Original Article

DOI: 10.14235/bas.galenos.2026.82085
Bezmialem Science

Comparative Analysis of Aesthetic Judgment Between Orthodontists, General Dentists, and Different Versions of Large Language Models Using the IOTN-AC Index

IOTN-AC İndeksi Kullanılarak Ortodontistler, Genel Diş Hekimleri ve Büyük Dil Modellerinin Farklı Sürümleri Arasındaki Estetik Değerlendirmelerin Karşılaştırmalı Analizi

Rümeysa Cemre ÖNCÜ, Aylin PAŞAOĞLU BOZKURT

İstanbul Aydın University Faculty of Dentistry, Department of Orthodontics, İstanbul, Türkiye

ABSTRACT

Objective: This study aimed to compare the performance and diagnostic accuracy of general dentists and different artificial intelligence (AI)-based large language models (LLMs), including their older and newer versions, in photograph-based IOTN-AC index assessments, using orthodontists' evaluations as the reference standard.

Methods: The mean IOTN-AC scores derived from orthodontists were established as the reference (gold) standard. Cases were evaluated by general dentists and both earlier and newer versions of the ChatGPT, Claude, and Gemini families. Performance was analyzed using mean absolute error (MAE), Spearman's correlation, Lin's concordance correlation coefficient (CCC), systematic bias (Bland-Altman analysis), and diagnostic accuracy metrics (sensitivity/specificity) based on the clinical treatment need threshold of IOTN-AC ≥4.0.

Results: General dentists demonstrated the closest agreement with the reference standard, exhibiting the lowest MAE (0.761) and the highest CCC (0.901, 95% confidence interval: 0.796-0.948). Conversely, all AI models exhibited a positive bias by systematically overestimating aesthetic impairment (+0.761 to +3.361). In the longitudinal analysis, although newer iterations of LLMs showed improved ranking performance (correlation), they regressed in absolute accuracy compared to earlier versions, exhibiting increased

ÖZ

Amaç: Bu çalışma, genel diş hekimlerinin ve yapay zeka (YZ) tabanlı farklı büyük dil modelllerinin (LLM) eski ve yeni versiyonlarıyla birlikte fotoğraf tabanlı IOTN-AC indeksi değerlendirme performanslarını ve tanısal doğruluklarını ortodontistlerin referans standartlarına göre karşılaştırmayı amaçlamıştır.

Yöntemler: Ortodontistlerin ortalama IOTN-AC puanları referans (altın) standart olarak kabul edilmiştir. Olgular genel diş hekimleri ile ChatGPT, Claude ve Gemini ailelerinin eski ve yeni versiyonları tarafından değerlendirilmiştir. Performanslar; ortalama mutlak hata (MAE), Spearman korelasyonu, Lin'in konkordans korelasyon katsayısı (CCC), sistematik bias (Bland-Altman) ve IOTN-AC ≥4,0 klinik eşiğine göre tanısal doğruluk (duyarlılık/özgüllük) metrikleri kullanılarak analiz edilmiştir.

Bulgular: Genel diş hekimleri, en düşük MAE (0,761) ve en yüksek CCC (0,901, %95 güven aralığı: 0,796-0,948) ile referans standarda en yakın uyumu göstermiştir. YZ modelleri ise estetik bozukluğu olduğundan yüksek değerlendirerek pozitif bias sergilemiştir (+0,761 ila +3,361). Boylamsal analizde, LLM'lerin yeni iterasyonları sıralama performansında (korelasyon) artış gösterse de, mutlak doğrulukta eski versiyonlara göre gerilemiş ve MAE değerleri artmıştır (örn. Gemini 3 Pro MAE 2,529'dan 3,406'ya yükselmiştir; p=0,023). IOTN-AC ≥4,0 eşiğinde LLM'ler yüksek duyarlılık (%84,6-100,0) gösterirken,

Address for Correspondence: Rümeysa Cemre Öncü, İstanbul Aydın University, Faculty of Dentistry, Department of Orthodontics, İstanbul, Türkiye
E-mail: rcemresahin@stu.aydin.edu.tr
ORCID IDs of the authors: R.C.Ö.: 0000-0002-7102-9324, A.P.B.: 0000-0001-8100-4684

Cite this article as: Öncü RC, Paşaoğlu Bozkurt A. Comparative analysis of aesthetic judgment between orthodontists, general dentists, and different versions of large language models using the IOTN-AC index. Bezmialem Science. [Epub Ahead of Print]

Received: 06.06.2026
Accepted: 23.07.2026
Epub: 02.09.2026

ABSTRACT

MAE values (e.g., Gemini 3 Pro MAE increased from 2.529 to 3.406; p=0.023). At the IOTN-AC ≥4.0 threshold, while LLMs demonstrated high sensitivity (84.6-100.0%), they critically failed to distinguish healthy cases, resulting in extremely low specificity (e.g., 5.9% for Gemini 3 Pro).

Conclusion: Within the limitations of this study's sample and pooled evaluation design, general dentists demonstrated higher reliability and closer agreement with the reference standard than AI models in orthodontic aesthetic assessment. The architectural advancements in newer model versions did not translate positively into absolute clinical accuracy, creating a performance paradox. Although current LLMs show potential as preliminary screening tools due to their high sensitivity, they carry a significant risk of over-diagnosis owing to their low specificity, remaining far from being independent clinical decision support systems.

Keywords: Artificial Intelligence, multimodal large language models, IOTN-AC, orthodontic treatment need, orthodontists, aesthetics

ÖZ

sağlıklı olguları ayırt etmede (özgüllük) kritik düzeyde başarısız olmuştur (örn., Gemini 3 Pro için %5,9).

Sonuç: Bu çalışmanın örneklemi ve havuzlanmış değerlendirme tasarımı sınırları içinde, genel diş hekimleri ortodontik estetik değerlendirmede YZ modellerinden daha yüksek güvenilirlik ve referans standarda daha yakın uyum göstermiştir. Modellerin yeni versiyonlarındaki mimari gelişim, mutlak klinik doğruluğa olumlu yansımamış ve bir performans paradoksu yaratmıştır. Güncel LLM'ler yüksek duyarlılıkları ile potansiyel bir tarama aracı olsalar da, düşük özgüllükleri nedeniyle ciddi bir "aşırı teşhis" (*over-diagnosis*) riski taşımakta ve henüz bağımsız bir klinik karar destek sistemi olmaktan uzak kalmaktadırlar.

Anahtar Kelimeler: Yapay zeka, çok modlu büyük dil modelleri, IOTN-AC, ortodontik tedavi ihtiyacı, ortodontistler, estetik

Introduction

Oral and dental health problems are among the most prevalent health conditions worldwide. Along with dental caries, periodontal diseases, and dental fluorosis, malocclusions represent one of the most frequently observed dental problems (1). Individuals with malocclusion may experience aesthetic concerns, speech difficulties, and masticatory dysfunction. Dental crowding is the primary reason patients seek orthodontic treatment (2). Achieving an attractive smile requires the correction of crowding and elimination of malocclusions, processes that are primarily managed by orthodontists (3,4).

Access to appropriate orthodontic treatment largely depends on the orthodontist's expertise and the quality of professional training. Orthodontic care may be pursued to improve both masticatory function and facial aesthetics, depending on the patient's awareness and expectations. General dentists play a crucial role in identifying patients who require referral, whereas orthodontists determine the necessity and scope of treatment. Postgraduate specialization has become increasingly important for enhancing clinical competence, improving diagnostic accuracy, promoting research awareness, and reducing the risk of complications (5).

Throughout history, dental aesthetics has held significant importance. Archaeological findings indicate that individuals have attempted to correct dental irregularities since early civilizations (6). Several indices and analytical tools have been developed to evaluate dentofacial aesthetics, including the Dental Aesthetic Index, Smile Index, Modified Smile Index, visual analog scale, and Q-sort analysis (7). However, the dental health component (DHC) and aesthetic component (AC) of the Index of Orthodontic Treatment Need (IOTN) were specifically designed for orthodontic assessment (8).

In orthodontic practice, standardized indices with established reliability and validity are used to assess treatment need. The IOTN provides a structured numerical scoring system that quantifies the degree of deviation from ideal occlusion and is internationally accepted for evaluating orthodontic treatment requirements (9,10).

In recent years, rapid advancements in digital health technologies and artificial intelligence (AI) systems have significantly influenced the field of dentistry (11,12). AI systems support clinical workflows through image analysis, pattern recognition, and decision-support mechanisms (13). In orthodontics, three-dimensional modeling and predictive simulations enable the anticipation of tooth movements prior to treatment initiation, and automated systems contribute to aesthetic evaluation processes (14).

Large language models (LLMs), such as ChatGPT, Gemini, and Claude, were originally developed for natural language processing tasks. However, recent multimodal versions have incorporated visual processing capabilities in addition to text-based reasoning (13). Unlike conventional convolutional neural network-based image classification systems that are specifically trained for diagnostic imaging tasks, multimodal LLMs integrate visual encoders with large-scale reasoning architectures. Although previous AI studies in orthodontics have primarily relied on task-specific, trained image analysis models, the potential of widely accessible, general-purpose multimodal LLMs to perform structured aesthetic evaluations based on standardized clinical indices remains insufficiently explored.

For AI systems to be considered reliable clinical decision-support tools, it is essential to assess not only their agreement with human evaluators but also the consistency of their outputs across different model versions. Model updates may influence evaluation behavior, thereby affecting clinical predictability and stability.

Therefore, the present study aimed to evaluate and compare the levels of aesthetic discomfort perceived by orthodontists, general dentists, and multimodal LLM-based AI systems using the IOTN-AC index. Additionally, this study investigated inter-group consistency and examined intra-model stability by comparing the aesthetic scoring performance of earlier and updated versions of the selected LLM systems.

Methods

This study was conducted following approval from the Institutional Ethics Committee of İstanbul Aydın University (approval date: May 22, 2026, decision no: 193/2026). Given the novel nature of evaluating multimodal LLMs using the IOTN-AC index, formal a priori sample size calculation was not performed, and this research was designed as an exploratory study. Standardized frontal intraoral photographs obtained from 30 patients were included in this study. To minimize selection bias, the 30 patient cases were selected from the institutional database using a non-consecutive, simple random sampling method. The inclusion criteria strictly comprised: (1) high-quality and clear frontal intraoral photographs, (2) patients with no history of previous orthodontic treatment, and (3) the absence of any congenital or craniofacial syndromes. Conversely, cases presenting with poor image resolution, inadequate lighting, or significant framing errors were excluded from the study. Demographic data (e.g., age and gender) of the patients were not recorded, as the photographs were completely de-identified and the study focused exclusively on the objective morphological assessment of intraoral visual features. According to the reference standard evaluations, the IOTN-AC distribution of the selected cases was as follows: 19 mild (IOTN-AC 1-4), 7 moderate (IOTN-AC 5-7), and 4 severe (IOTN-AC 8-10) cases.The images were evaluated by 50 orthodontic specialists, 50 general dentists, and selected multimodal AI models. Each human evaluator assessed the same set of photographs in a single, independent session to determine the degree of aesthetic impairment. To evaluate the collective performance of the general dentist group, the pooling method was conducted by calculating the arithmetic mean of the 50 individual scores assigned to each specific photograph. This resulting mean value was then adopted as the representative final score of the general dentist group for that respective case.

Prior to evaluation, orthodontists and general dentists were provided with the internationally accepted IOTN-AC reference chart and were instructed to perform all assessments strictly according to this standardized scale.

The IOTN, developed by Brook and Shaw (9), is an objective method for assessing orthodontic treatment requirements. It evaluates malocclusion severity, degree, and aesthetic appearance (9,10). Following modifications recommended by the Swedish Medical Council, the index consists of two components: the DHC and the AC (15). The AC is based on a 10-point photographic scale representing varying levels of dental attractiveness. A score of 1 indicates little or no perceived treatment need, whereas a score of 10 reflects a definite need for orthodontic intervention (16,17). The DHC evaluates occlusal characteristics, emphasizing the most severe clinically significant trait when determining overall treatment need (9).

For AI evaluation, only multimodal LLM versions with integrated image interpretation capabilities were included. The AI evaluations were conducted following ethics committee approval, with all access and diagnostic assessments completed on May 25, 2026. The evaluations were performed using the premium web-based interfaces of the respective providers, without application programming interface integration. The specific models utilized were ChatGPT-4o and ChatGPT-5.2 (OpenAI; accessed via ChatGPT Plus subscription), Claude 3.5 Sonnet and Claude 4.5 Sonnet (Anthropic; accessed via Claude Pro subscription), and Gemini 1.5 Pro and Gemini 3 Pro (Google; accessed via Gemini Advanced subscription). Prior to assessment, the models were queried regarding their familiarity with the IOTN-AC index to ensure conceptual understanding. No task-specific retraining, fine-tuning, or dataset optimization was performed. This zero-shot evaluation approach was intentionally adopted to reflect real-world conditions of publicly accessible, general-purpose AI systems rather than specialized orthodontic image-analysis models. Both earlier and updated versions of the selected AI systems were evaluated independently. Each AI version assessed the identical photograph set only once in a single, independent run using the same IOTN-AC reference scale, without multiple iterations, prompt modifications, or regeneration of responses. Version-based comparisons were conducted to analyze the impact of model updates on aesthetic scoring behavior and to examine intra-model stability.

Orthodontists and general dentists were presented with standardized frontal intraoral photographs and were asked to identify the reference image on the IOTN-AC scale that most closely corresponded to each case. The same images were provided to the AI models, which were instructed to assign aesthetic scores according to the IOTN-AC index.

All selected photographs were high-resolution, clear, and standardized images. Data were organized in tabular format and presented as frequencies and percentages. Participation of clinicians was voluntary and based on informed consent.

Statistical Analysis

Statistical analyses were performed using Python-based data analysis, utilizing standard scientific computing libraries including Pandas, NumPy, SciPy, and Scikit-learn, with the significance level set at $p<0.05$ for all tests. The normality of data distribution was assessed using the Shapiro-Wilk test. The mean IOTN-AC scores obtained from 50 orthodontists served as the reference standard. Agreement between the evaluator groups and the reference standard was analyzed using Spearman's rank correlation coefficient (ρ), Lin's concordance correlation coefficient (CCC), mean absolute error (MAE), and Bias (systematic deviation) metrics. Performance differences between AI model versions were compared using the Wilcoxon signed-rank test. Diagnostic performance metrics [sensitivity, specificity, positive predictive value (PPV), negative predictive value, and accuracy] were calculated based on the clinical treatment need threshold of IOTN-AC ≥4.0. Limits of agreement between methods were evaluated using Bland-Altman analysis. Additionally, model performance was examined across case difficulty subgroups, stratified into "clear cases" [high consensus, standard deviation (SD) ≤0.508] and "ambiguous cases" (low consensus, SD >0.508).

Results

The comparative performance metrics of general dentists and LLM-based AI systems, based on the mean IOTN-AC scores accepted as the reference standard from orthodontists, are presented in Table 1. According to the analysis results, general dentists exhibited the highest level of agreement with the orthodontists. This group demonstrated the lowest (MAE: 0.761) and the highest Lin's (CCC: 0.901; 95% confidence interval: 0.796-0.948), indicating the absolute scoring closest to the reference standard in aesthetic assessment. In contrast, the absolute agreement of LLMs with orthodontist scores was found to be lower. Although the Spearman rank correlation coefficients (ρ) of all models were calculated to be statistically significant ($p<0.05$), positive bias (systematic deviation) values indicating that aesthetic impairment was overestimated were detected in all models. Notably, in the Gemini 3 Pro model, this deviation exceeded +3 points, measured at a level that could potentially alter patients' IOTN treatment need classification.

The performance differences between older and newer versions of the same AI models were evaluated using the Wilcoxon signed-rank test (Table 2). The analysis revealed that in the newer versions of the ChatGPT and Gemini families (ChatGPT-5.2 and Gemini 3 Pro), Spearman correlation coefficients increased, indicating an improvement in ranking performance; however, in terms of absolute accuracy (MAE), error rates increased in the new iterations of all three model families. Specifically, Gemini 3 Pro showed a statistically significant increase in MAE compared to its predecessor, Gemini 1.5 Pro (MAE increased from 2.529 to 3.406; $p=0.023$). The error rate also increased in the ChatGPT-5.2 model compared to ChatGPT-4o (+0.296 points), but this increase did not reach statistical significance ($p=0.422$). The most pronounced performance loss between versions was detected in the Anthropic family. The newer model, Claude 4.5 Sonnet, exhibited a marked increase in error (+0.513 points; $p=0.059$) compared to Claude 3.5 Sonnet. Furthermore,

Table 1. Agreement of general dentists and large language models with orthodontists' mean IOTN-AC scores

Rater group	Lin's CCC (95% CI)	Spearman's ρ (p-value)	MAE (95% CI)	Bias (95% CI)
General dentists (pooled)	0.901 (0.796 to 0.948)	0.949 (<0.001)	0.761 (0.589 to 0.965)	+0.659 (0.437 to 0.898)
ChatGPT-4o	0.574 (0.243 to 0.786)	0.587 (<0.001)	1.415 (0.929 to 2.043)	+0.761 (0.123 to 1.473)
ChatGPT-5.2	0.502 (0.292 to 0.677)	0.731 (<0.001)	1.711 (1.309 to 2.159)	+1.494 (0.970 to 2.045)
Claude 3.5 Sonnet	0.528 (0.255 to 0.719)	0.584 (<0.001)	1.861 (1.417 to 2.389)	+1.361 (0.733 to 2.019)
Claude 4.5 Sonnet	0.351 (0.095 to 0.565)	0.377 (0.040)	2.374 (1.856 to 2.932)	+1.394 (0.481 to 2.245)
Gemini 1.5 Pro	0.329 (0.127 to 0.519)	0.587 (<0.001)	2.529 (2.001 to 3.190)	+2.361 (1.744 to 3.091)
Gemini 3 Pro	0.214 (0.079 to 0.382)	0.600 (<0.001)	3.406 (2.752 to 4.114)	+3.361 (2.655 to 4.087)

Note: MAE indicates mean absolute error relative to orthodontists' mean scores. Bias represents the mean signed difference, with positive values indicating systematic overestimation of aesthetic severity. Spearman's ρ reflects rank-order agreement, while Lin's concordance correlation coefficient (CCC) evaluates both precision and absolute agreement
IOTN: Index of Orthodontic Treatment Need, AC: Aesthetic component, MAE: Mean absolute error, CI: Confidence interval

Table 2. Comparison of model versions: analysis of performance change (Wilcoxon signed-rank test)

Model comparison (old → new)	MAE (old)	MAE (new)	Change in MAE	Result	p-value*
ChatGPT-4o → ChatGPT-5.2	1.415	1.711	+0.296	Degradation	0.422
Claude 3.5 Sonnet → Claude 4.5 Sonnet	1.861	2.374	+0.513	Degradation	0.059
Gemini 1.5 Pro → Gemini 3 Pro	2.529	3.406	+0.877	Significant degradation	0.023

Note: MAE: Mean absolute error (lower is better). Change in MAE: Positive values indicate an increase in error (worsening performance)
*: Statistically significant difference ($p<0.05$), MAE: Mean absolute error

Claude 4.5 Sonnet demonstrated a weaker agreement with the reference standard, showing a higher MAE (2.37 vs. 1.86), a lower Spearman correlation coefficient (0.38 vs. 0.58), and a decreased Lin's CCC (0.35 vs. 0.53) with wider confidence intervals.

Bland-Altman analysis demonstrated differences in individual reliability among the groups (Figure 1). General dentists exhibited high individual agreement with the reference standard, showing low systematic error (Bias: +0.659) and narrow limits of agreement (-0.64 to +1.96). In all AI models, a systematic tendency toward "overestimation" (positive bias) was detected, and this deviation reached a level of +3.36 in the Gemini 3 Pro model. Additionally, the limits of agreement for AI models were found to be wider compared to human evaluators. For example, the wide

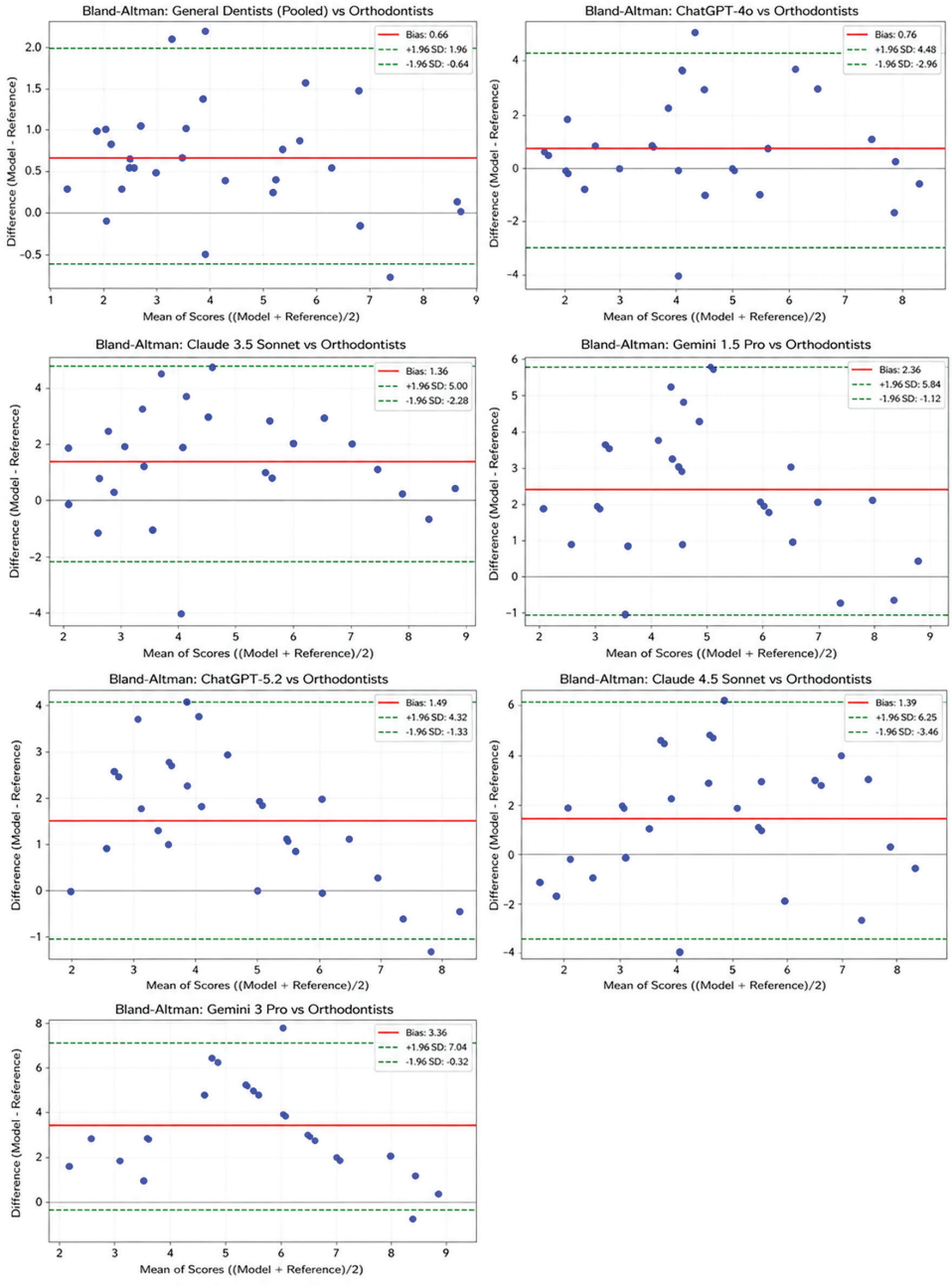

Figure 1. Bland-Altman plots evaluating the agreement between the raters (general dentists and AI models) and the reference standard (orthodontists). The solid red line represents the mean bias (systematic difference between the rater and the reference standard). The upper and lower dashed green lines indicate the 95% limits of agreement, calculated as mean bias ±1.96 SDs. The x-axis represents the mean of the scores given by the specific rater and the reference standard [(model + reference) / 2), while the y-axis shows the absolute difference between them (model - reference)]

SD: Standard deviation, AI: Artificial intelligence

range calculated for ChatGPT-4o (-2.95 to +4.47) indicates a fluctuation of approximately 7 points in individual cases. In the graphical distributions in Figure 1, proportional bias was observed, where the magnitude of error systematically changed with case severity.

The results of the bias analysis stratified by case severity indicated that the error characteristics of the AI models varied according to case categories (Table 3). While general dentists exhibited a positive bias of +0.78 and +0.71 in mild (IOTN 1-4) and moderate (IOTN 5-7) severity cases respectively, this value was calculated as -0.21 in severe cases (IOTN 8-10). In the AI models, an inverse relationship was detected between case severity and error magnitude. All models showed a distinct "overestimation" tendency in mild cases; this deviation reached +4.16 points in the Gemini 3 Pro model and +2.95 points in the Gemini 1.5 Pro model. In severe cases, the bias values of the same models decreased significantly, approaching zero or turning negative (e.g., Claude 3.5 Sonnet: -0.01; ChatGPT-4o: -0.67).

Model performances were evaluated under varying levels of clinical consensus (expert agreement), and the dataset was stratified into "clear cases" (high expert consensus, SD ≤0.508, n=15) and "ambiguous cases" (low expert consensus, SD >0.508, n=15) (Table 4). According to the analysis results, general dentists exhibited low error rates in both categories; the MAE was 0.976 in clear cases and 0.547 in ambiguous cases. In the AI models, while higher error rates and positive bias (e.g., Gemini 3 Pro: +3.911; ChatGPT-5.2: +2.244) were detected in clear cases where experts had a strong consensus, error margins were observed to decrease in ambiguous cases where expert consensus was lower. For instance, while ChatGPT-4o had a bias of +1.311 in the clear cases group, this value was measured at +0.211 in the ambiguous cases group, which was lower than the bias value of general dentists in the same group (+0.355). This trend was consistent across other AI models, and MAE values were systematically calculated to be lower in the ambiguous cases category (e.g., ChatGPT-5.2 MAE: 2.263 in clear cases vs. 1.160 in ambiguous cases).

An IOTN-AC ≥4.0 threshold was utilized to determine treatment need in the diagnostic performance evaluation (Table 5). The analysis revealed that general dentists demonstrated 80.0% accuracy and 70.6% specificity. The sensitivity rates of the AI models were measured between 84.6% and 100%, with the ChatGPT-5.2 and Gemini 3 Pro models achieving 100% sensitivity. Conversely, the specificity values of the models were found to be low; this

Table 3. Systematic bias stratified by case severity (IOTN-AC categories)

Rater group	Mild cases[a]	Moderate cases[b]	Severe cases[c]
General dentists (pooled)	+0.780	+0.710	-0.210
ChatGPT-4o	+1.370	-0.160	-0.670
Claude 3.5 Sonnet	+1.740	+0.970	-0.010
ChatGPT-5.2	+2.210	+0.720	-1.010
Claude 4.5 Sonnet	+1.900	+1.090	-1.010
Gemini 1.5 Pro	+2.950	+1.970	-0.340
Gemini 3 Pro	+4.160	+2.590	+0.330

Note: Values represent mean bias, calculated as the rater score minus the orthodontist score, where positive values indicate overestimation of aesthetic severity and negative values indicate underestimation
[a]: Mild cases: Reference standard mean score <4.5 (n=19), [b]: Moderate cases: Reference standard mean score ≥4.5 and ≤7.5 (n=7), [c]: Severe cases: reference standard mean score >7.5 (n=4), IOTN: Index of Orthodontic Treatment Need, AC: Aesthetic component

Table 4. Comparison of performance metrics in "clear cases" (low observer variability) versus "ambiguous cases" (high observer variability)

Rater group	Clear cases (low SD) MAE	Clear cases bias	Ambiguous cases (high SD) MAE	Ambiguous cases bias
General dentists	0.976	+0.963	0.547	+0.355
ChatGPT-4o	1.567	+1.311	1.264	+0.211
ChatGPT-5.2	2.263	+2.244	1.160	+0.744
Claude 3.5 Sonnet	1.897	+1.577	1.824	+1.144
Claude 4.5 Sonnet	2.260	+1.844	2.488	+0.944
Gemini 1.5 Pro	2.855	+2.711	2.203	+2.011
Gemini 3 Pro	3.911	+3.911	2.901	+2.811

Note: Clear cases (high consensus) are defined as images where the standard deviation of orthodontist scores was less than or equal to 0.508 (n=15), indicating strong agreement among experts. Ambiguous cases (low consensus) are defined as images where the standard deviation exceeded 0.508 (n=15), indicating disagreement or clinical ambiguity
MAE: Mean absolute error (lower is better), Bias: Mean difference (positive values indicate overestimation), SD: Standard deviation

rate was calculated as 5.9% for Gemini 3 Pro and 11.8% for Gemini 1.5 Pro. The PPVs of the AI models ranged from 44.4% to 63.2%.

Discussion

This study is among the first investigations directly compares orthodontists, general dentists, and LLM-based AI systems in aesthetic assessment based on the IOTN-AC index. While traditional orthodontic diagnosis relies heavily on clinical experience and professional judgment, AI is increasingly being integrated into diagnosis and treatment processes by analyzing large datasets. However, the IOTN-AC index is not solely based on measurable morphological parameters; it relies on a subjective assessment shaped by aesthetic sensibility and the interpretation of visual stimuli (18). By exhibiting the highest agreement with orthodontists, general dentists have supported the central role of clinical training and experience-based judgment in aesthetic assessment. This result suggests that aesthetic perception is likely not merely dependent on visual pattern recognition, but on an integrated clinical understanding that remains uniquely human. In this context, the role of LLMs in such clinical tasks that require human perception rather than objective measurement remains a subject of significant debate (18,19).

In our study, LLMs were conceptualized not as technical tools intentionally designed to classify images objectively or simply as "seeing" systems, but as multimodal systems that shape human aesthetic perception, interpreting, describing, and categorizing visual information in a human-like manner. From this perspective, the primary objective of the study was not to test image recognition accuracy, but to explore whether LLMs could meaningfully represent aesthetic severity within a human-centered conceptual framework (19).

The wide limits of agreement and distinct proportional error observed in LLMs suggest significant limitations regarding the clinical reliability of these models. In particular, the marked increase in the margin of error as case severity decreases suggests that the models experience serious difficulties in distinguishing healthy cases with low visual complexity in the present dataset. This pattern indicates that although LLMs can approximate certain aspects of human perceptual reasoning, they lack the calibration required for reliable absolute clinical scoring. These deviations, which can exceed several points in some cases, have the potential to significantly alter IOTN-based clinical classification. Therefore, it is not sufficient for current LLMs to merely demonstrate average accuracy; the high variability at the individual case level limits the use of these systems as independent clinical decision-makers at the current stage.

The most striking finding of our study, which contradicts general expectations in the literature, is the systematic regression observed in IOTN index scoring of new generation LLMs compared to their predecessors. In particular, the statistically significant performance loss observed in the Gemini 3 Pro model (p=0.023), along with similar deterioration trends in other models, suggests that advances in AI development do not inherently and always equate to "clinical improvement". This finding shows that increased model complexity does not guarantee enhanced performance in perception-based medical tasks and serves as a critical warning for medical AI integration. It is becoming a necessity that LLM-based tools designed for healthcare systems undergo *de novo* validation tests after every software update, just like the validation of an entirely new medical device. This paradoxical result can be attributed to the "reinforcement learning from human feedback" and "safety alignment" strategies, which are increasingly prioritized in the training processes of current LLMs (20,21).

Another fundamental weakness of LLMs in our study was their low specificity. The fact that the models demonstrated a sensitivity approaching that of experts in detecting severe malocclusions (IOTN 8-10) suggests that AI has the capacity to recognize distinct anatomical deviations such as severe crowding or overjet. Although sensitivity rates reaching 100% indicate that the models do not miss cases requiring treatment, this success is largely overshadowed by specificity rates dropping as low as 5.9%. This situation may

Table 5. Diagnostic accuracy metrics for orthodontic treatment need (threshold: IOTN-AC ≥4.0)

Rater group	Sensitivity	Specificity	PPV	NPV	Accuracy
General dentists	0.923	0.706	0.706	0.923	0.800
ChatGPT-4o	0.923	0.588	0.632	0.909	0.733
ChatGPT-5.2	1.000	0.176	0.481	1.000	0.533
Claude 3.5 Sonnet	0.846	0.412	0.524	0.778	0.600
Claude 4.5 Sonnet	0.923	0.412	0.545	0.875	0.633
Gemini 1.5 Pro	0.923	0.118	0.444	0.667	0.467
Gemini 3 Pro	1.000	0.059	0.448	1.000	0.467

Note: Threshold for treatment need is defined as an IOTN-AC score of 4.0 or higher
Sensitivity: Ability to identify cases requiring treatment, Specificity: Ability to identify healthy cases correctly, PPV: Positive predictive value, NPV: Negative predictive value, IOTN: Index of Orthodontic Treatment Need, AC: Aesthetic component

indicate that AI acts more like an overly sensitive and flawed alarm system rather than an objective diagnostic tool. The systematic and over-scoring tendency observed in mild cases (IOTN 1-4) indicates that these systems currently lack the calibration needed to distinguish the fine line between "clinical norm" and "pathology". This tendency towards "over-diagnosis" can lead to a significant burden resulting from unnecessary specialist referrals in clinical practice. Considering the superior level of specificity exhibited by general dentists, LLMs should be utilized as auxiliary tools under professional supervision solely to benefit from their high sensitivity. While the models erroneously labeling mild cases as "severe" creates an unnecessary referral burden on the system and unwarranted anxiety for the patient, the success of general dentists suggests that human common sense and the principle of *primum non nocere* (first, do no harm) cannot yet be replicated by algorithms. The safety-oriented approach of the developers appears to have created a strict error-avoidance behavior in the models; this may have replaced technical accuracy with an "overprotective" and artificial score inflation that perceives even the slightest deviations as a risk.

While LLMs failed in low-variation "clear cases" where orthodontists reached full consensus, they exhibited a performance approaching, and in some cases even exceeding, that of general dentists in high-variation "ambiguous cases" characterized by expert disagreements. The natural indecision among human observers in "ambiguous cases", which generally represent individuals with borderline treatment needs, can cause AI predictions to be recorded statistically as "different but valid opinions" rather than "errors". Furthermore, in the face of the density of "ideal aesthetic" and "severe anomaly" images in LLM training datasets, the more limited data on "borderline" cases may have led the models to manage this ambiguity through an averaging method known as the "regression effect"; this situation suggests that AI's strategy for managing clinical uncertainty is fundamentally different from that of humans, exhibiting a "generalizing" or averaging behavior where humans adopt a "conservative" approach in the face of uncertainty.

When compared with current studies in the literature, our findings add a deeper and more critical dimension to the ongoing debates regarding the role of AI in orthodontic diagnosis. While Kazimierczak et al. (22) paint a general picture of optimism by highlighting the potential of AI to revolutionize orthodontic diagnosis and treatment planning, our study reveals with concrete data that this "revolution" is not yet complete, especially when it comes to subjective aesthetic perception.

In particular, when examining the relationship between model performance and version updates, our findings form a striking contrast with the results of Kılınç and Mansız (23). While Kılınç and Mansız (23) reported that the reliability and readability of orthodontic information increased as text-based ChatGPT versions were updated, our study found that new generation models focusing on visual analysis capabilities (e.g., Gemini 3 Pro) experienced a statistically significant performance loss compared to their predecessors. This situation suggests that text-based improvements do not directly translate to visual aesthetic judgment; on the contrary, factors such as "safety alignment" can lead to a paradoxical regression in visual analysis.

Similarly, Perez-Pino et al. (24) stated that virtual assistants exhibited inconsistencies in routine questions and required professional validation; our study supports this view by indicating that a potential reason for this inconsistency is the extreme sensitivity and systematic "over-diagnosis" tendency developed by the models due to the strict safety constraints.

In addition, Stetzel's (25) thesis study on IOTN-AC estimation pointed out the difficulties in automating this index. Our study, on the other hand, suggested that this difficulty might stems not only from a technical inadequacy but also from the models' potential inability to fully imitate the "contextual flexibility" specific to human perception (as we see in general dentists).

Study Limitations

This study has certain limitations. First, given the novel and exploratory nature of evaluating multimodal LLMs using the IOTN-AC index, a formal a priori sample size calculation (power analysis) was not performed. Consequently, the relatively small sample size limits the broad generalizability of the findings, and the present research should be considered primarily as an exploratory or pilot investigation. Second, aesthetic assessments were conducted solely on standardized two-dimensional frontal intraoral photographs obtained from the 30 included patients, and the DHC of the IOTN index, which evaluates overall treatment need, was not included in the analysis. Third, when testing the AI models, no task-specific retraining, fine-tuning, or dataset optimization was performed; instead, a "zero-shot" evaluation approach was intentionally adopted to reflect the real-world conditions of publicly accessible systems. While this methodological choice reflects the general-purpose capabilities of the models, it does not represent the potential performance of algorithms specifically optimized for orthodontic image analysis.

Conclusion

In conclusion, within the specific limitations of this study's sample and pooled evaluation design although current LLM systems show promise in identifying pathology (sensitivity), their weakness in distinguishing healthy cases (specificity) suggests that they may still be inadequate at fully replicating the depth and contextual sensitivity of human aesthetic judgment. This situation limits the use of these models as autonomous screening tools; rather, it supports

their positioning as preliminary assessment adjuncts that require clinician validation. Human expertise remains the ultimate reference standard in aesthetic orthodontic assessment; therefore, the integration of LLMs into clinical workflows should be approached with caution, carefully considering these inherent structural limitations.

Ethics

Ethics Committee Approval: This study was conducted following approval from the Institutional Ethics Committee of İstanbul Aydın University (approval date: May 22, 2026, decision no: 193/2026).

Informed Consent: Participation of clinicians was voluntary and based on informed consent.

Footnotes

Authorship Contributions

Surgical and Medical Practices: R.C.Ö., A.P.B., Concept: R.C.Ö., A.P.B., Design: R.C.Ö., A.P.B., Data Collection or Processing: R.C.Ö., Analysis or Interpretation: R.C.Ö., A.P.B., Literature Search: R.C.Ö., Writing: R.C.Ö., A.P.B.

Conflict of Interest: No conflict of interest was declared by the authors.

Financial Disclosure: The authors declared that this study received no financial support.

References

1. Sureshbabu A, Chandu G, Shafiulla M. Prevalence of malocclusion and orthodontic treatment needs among 13-15-year-old schoolchildren of Davangere city, Karnataka. J Indian Assoc Public Health Dent. 2005;5:32.
2. Buschang PH, Shulman JD. Incisor crowding in untreated persons 15-50 years of age: United States, 1988-1994. Angle Orthod. 2003;73:502-8.
3. Sam G. Orthodontics as a prospective career choice among undergraduate dental students: a prospective study. J Int Soc Prev Community Dent. 2015;5:290-5.
4. Dewey M. Who is an "orthodontist"? 1915. Am J Orthod Dentofacial Orthop. 2015;147:523-4.
5. Paquette JM, Sheets CG. The second "DDS" degree: a formula for practice success. J Am Dent Assoc. 2004;135:1321-5.
6. Wahl N. Orthodontics in 3 millennia. Chapter 1: antiquity to the mid-19th century. Am J Orthod Dentofacial Orthop. 2005;127:255-9.
7. American Dental Association, National Commission on Recognition of Dental Specialties and Certifying Boards. Recognized dental specialties. American Dental Association. Published 2018. Accessed June 1 [cited September 2, 2026], 2025. https://www.ada.org
8. Evans R, Shaw W. Preliminary evaluation of an illustrated scale for rating dental attractiveness. Eur J Orthod. 1987;9:314-8.
9. Brook PH, Shaw WC. The development of an index of orthodontic treatment priority. Eur J Orthod. 1989;11:309-20.
10. Shaw WC, Richmond S, O'Brien KD. The use of occlusal indices: a European perspective. Am J Orthod Dentofacial Orthop. 1995;107:1-10.
11. Hwang JJ, Jung YH, Cho BH, Heo MS. An overview of deep learning in the field of dentistry. Imaging Sci Dent. 2019;49:1-7.
12. Sener E, Baksi Sen G. Role of artificial intelligence applications in dentomaxillofacial radiology: part 1 [Dentomaksillofasiyal radyolojide yapay zeka uygulamalarının rolü: bölüm 1]. Selcuk Dent J. 2022;9:713-20. Turkish.
13. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023;29:1930-40.
14. Richmond S, Shaw WC, O'Brien KD, Buchanan IB, Jones R, Stephens CD, et al. The development of the PAR Index (Peer Assessment Rating): reliability and validity. Eur J Orthod. 1992;14:125-39.
15. 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;13:735.
16. Richmond S, Aylott NA, Panahei ME, Rolfe B, Tausche E. A two-center comparison of orthodontists' perceptions of orthodontic treatment difficulty. Angle Orthod. 2001;71:404-10.
17. de Almeida AB, Leite IC. Orthodontic treatment need for Brazilian schoolchildren: a study using the Dental Aesthetic Index. Dental Press J Orthod. 2013;18:103-9.
18. Wang L, Li J, Zhuang B, Huang S, Fang M, Wang C, et al. Accuracy of large language models when answering clinical research questions: systematic review and network meta-analysis. J Med Internet Res. 2025;27:e64486.
19. Makrygiannakis MA, Giannakopoulos K, Kaklamanos EG. Evidence-based potential of generative artificial intelligence large language models in orthodontics: a comparative study of ChatGPT, Google Bard, and Microsoft Bing. Eur J Orthod. 2025;48:cjae017.
20. Casper S, Davies X, Shi C, Gilbert TK, Scheurer J, Rando J, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. Preprint. arXiv. 2023.
21. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright CL, Mishkin P, et al. Training language models to follow instructions with human feedback. Preprint. arXiv. 2022.
22. Kazimierczak N, Kazimierczak W, Serafin Z, Nowicki P, Nożewski J, Janiszewska-Olszowska J. AI in orthodontics: revolutionizing diagnostics and treatment planning-a comprehensive review. J Clin Med. 2024;13:344.
23. Kılınç DD, Mansız D. Examination of the reliability and readability of chatbot generative pretrained transformer's (ChatGPT) responses to questions about orthodontics and the evolution of these responses in an updated version. Am J Orthod Dentofacial Orthop. 2024;165:546-55.
24. Perez-Pino A, Yadav S, Upadhyay M, Cardarelli L, Tadinada A. The accuracy of artificial intelligence-based virtual assistants in responding to routinely asked questions about orthodontics. Angle Orthod. 2023;93:427-32.
25. Stetzel L. Artificial intelligence (AI) for predicting the aesthetic component (AC) of the Index of Orthodontic Treatment Need (IOTN) (dissertation). Columbus, Ohio: The Ohio State University. 2023.