist scores rather than improved AI performance.

Lay observers did not detect a significant change after treatment, and their post-treatment scores remained lower than orthodontists'. This agrees with previous studies reporting that dental professionals are more sensitive than laypeople to differences in smile esthetics (7,23,24). Lay observers also rated pre-treatment smiles higher than orthodontists, suggesting that laypeople were less critical of the dentofacial features of untreated Class II malocclusion.

Limitations

The limitations of the study must be acknowledged. First, the study was conducted at a single center and included a small sample of patients with a particular type of malocclusion and treatment, which limits the generalizability of the findings. Second, the study lacked an untreated control group, so observed changes cannot be attributed solely to treatment, as natural growth during adolescence may contribute to esthetic changes. However, the study's primary aim was method comparison rather than treatment efficacy, and the pre-post design provided a range of esthetic presentations for evaluating inter-method agreement. The images were standardized and converted to greyscale, which enhances de-identification but may not reflect how these tools perform with full-face or profile photographs. Also, because both AI tools are cloud-based, uploading patient images raises data governance considerations. However, only de-identified images were used, and ethical approval and consent were obtained; future studies should consider locally deployed AI models to further mitigate privacy risks. Finally, AI tools are rapidly evolving, and the performance characteristics may differ in future versions and newly developed tools, including the specific ChatGPT tool evaluated here.

Overall, the present study found that commercially available, generic AI tools have limited ability to detect treatment-induced changes. This suggests that current generic AI tools should not be used as standalone metrics for evaluating esthetic outcomes of orthodontic treatment. Orthodontist assessment remains the most sensitive modality for detecting treatment-induced improvements. Lay observers did not perceive esthetic changes that clinicians can detect. This must be considered when discussing expected treatment outcomes.

Conclusion

Assessment modalities showed limited individual-level agreement. Orthodontists exhibited higher sensitivity to treatment-induced esthetic changes compared to other raters. Lay observers did not demonstrate a statistically significant change in smile attractiveness scores following treatment. Generic AI tools, such as ChatGPT and Face Rate, showed restricted capability in reliably capturing subtle orthodontic outcomes; thus, their use should be approached with caution as supplementary rather than standalone metrics in clinical practice.

Acknowledgments

None.

Conflict of Interests

The authors have no competing interests to declare.

Data Availability

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

Funding

None.

Authors' Contributions

Muskan Chanduka: Methodology, Formal analysis, Writing – original draft, Data curation. Shweta Nagesh: Conceptualization, Methodology, Formal analysis, Writing – review and editing, Supervision

References

1. Coppola G, Christopoulou I, Gkantidis N, Verna C, Pandis N, Kanavakis G. The effect of orthodontic treatment on smile attractiveness: a systematic review. Prog Orthod. 2023;24(1):4. doi: 10.1186/s40510-023-00456-5.
2. Umar D., Husain A., Ashraf M., Shetty S., Koya S. Esthetic impact of alignment on smile components and attractiveness. Iran J Orthod. 2025;20(1):1–9. doi: 10.32592/ijorth.20.1.1196
3. Paduano S, Rongo R, Bucci R, Carvelli G, Cioffi I. Impact of functional orthodontic treatment on facial attractiveness of children with Class II division 1 malocclusion. Eur J Orthod. 2020;42(2):144-150. doi: 10.1093/ejo/cjz076. PMID: 31586412.
4. Sena LMF, Damasceno E Araújo LAL, Farias ACR, Pereira HSG. The influence of sagittal position of the mandible in facial attractiveness and social perception. Dental Press J Orthod. 2017;22(2):77-86. doi: 10.1590/2177-6709.22.2.077-086.oar.
5. Reis GM, de Freitas DS, Oliveira RC, de Oliveira RCG, Pinzan-Vercelino CRM, Freitas KMS, Valarelli FP. Smile attractiveness in class III patients after orthodontic camouflage or orthognathic surgery. Clin Oral Investig. 2021;25(12):6791-6797. doi: 10.1007/s00784-021-03966-w.
6. Vasudevan R, Selvam S, Nagesh S. Comparison of smile parameters, maxillary arch changes and perceived smile attractiveness in patients treated with conventional and self-ligating bracket systems: an observational study. J Clin Diagn Res 2025;19(4):ZC33–8. https://doi.org/10.7860/JCDR/2025/74240/20872.
7. Flores-Mir C, Silva E, Barriga MI, Lagravere MO, Major PW. Lay person's perception of smile aesthetics in dental and facial views. J Orthod. 2004;31(3):204-9. doi: 10.1179/146531204225022416.
8. Al-Bitar Z, Attari N, Abu Arqub S, Al-Omiri M. Analyzing how personality factors influence the perception of smile aesthetics across diverse dental educational backgrounds and lay audiences. Prog Orthod. 2025;26(1):52. doi: 10.1186/s40510-025-00599-7.
9. Pham TAV, Nguyen PA. Morphological features of smile attractiveness and related factors influence perception and gingival aesthetic parameters. Int Dent J. 2022;72(1):67-75. doi: 10.1016/j.identj.2021.02.001.
10. Raj P, Nagesh S, Boyapati R. Impact of gingival pigmentation on laypersons' perception of smile aesthetics: an observational study. Explor Med. 2025; 6:1001274. doi: 10.37349/emed.2025.1001274
11. Subramanian AK, Marrapodi MM, Franco R, et al. Role of artificial intelligence (AI) in orthodontic diagnosis and treatment planning: a systematic review. Bull Stomatol Maxillofac Surg 2024;20(3):64–84.doi:10.58240/1829006X-2024.3-64.
12. LA Rosa S, Uzunçibuk H, Almeida LE, Meto A, Veeraraghavan VP, Heboyan A. The impact of artificial intelligence on orthodontics: a systematic review of applications and implications. Minerva Dent Oral Sci. 2025;74(3):195-208. doi: 10.23736/S2724-6329.24.04930-1.
13. Ghaffar RSA, Malik R, Amjad M, Nadeem R, Admani MS, Sanaullah A, Altamash S, Ahmed N, Heboyan A, Jouhar R. Awareness, usage, and perspectives on ChatGPT in dental education among graduate and undergraduate students. Sci Rep. 2025;16(1):3747. doi: 10.1038/s41598-025-33704-1.
14. Kramer RSS. Comparing ChatGPT with human judgements of social traits from face photographs. Comput Hum Behav Artif Hum 2025;4:100156. doi:10.1016/j.chbah.2025.100156.
15. Moridani MK, Jamiee N, Saghafi S. Human-like evaluation by facial attractiveness intelligent machine. Int J Cogn Comput Eng 2023;4:160–9. doi: 10.1016/j.ijcce.2023.04.001

16. Patcas R, Bernini DAJ, Volokitin A, Agustsson E, Rothe R, Timofte R. Applying artificial intelligence to assess the impact of orthognathic treatment on facial attractiveness and estimated age. Int J Oral Maxillofac Surg. 2019;48(1):77-83. doi: 10.1016/j.ijom.2018.07.010.
17. Obwegeser D, Timofte R, Mayer C, Eliades T, Bornstein MM, Schätzle MA, Patcas R. Using artificial intelligence to determine the influence of dental aesthetics on facial attractiveness in comparison to other facial modifications. Eur J Orthod. 2022;44(4):445-451. doi: 10.1093/ejo/cjac016.
18. Kottner J, Audigé L, Brorson S, Donner A, Gajewski BJ, Hróbjartsson A, Roberts C, Shoukri M, Streiner DL. Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. J Clin Epidemiol. 2011;64(1):96-106. doi: 10.1016/j.jclinepi.2010.03.002.
19. Muralidharan V, Burgart A, Daneshjou R, Rose S. Recommendations for the use of pediatric data in artificial intelligence and machine learning ACCEPT-AI. NPJ Digit Med. 2023;6(1):166. doi: 10.1038/s41746-023-00898-5.
20. Thiruvenkatachari B, Javidi H, Griffiths SE, Shah AA, Sandler J. Extraction of maxillary canines: Esthetic perceptions of patient smiles among dental professionals and laypeople. Am J Orthod Dentofacial Orthop. 2017;152(4):509-515. doi: 10.1016/j.ajodo.2017.02.015.
21. Akyalcin S, Frels LK, English JD, Laman S. Analysis of smile esthetics in American Board of Orthodontic patients. Angle Orthod. 2014;84(3):486-91. doi: 10.2319/072813-562.1.
22. Gracea RS, Winderickx N, Vanheers M, Hendrickx J, Preda F, Shujaat S, Cadenas de Llano-Pérula M, Jacobs R. Artificial intelligence for orthodontic diagnosis and treatment planning: A scoping review. J Dent. 2025;152:105442. doi: 10.1016/j.jdent.2024.105442.
23. Ravichandra S, Riyaz K, Manjappa LS, Talwar A, Narsimha Mogaveera NK, Ghosh PP, Gupta S. Comparative assessment of smile attractiveness perception and treatment need across dental professionals and laypersons. Cureus. 2025;17(10):e94037. doi: 10.7759/cureus.94037.
24. Cracel-Nogueira F, Pinho T. Assessment of the perception of smile esthetics by laypersons, dental students and dental practitioners. Int Orthod. 2013;11(4):432-44. doi: 10.1016/j.ortho.2013.09.007.