diagnostics

MDPI

Article

Evaluating Large Language Models in Interdisciplinary Orthodontic–Periodontal Clinical Decision-Making: A Synthetic Case-Based Comparison with Expert Decisions

Zeliha Başak Çakır Erdil [1], Seda Sağoğlu [2,*], Hüseyin Akıllı [1], Emine Şişman [2], Şahin Altuğ [1] and Gül Demir [2]

[1] Department of Periodontology, Faculty of Dentistry, Adıyaman University, 02230 Adıyaman, Türkiye; bskcakirr@gmail.com (Z.B.Ç.E.)
[2] Department of Orthodontics, Faculty of Dentistry, Necmettin Erbakan University, 42090 Konya, Türkiye
* Correspondence: ssagoglu@gmail.com

Abstract

Background/Objectives: This study evaluated agreement between structured decisions generated by large language models (LLMs) and expert-based clinical reference decisions in interdisciplinary orthodontic–periodontal decision-making and whether agreement varied across clinical decision domains. **Methods**: Fifty synthetic multimodal clinical cases integrating orthodontic and periodontal variables were developed. After excluding two cases used for pilot testing and calibration, 48 cases were independently evaluated by three orthodontists and three periodontists to derive expert-based reference decisions. The same cases were evaluated by ChatGPT-5.6, Claude Opus 4.8 High, and Gemini Pro using standardized specialty-specific prompts. LLM–expert agreement was assessed using exact agreement rates and Cohen's κ, and inter-expert reliability was assessed using Krippendorff's α. **Results**: Inter-expert reliability and LLM–expert agreement varied across decision domains. For shared items, exact agreement ranged from 73.5–91.2% for periodontal readiness, 66.7–87.2% for periodontal risk, and 51.4–75.7% for modification of the orthodontic approach. For specialty-specific decisions, agreement was 87.5–89.6% ($\kappa = 0.750$–0.792) for orthodontic treatment-plan modification, 78.7–83.3% ($\kappa = 0.739$–0.798) for treatment timing, 64.6–75.0% ($\kappa = 0.516$–0.552) for periodontal disease control, and 69.6–84.8% ($\kappa = 0.289$–0.664) for mucogingival surgery. Models also differed when selecting among multiple interventions. **Conclusions**: LLM–expert agreement varied across stages of interdisciplinary decision-making, suggesting that domain-specific assessment may be more informative than a single overall metric. Given the synthetic, standardized design, these findings should not be interpreted as evidence of independent diagnostic or treatment-planning performance in real-world clinical settings.

Keywords: artificial intelligence; clinical decision-making; expert agreement; large language models; orthodontics; periodontology

Check for updates

Academic Editor: Giuseppina Malcangi

Received: 30 September 2026
Revised: 16 November 2026
Accepted: 20 November 2026
Published: 27 November 2026

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

1. Introduction

Large language models (LLMs) are increasingly being used across a range of healthcare applications, including clinical decision support, patient education, and healthcare management, while their use in diagnostic and interpretive applications in dentistry continues to expand [1–4]. However, the variability in diagnostic accuracy and consistency across different clinical tasks underscores the need to evaluate not only the capacity of these models to provide information, but also their ability to integrate complex clinical

findings and generate clinically appropriate decisions [5]. Previous studies have reported that ChatGPT may have difficulty recognizing certain atypical conditions in patients with complex medical or surgical histories and selecting the most appropriate option among alternative additional investigations [6].

Because orthodontic diagnosis requires the integrated interpretation of clinical examination findings, cephalometric assessments, and other patient records, conventional diagnostic evaluations can be time-consuming and may be subject to inter-clinician variability. Consequently, multimodal artificial intelligence (AI) approaches capable of integrating different types of clinical and imaging data are increasingly being investigated for orthodontic diagnosis and clinical decision support [7]. Although LLMs may provide satisfactory responses to general knowledge questions in orthodontics, this performance cannot necessarily be extrapolated to more complex and individualized treatment-planning tasks, and general-purpose LLMs have been reported to exhibit uncertainty when applied to clinical cases [8]. Accordingly, the clinical applicability of LLMs should be evaluated not solely on the basis of factual accuracy, but also through treatment decisions involving patient-specific characteristics and multiple interacting risk factors.

Orthodontic treatment planning in periodontally compromised patients should be based not only on identifying existing periodontal disease, but also on the integrated assessment of the patient's primary malocclusion and any secondary malocclusion resulting from periodontal tissue loss. The combined consideration of periodontal and orthodontic diagnoses represents a fundamental step in establishing functional and esthetic treatment goals [9–11]. In this context, incorporating periodontal assessment, and when indicated, consultation with a periodontist, into treatment planning before initiating orthodontic therapy is particularly important [11,12]. Factors such as periodontal tissue phenotype, dimensions of keratinized tissue, existing alveolar bone support, tooth position within the alveolar bone, the direction and magnitude of the planned orthodontic movement, and the patient's level of oral hygiene should be considered collectively when anticipating the periodontal tissue response during orthodontic treatment [3,11,12]. Furthermore, in individuals who are susceptible to periodontal disease or are at high risk but have not yet been diagnosed, failure to adequately determine periodontal status before the application of orthodontic forces may be clinically relevant because of the potential for the progression of existing periodontal tissue loss. Interdisciplinary consultation between the orthodontist and periodontist is therefore important not only for the management of existing periodontal disease, but also for planning orthodontic treatment within periodontal boundaries and preserving periodontal tissues throughout treatment.

The possibility of limited agreement even among experienced periodontists further highlights the potential relevance of clinical decision-support systems [13]. In this context, rather than replacing expert periodontal assessment, AI-based systems could potentially support orthodontists in preliminary evaluation, particularly in clinical settings where immediate periodontal consultations are not readily available. Such support could facilitate the recognition of findings suggestive of periodontal risk, the identification of patients requiring further evaluation, and the determination of the need for periodontal consultation. Nevertheless, periodontology remains one of the comparatively underexplored areas of dental LLM applications, and existing studies have reported variable performance in determining periodontal stage, grade, and extent [13,14].

The potential role of LLMs in dentistry is expanding beyond general information generation toward prognostic and clinical decision-support tasks that require the integration of multiple clinical variables. Nevertheless, evidence regarding the extent to which LLM outputs align with expert clinicians' decisions in complex dental scenarios remains limited [15]. A recent systematic review highlighted that the existing literature has predom-

inantly compared performance across different LLMs, whereas direct comparisons using qualified clinicians as the reference remain relatively scarce [8]. It is therefore important to evaluate LLM performance not only in terms of general question-answering accuracy, but also through direct comparison with expert decisions in standardized clinical cases. Previous research has further emphasized the need to assess whether LLMs can translate knowledge derived from clinical consensus into decisions applicable to real-world clinical scenarios. This distinction is particularly relevant in interdisciplinary cases requiring the simultaneous interpretation of multiple clinical variables [16].

The aim of this study was to investigate the agreement between structured clinical decisions generated by three different LLMs and expert-based clinical reference decisions derived from independent assessments by specialists in the relevant disciplines, using standardized synthetic clinical scenarios requiring the integrated evaluation of orthodontic and periodontal factors. In addition, by assessing both inter-expert and interdisciplinary agreement, we sought to determine whether LLM performance varied according to the clinical decision domain and specialty perspective. To the best of our knowledge, this is the first study to simultaneously compare LLM-generated orthodontic and periodontal decisions for the same clinical cases with specialty-specific expert references.

2. Materials and Methods

2.1. Study Design and Ethical Considerations

This study was designed as a synthetic case-based comparative methodological study to evaluate the agreement between decisions generated by LLMs and those made by expert clinicians in clinical scenarios requiring the integrated assessment of orthodontic and periodontal considerations.

All clinical scenarios and accompanying visual materials used in the study were synthetically generated specifically for research purposes. No interactions with real patients were conducted during the study, and no clinical records, medical or dental history information, radiographs, clinical photographs, biological materials, personal health information, or other identifiable data from real patients were used. Furthermore, no diagnostic or therapeutic interventions were performed on any patient. Because the study was conducted exclusively using synthetic cases, involved neither real patients nor patient-derived data, and included no clinical interventions, no application for ethics committee approval for human-subject research was submitted.

2.2. Development of Synthetic Clinical Cases

At the initial stage of the study, a total of 50 synthetic clinical cases were developed. The number of cases was determined pragmatically, with the aim of ensuring sufficient representation of the range of clinical conditions and levels of decision-making complexity that may be encountered in multidisciplinary orthodontic–periodontal decision-making. No formal sample size calculation or statistical power analysis was performed before the study.

Orthodontic and periodontal specialists' clinical experience informed the development of the case pool. Two of the investigators involved in the case-development process (ZBÇE and SS) subsequently participated in the independent evaluation of cases within their respective specialties. Initially, the orthodontists identified plausible case presentations and clinical characteristics that may be encountered in routine clinical practice and in which periodontal status could influence orthodontic treatment decisions. The orthodontists constructed the distribution of malocclusions across the cases with consideration of the prevalence rates previously reported for the Turkish population [17]. However, the case pool was not intended to reproduce the prevalence of clinical conditions in a real-world patient population; rather, it was designed to provide sufficient representation of a diverse range of

clinical scenarios. The periodontists, in turn, defined the clinical diversity to be represented in the case pool with respect to periodontal disease status, loss of periodontal support, degree of inflammation, mucogingival conditions, and periodontal treatment needs.

The clinical components identified by the two specialties were subsequently integrated to develop synthetic cases requiring the concurrent assessment of orthodontic and periodontal variables within the same clinical scenario. The scenarios were reviewed by the investigators for clinical coherence, duplication of case characteristics, and diversity of multidisciplinary decision-making requirements. Where necessary, individual case components were revised to produce a final pool of 50 synthetic clinical scenarios encompassing varying levels of clinical severity, periodontal risk, and treatment needs (Supplementary Material S1). The investigators responsible for generating the synthetic images (ZBÇE and SS) reviewed them to assess their visual consistency with the clinical characteristics predefined for the corresponding case scenario. This process was performed as a quality-control step to identify evident discrepancies between the synthetic images and their corresponding case scenarios and was not intended to constitute independent image validation.

The first synthetic case was used as a pilot case to develop the assessment form, define the clinical scope of the individual items, and establish the response categories. The assessment form was finalized based on expert feedback, and this case was excluded from the final analyses.

Synthetic visual materials accompanying each clinical scenario were generated using the AI-assisted image-generation functionality of ChatGPT Images (OpenAI) to correspond to the clinical characteristics predefined for each case. The images were generated using prompts written in Turkish to visually represent the specified case characteristics. Accordingly, neither the case descriptions nor their accompanying images contained personal or identifiable information from any real patient. As the primary objective of the study was to evaluate clinical decision agreement between experts and AI models, the findings were assessed using the observed agreement measures together with their corresponding 95% confidence intervals.

2.3. Development of the Structured Clinical Decision Assessment Form

The assessment instrument used in this study was not an existing validated psychometric scale but rather a "Structured Case-Based Clinical Decision Assessment Form" specifically developed to address the research question (Supplementary Material S2).

The pilot case was used to guide the development of the form. Using this case, orthodontic and periodontal specialists discussed the key clinical decision domains that should be considered in a multidisciplinary assessment. The assessment items were developed based on expert consensus and clinical principles reported in the relevant literature.

The form consisted of two sections: (1) items completed by both orthodontists and periodontists to assess decision domains shared across the multidisciplinary decision-making process, and (2) specialty-specific items addressing decisions relevant to the clinical assessment and treatment-planning process within each specialty. This structure was designed to enable the evaluation of both shared decisions across specialties and specialty-specific assessments within the same clinical case.

2.3.1. Shared Clinical Decision Items

To enable orthodontists and periodontists to provide directly comparable decisions for the same case, four shared assessment items were developed based on the literature:

- S1—Periodontal readiness to initiate orthodontic treatment: This item was based on the role of periodontal inflammation control and periodontal stabilization in determining readiness to initiate orthodontic treatment [1,9,18–21].

- S2—Risk of adverse periodontal effects during orthodontic treatment: Periodontal, biological, and treatment-related factors that may influence the risk of periodontal deterioration were considered [3,9,21–23].
- S3—Need for modification of the orthodontic approach based on periodontal conditions: This item was based on the effects of periodontal support and tissue conditions on orthodontic tooth movement and the biomechanical approach [2,11,21,24–26].
- S4—Level of confidence in the clinical assessment: This item was included to separately assess the evaluator's confidence in the clinical decision-making process [21,27].

2.3.2. Orthodontic-Specific Items

Specialty-specific items were developed for completion by orthodontists to assess the influence of periodontal findings on orthodontic treatment planning:

- O1a—Need to modify the orthodontic treatment plan: This item was developed to determine whether the existing orthodontic treatment plan required modification because of periodontal findings [10,11,21,28].
- O1b—Nature of the required modification to the orthodontic treatment plan: This item was developed based on the literature addressing the effects of periodontal support loss on orthodontic treatment planning [11,19,28–30].
- O2—Timing and sequencing of orthodontic treatment: This item was developed in accordance with clinical principles concerning the appropriate timing of periodontal treatment and stabilization relative to orthodontic treatment and the sequencing of orthodontic therapy in relation to periodontal interventions [11,20,22,24,31].

2.3.3. Periodontal-Specific Items

Specialty-specific items were developed for completion by periodontists to assess periodontal requirements before orthodontic treatment:

- P1—Need for periodontal disease control: This item was based on clinical principles concerning the control of periodontal inflammation and disease, periodontal stabilization, and the need for periodontal treatment before orthodontic therapy [1,11,23,28].
- P2a—Need for mucogingival surgical intervention: This item was developed to determine whether surgical intervention was required based on the existing mucogingival findings and the planned orthodontic tooth movements [11,25,26,32,33]
- P2b—Type of mucogingival surgical intervention: For cases in which an intervention was considered necessary in P2a, this item was developed to assess the preferred surgical approach. The intervention options were defined based on the literature concerning mucogingival surgical approaches [25,26,32–36].
- P2c—Timing of mucogingival surgical intervention: For cases in which an intervention was considered necessary in P2a, this item was developed to assess the timing of the intervention relative to orthodontic treatment [11,25,37–39].

P2 was structured as a hierarchical clinical decision pathway comprising P2a–P2c. In the first step, P2a assessed the need for mucogingival surgical intervention; P2b and P2c were evaluated only when an intervention was considered necessary in P2a. P2b was defined as a multiple-response variable allowing more than one option to be selected simultaneously, whereas P2c was defined as a nominal categorical variable with no inherent ordering of severity. Each item in the form was treated as an independent clinical decision variable, and no overall or composite scale score was calculated by combining the individual items.

2.4. Evaluator Calibration

The second synthetic case was used for evaluator calibration. The shared items were jointly reviewed by the orthodontic and periodontal specialists, whereas the specialty-specific items were reviewed within the respective specialty groups to clarify the scope of each item and the interpretation of the response categories. Calibration was performed solely to standardize the assessment procedure; no sharing of responses or consensus process among experts was permitted for the subsequent cases.

The calibration case was excluded from the analysis. Following the exclusion of the pilot and calibration cases, the final analyses were conducted using the remaining 48 synthetic clinical cases.

2.5. Expert Assessments and Derivation of Expert-Based Reference Decisions

Three orthodontists and three periodontists independently evaluated the 48 synthetic clinical cases included in the analysis using the standardized assessment form. Both specialist groups completed the shared S1–S4 items. In addition, the orthodontists completed O1a, O1b, and O2, whereas the periodontists completed P1, P2a, P2b, and P2c. The evaluators had no access to one another's responses and conducted no subsequent reassessment or discussion to establish consensus for the cases included in the analysis.

P2 was structured as a hierarchical clinical decision pathway; P2b and P2c were evaluated only when mucogingival surgical intervention was considered necessary in P2a. P2b was treated as a multiple-response variable, whereas P2c was treated as a nominal categorical variable. Each item on the form was analyzed as a separate clinical decision variable, and no overall scale score was calculated.

For the expert reliability analysis, a periodontist's P2c response was considered applicable only when that same periodontist had indicated in P2a that mucogingival surgery was necessary (P2a = 1). Responses associated with P2a = 0 or P2a = 9 were treated as structurally not applicable and coded as missing. Five P2c entries recorded despite P2a ≠ 1 were excluded. Krippendorff's α for P2c was calculated using cases with applicable responses from at least two of the three periodontists. For the primary analysis of the shared items (S1–S4), a single multidisciplinary expert-based reference was derived from the independent responses of all six experts. A reference category was assigned when at least four of the six experts selected the same category. Cases that did not meet this consensus threshold were excluded from the primary AI–reference agreement analysis for the respective item. This resulted in item-specific analytical denominators of 34 for S1, 39 for S2, 37 for S3, and 32 for S4. The same multidisciplinary reference and the same eligible cases were used for both AI role sessions and all three models. For specialty-specific items, expert-based reference decisions were derived from the three experts in the corresponding specialty. For ordinal variables, the median of the three experts' ratings was used; for binary variables, the majority decision was used; and for nominal variables, the mode was used. For multiple-response items (O1b and P2b), each response option was treated as a separate binary variable, and a majority decision was determined for each option. When all three experts selected different categories for a nominal item, no unique mode could be established; such cases were excluded from the corresponding AI–expert agreement analysis. Individual expert responses were retained for analyses of inter-expert agreement and comparisons between individual experts and the AI models.

A sensitivity analysis was conducted for the shared items using specialty-specific references derived separately from the three orthodontists and the three periodontists. The median was used for ordinal variables, and the mode was used for nominal variables. When all three experts selected different categories for S3, no unique mode could be established, and that case was excluded from the corresponding sensitivity analysis. The denominators

for this analysis were 48 for S1, S2, and S4 in both perspectives, and 47 and 46 for S3 in the orthodontist and periodontist panels, respectively.

2.6. Evaluation of the AI Models

After completion of the expert assessments, the 48 synthetic clinical cases included in the statistical analysis were evaluated by three different LLMs: Claude Opus 4.8 High (Anthropic, San Francisco, CA, USA), ChatGPT-5.6 (OpenAI, San Francisco, CA, USA), and Gemini Pro 4 Hydrogen (Google, Mountain View, CA, USA). All AI evaluations were completed on 19 September 2026 within the same assessment period between 08:00 and 18:00 Turkey Time (TRT, UTC+3). The synthetic cases, assessment instructions, and structured clinical decision form were presented to all models in Turkish. A single response was analyzed for each model–case–specialty combination, and no repeated model queries were performed.

To minimize potential contextual effects arising from previous user interactions, account history, or possible personalization, new user accounts were created using email addresses different from those associated with the expert clinicians' existing personal accounts, and all evaluations were conducted through these newly created accounts. Before the evaluations, no expert case responses, expert-based reference decisions, or outputs from the other AI models were provided to these accounts.

The complete and unmodified Turkish-language role-based prompts used for the orthodontic and periodontal evaluations, including the specialty-specific instructions, predefined response categories, and coding rules, are provided verbatim in Supplementary Material S2. In each specialty-specific assessment session, the corresponding standardized role-based prompt was provided before the clinical cases. Subsequently, the case-specific clinical information, predefined orthodontic treatment plan, and accompanying clinical images were presented to the model for evaluation.

All AI evaluations were conducted through the models' standard web-based chat interfaces without API access. No generation parameters, including temperature, top-p, or maximum output tokens, were manually configured or modified. Accordingly, the evaluations were performed using the default settings available through each platform's user interface at the time of assessment.

Each AI model was evaluated in separate and independent sessions for orthodontics and periodontology. Before case evaluation, standardized role-based prompting was applied. For the periodontal assessments, the models were instructed to act as experienced periodontists and evaluate the cases from a periodontal perspective; for the orthodontic assessments, they were instructed to act as experienced orthodontists and evaluate the cases from an orthodontic perspective. In both assessment settings, the definitions, response options, and coding rules for the shared S1–S4 items were provided to the models. The periodontal sessions additionally included P1, P2a, P2b, and P2c, whereas the orthodontic sessions included O1a, O1b, and O2. For each case, the models were instructed to report their decisions using the predefined response categories.

Each model–case–specialty combination was evaluated in a separate, chat session. Accordingly, each of the 48 synthetic clinical cases was independently assessed under both the orthodontist and periodontist roles by all three AI models, resulting in 288 separate chat sessions. In each session, the corresponding standardized specialty-specific prompt was provided before the case-specific clinical information, predefined orthodontic treatment plan, and accompanying synthetic images (Supplementary Material S2). No conversational history or responses from previously evaluated cases or from the alternative specialty perspective were carried over between sessions. Furthermore, the models were not provided with the individual assessments of expert clinicians or outputs generated by the other

AI models. This procedure was intended to minimize conversational carryover effects while maintaining, as consistently as possible, the same case content, assessment language, specialty role, clinical decision criteria, and response format across model comparisons.

The procedure did not involve model training, retraining, or fine-tuning. Instead, pretrained general-purpose LLMs were evaluated from two different specialty perspectives using standardized role- and task-specific prompts. Responses obtained from each model under the orthodontic and periodontal perspectives were recorded separately, and the structured decisions derived according to the predefined coding system were used in the statistical analyses.

2.7. Outcome Measures

The primary outcome of the study was the agreement between the structured clinical decisions generated by the AI models and the expert-based reference decisions derived from independent assessments by specialists in the corresponding discipline. These reference decisions were not considered an absolute diagnostic or therapeutic gold standard but rather an expert-based operational clinical reference established for comparison with the AI models. AI–expert agreement was assessed using agreement coefficients appropriate to the measurement level of each variable together with the exact agreement rates. For ordinal variables, agreement within the ±1 category was additionally examined, and agreement coefficients were reported with the 95% confidence interval (CI).

Secondary outcomes included inter-expert agreement within each specialty, orthodontist–periodontist agreement for the shared S1–S3 items, patterns of model deviation from the expert-based reference decisions, agreement with individual experts, and comparisons of inter-model agreement. Because S4 represented subjective confidence in a decision rather than a clinical decision itself, it was treated as a secondary outcome and interpreted separately from clinical decision agreement.

2.8. Statistical Analysis

Statistical analyses were performed on the 48 synthetic clinical cases remaining after exclusion of the first two cases, which had been used for pilot form development and evaluator calibration. All analyses were conducted using Python 3.12. The scikit-learn (version 1.6.1), statsmodels (version 0.14.4), SciPy (version 1.15.2), krippendorff (version 0.8.0), and irrCAC (version 0.4.4) packages were used for agreement and reliability analyses.

Variables were classified as ordinal, binary, or nominal categorical according to their measurement properties, and categorical data were summarized as counts and percentages. S3 was treated as a nominal categorical variable because its response options did not follow a unidirectional ordering. For the multiple-response O1b and P2b items, each response option was treated as a separate binary variable. The expert-based reference for each option was derived from the three experts in the corresponding specialty using the majority rule, and AI–reference agreement was evaluated separately for each option. For the shared items S1–S4, the primary analysis used a single multidisciplinary reference based on agreement by at least four of the six experts. Cases without this level of agreement were excluded from the analysis of the respective item. A sensitivity analysis using specialty-specific references for S1–S4 was also conducted.

Inter-rater agreement among the three orthodontists, and separately among the three periodontists within their respective specialties was evaluated using the observed exact agreement rate and Krippendorff's alpha (α), with the metric specified according to the measurement level of each variable. An ordinal metric was used for ordinal items (S1, S2, S4, O2, and P1), whereas a nominal metric was used for nominal items (S3, O1a, P2a, and P2c). For multiple-response items, each response option was analyzed as a separate

binary variable. No composite score was calculated from the S1–S4 items, and each item was analyzed separately as a distinct clinical decision variable. Krippendorff's α values were reported with 95% CIs.

Agreement between the AI models and the expert-based reference decisions was evaluated using the exact agreement rate (n/N and %) and Cohen's kappa (κ). Linearly weighted Cohen's κ was used for ordinal variables, whereas unweighted Cohen's κ was used for binary and nominal variables. For Cohen's κ and Krippendorff's α coefficients, 95% CIs were estimated using case-level bootstrap resampling with 5000 replicates. For ordinal variables, agreement within the ±1 category and category-level deviations from the expert-based reference decision were also examined descriptively.

Responses coded as "9 = Unable to assess" were excluded from ordinal analyses, whereas for nominal variables, they were retained as a separate category, with their frequencies reported as counts and percentages. For the conditional P2c item, responses that were structurally not applicable were coded as missing and excluded from the corresponding analysis. For expert reliability analysis, a case contributed to Krippendorff's α only when at least two periodontists had applicable P2c responses. Structurally non-applicable responses were not treated as missing data. P2b and P2c were analyzed only when surgical intervention was considered necessary in P2a; consequently, the number of evaluable cases could vary across models. When a model provided an invariant response across evaluable cases but the reference responses varied, κ was reported as 0.000. When both the model and reference responses were invariant and identical, κ was not estimable because the expected agreement was 1.

As an additional descriptive analysis, option-specific exact agreement for O1b and P2b was evaluated using cases jointly applicable to the corresponding specialty-specific expert reference and all three AI models. Each option was treated as a binary decision (selected vs. not selected), and the expert reference was determined by the majority of valid expert responses. Option–case comparisons without an expert majority were excluded. Agreement was reported as n/N (%).

An additional test–retest analysis was performed to assess within-model response consistency using a purposively selected subset of 12 synthetic clinical cases. For each AI model and specialty-role session, responses from the original assessment were compared with those obtained during the repeat assessment. The repeat evaluations were conducted during the revision process using the same model versions as in the initial assessments. The statistical analysis of the repeated responses was completed on 10 November 2026. For single-response items, consistency was summarized using exact agreement rates and Cohen's κ coefficients, where estimable. Linear-weighted Cohen's κ was used for ordinal items, whereas unweighted Cohen's κ was used for nominal and binary items. For multiple-response items (O1b and P2b), exact response-set agreement and Jaccard similarity were calculated. Analyses of conditionally applicable items were restricted to eligible paired responses. Pooled exact agreement across single-response items was reported descriptively as agreement counts (n/N) and percentages, without confidence intervals, because multiple decision items derived from the same synthetic cases were not statistically independent. The test–retest analysis was descriptive, and no formal comparisons between models were performed.

3. Results

3.1. Inter-Expert Reliability

Inter-expert reliability, assessed using Krippendorff's α, was 0.946, 0.707, and 0.724 for S1; 0.880, 0.824, and 0.797 for S2; and 0.804, 0.579, and 0.544 for S3 among the orthodontists, periodontists, and multidisciplinary group, respectively. For S4, Krippendorff's α was 0.586

among the orthodontists, 0.734 among the periodontists, and 0.218 in the multidisciplinary group. For the specialty-specific items, α values were 0.945 for O1a, 0.934 for O2, 0.530 for P1, and 0.777 for P2a. For the conditional P2c item, Krippendorff's α was 0.617 (95% CI, 0.185–0.873) based on 18 cases with applicable responses from at least two periodontists (Table 1, Figure 1).

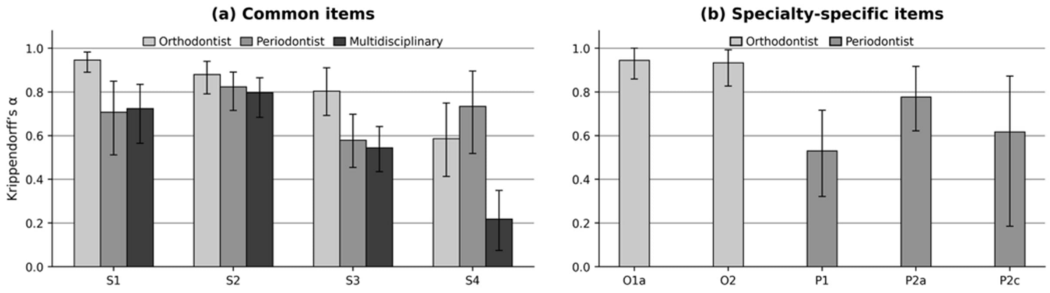


Figure 1. Inter-expert reliability assessed using Krippendorff's α. (**a**) Reliability estimates for shared clinical decision items (S1–S4) among orthodontists, periodontists, and the multidisciplinary expert panel. (**b**) Reliability estimates for specialty-specific clinical decision items among the corresponding expert groups.

Table 1. Inter-expert reliability.

Item	Expert Group	Number of Experts	Measurement Level	Evaluable Cases (*n*)	Krippendorff's α (95% CI)
Panel A. Common clinical decision items					
S1	Orthodontists	3	Ordinal	48	0.946 (0.890–0.983)
	Periodontists	3	Ordinal	48	0.707 (0.512–0.849)
	All experts	6	Ordinal	48	0.724 (0.565–0.835)
S2	Orthodontists	3	Ordinal	48	0.880 (0.791–0.940)
	Periodontists	3	Ordinal	48	0.824 (0.715–0.891)
	All experts	6	Ordinal	48	0.797 (0.684–0.865)
S3	Orthodontists	3	Nominal	48	0.804 (0.692–0.911)
	Periodontists	3	Nominal	48	0.579 (0.454–0.698)
	All experts	6	Nominal	48	0.544 (0.435–0.642)
S4	Orthodontists	3	Ordinal	48	0.586 (0.413–0.749)
	Periodontists	3	Ordinal	48	0.734 (0.518–0.896)
	All experts	6	Ordinal	48	0.218 (0.074–0.349)
Panel B. Specialty-specific clinical decision items					
O1a	Orthodontists	3	Nominal	48	0.945 (0.859–1.000)
O2	Orthodontists	3	Ordinal	48	0.934 (0.827–0.993)
P1	Periodontists	3	Ordinal	48	0.530 (0.321–0.716)
P2a	Periodontists	3	Nominal	48	0.777 (0.622–0.917)
P2c	Periodontists	3	Nominal	18	0.617 (0.185–0.873)

α, Krippendorff's alpha; CI, confidence interval. Ordinal and nominal Krippendorff's α coefficients were calculated according to the measurement level of each item. The 95% CIs were estimated using case-level bootstrap resampling with 5000 repetitions. For ordinal items, responses coded as 9 ("Unable to assess") were excluded from ordinal distance calculations. For the conditional P2c item, *n* represents the number of cases with valid responses from at least two raters that contributed to the Krippendorff's α calculation. For this item, a response was considered applicable only when the same rater had indicated that surgery was necessary in P2a (P2a = 1). Responses associated with P2a = 0 or P2a = 9 were treated as structurally not applicable and coded as missing. Five P2c entries recorded despite P2a ≠ 1 were excluded. The reliability analysis included cases with applicable P2c responses from at least two raters ($n = 18$). "All experts" represents an additional inter-rater agreement analysis including all six experts and was not used as a multidisciplinary reference for AI comparisons.

3.2. Agreement of AI Models with the Multidisciplinary Expert-Based Reference for Shared Items

For S1, exact agreement rates ranged from 73.5% to 91.2%, with Cohen's κ values ranging from 0.732 to 0.909. For S2, exact agreement rates ranged from 66.7% to 87.2%, and κ values ranged from 0.664 to 0.877. For S3, which was analyzed as a nominal variable, exact agreement rates ranged from 51.4% to 75.7%, with κ values ranging from 0.231 to 0.605. For S4, which assessed confidence in the clinical decision, exact agreement rates ranged from 18.8% to 68.8%. Cohen's κ could not be calculated for ChatGPT-5.6 (OpenAI) because of the lack of variability in the model's responses. For Gemini, κ values ranged from 0.146 to 0.184, whereas Claude Opus 4.8 High (Anthropic) yielded a κ value of 0.011 (Table 2).

Table 2. Agreement of AI models with the multidisciplinary expert-based reference for shared items (S1–S4).

Item			Orthodontist Perspective		Periodontist Perspective	
	Model	***n***	**Exact Agreement, % (95% CI)**	**κ (95% CI)**	**Exact Agreement, % (95% CI)**	**κ (95% CI)**
S1	ChatGPT-5.6 (OpenAI)	34	88.2 (73.4–95.3)	0.878 (0.754–0.970)	85.3 (69.9–93.6)	0.843 (0.712–0.962)
	Gemini Pro (Google)	34	73.5 (56.9–85.4)	0.732 (0.566–0.875)	88.2 (73.4–95.3)	0.879 (0.763–0.970)
	Claude Opus 4.8 High (Anthropic)	34	91.2 (77.0–97.0)	0.909 (0.800–1.000)	91.2 (77.0–97.0)	0.909 (0.797–1.000)
S2	ChatGPT-5.6 (OpenAI)	39	69.2 (53.6–81.4)	0.672 (0.499–0.811)	71.8 (56.2–83.5)	0.677 (0.525–0.812)
	Gemini Pro (Google)	39	87.2 (73.3–94.4)	0.877 (0.755–0.973)	79.5 (64.5–89.2)	0.798 (0.645–0.920)
	Claude Opus 4.8 High (Anthropic)	39	66.7 (51.0–79.4)	0.664 (0.531–0.795)	66.7 (51.0–79.4)	0.664 (0.532–0.785)
S3	ChatGPT-5.6 (OpenAI)	37	54.1 (38.4–69.0)	0.234 (0.102–0.382)	51.4 (35.9–66.6)	0.231 (0.113–0.361)
	Gemini Pro (Google)	37	75.7 (59.9–86.6)	0.605 (0.402–0.807)	59.5 (43.5–73.7)	0.419 (0.264–0.594)
	Claude Opus 4.8 High (Anthropic)	37	73.0 (57.0–84.6)	0.575 (0.402–0.766)	73.0 (57.0–84.6)	0.575 (0.392–0.757)
S4	ChatGPT-5.6 (OpenAI)	32	18.8 (8.9–35.3)	0.000	18.8 (8.9–35.3)	0.000
	Gemini Pro (Google)	32	65.6 (48.3–79.6)	0.146 (−0.185–0.459)	68.8 (51.4–82.0)	0.184 (−0.164–0.535)
	Claude Opus 4.8 High (Anthropic)	32	18.8 (8.9–35.3)	0.011 (0.000–0.046)	18.8 (8.9–35.3)	0.011 (0.000–0.047)

The same multidisciplinary expert-based reference was used for both AI role sessions. The "orthodontist perspective" and "periodontist perspective" columns refer to the role assigned to the AI during its separate assessment sessions, not to different expert references. A reference category was assigned when at least four of the six experts selected the same category. Cases that did not meet this consensus threshold were excluded from the primary agreement analysis for the respective item. *n* indicates the number of cases included in the primary analysis. Linear-weighted Cohen's κ was used for ordinal items (S1, S2, and S4), and unweighted Cohen's κ was used for the nominal item (S3). Exact-agreement 95% confidence intervals were calculated using the Wilson method, and κ confidence intervals were estimated using case-level bootstrap resampling with 5000 replicates. For ChatGPT-5.6, κ for S4 was 0.000 because the model provided the same response across all cases while the expert-based reference responses varied; a confidence interval was not estimated. CI, confidence interval.

Specialty-specific sensitivity-analysis results are presented in Supplementary Table S1. In the sensitivity analysis, each AI role session was compared with the reference derived from the three experts in the corresponding specialty. The denominators were 48 for S1, S2, and S4 in both perspectives; for S3, they were 47 for the orthodontist panel and 46 for the periodontist panel. Agreement estimates differed from those in the primary analysis, particularly for the periodontist-perspective session. Notably, for S1 in the ChatGPT-5.6 periodontist-role session, exact agreement decreased from 85.3% (29/34) in the primary analysis to 56.3% (27/48) in the sensitivity analysis, indicating sensitivity to the reference definition and included case set. Full results, including exact agreement rates and Cohen's κ values with 95% confidence intervals, are presented in Supplementary Table S1.

3.3. Agreement of AI Models with the Expert-Based Reference for Specialty-Specific Items

For the orthodontic-specific items, exact agreement rates ranged from 87.5% to 89.6% for O1a and from 78.7% to 83.3% for O2. The corresponding κ values ranged from 0.750 to 0.792 and from 0.739 to 0.798, respectively. For the periodontal-specific items, exact agreement rates for P1 ranged from 64.6% to 75.0%, with κ values ranging from 0.516 to 0.552; for P2a, exact agreement rates ranged from 69.6% to 84.8%, with κ values ranging from 0.289 to 0.664. In the P2c analysis, the number of evaluable cases was small and varied across models (n = 4, 12, and 10), and the confidence intervals were wide. These estimates should therefore be interpreted cautiously and do not support firm comparisons between models. Notably, the 100.0% exact agreement observed for ChatGPT-5.6 was based on only four evaluable cases and was accompanied by a wide 95% confidence interval (51.0–100.0%). Detailed results for the specialty-specific items are presented in Table 3. Exact agreement rates for the shared and specialty-specific items are shown in Figure 2, and the corresponding Cohen's κ values are presented in Figure 3.

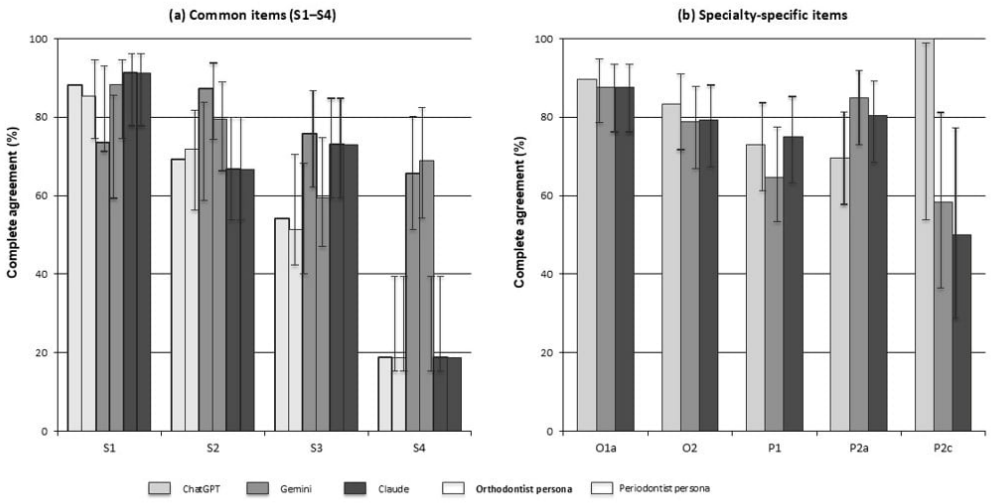


Figure 2. Exact agreement rates between the AI models and the expert-based reference decisions (95% CI).

For the multiple-response items O1b and P2b, the frequency with which each option was selected was calculated based on the applicable observations. For O1b, the options most frequently selected by the orthodontists were "avoidance of specific tooth movement" (73/73; 100.0%) and "limitation of the amount of tooth movement" (47/73; 64.4%). For P2b, the options most frequently selected by the periodontists were "connective tissue grafting/augmentation" (36/57; 63.2%) and "root coverage" (29/57; 50.9%). Selection frequencies for ChatGPT-5.6 (OpenAI), Gemini, and Claude Opus 4.8 High (Anthropic), based on the corresponding specialty persona and the number of applicable observations for

each model, are presented in Table 4. An additional descriptive analysis of option-specific exact agreement was performed using 19 common evaluable cases for O1b and 5 for P2b, with the results presented in Supplementary Table S2.

Table 3. Agreement of AI models with the expert-based reference for specialty-specific items.

Item	Model	*n*	Exact Agreement, % (95% CI)	κ (95% CI)
O1a	ChatGPT-5.6 (OpenAI)	48	89.6 (77.8–95.5)	0.792 (0.594–0.958)
	Gemini Pro (Google)	48	87.5 (75.3–94.1)	0.750 (0.543–0.916)
	Claude Opus 4.8 High (Anthropic)	48	87.5 (75.3–94.1)	0.751 (0.553–0.916)
O2	ChatGPT-5.6 (OpenAI)	48	83.3 (70.4–91.3)	0.798 (0.641–0.933)
	Gemini Pro (Google)	47	78.7 (65.1–88.0)	0.755 (0.596–0.887)
	Claude Opus 4.8 High (Anthropic)	48	79.2 (65.7–88.3)	0.739 (0.568–0.882)
P1	ChatGPT-5.6 (OpenAI)	48	72.9 (59.0–83.4)	0.528 (0.337–0.720)
	Gemini Pro (Google)	48	64.6 (50.4–76.6)	0.516 (0.325–0.702)
	Claude Opus 4.8 High (Anthropic)	48	75.0 (61.2–85.1)	0.552 (0.313–0.761)
P2a	ChatGPT-5.6 (OpenAI)	46	69.6 (55.2–80.9)	0.289 (0.061–0.514)
	Gemini Pro (Google)	46	84.8 (71.8–92.4)	0.664 (0.410–0.858)
	Claude Opus 4.8 High (Anthropic)	46	80.4 (66.8–89.3)	0.559 (0.306–0.795)
P2c	ChatGPT-5.6 (OpenAI)	4	100.0 (51.0–100.0)	NE
	Gemini Pro (Google)	12	58.3 (32.0–80.7)	0.104 (0.000–0.333)
	Claude Opus 4.8 High (Anthropic)	10	50.0 (23.7–76.3)	0.091 (0.000–0.286)

n indicates the number of evaluable cases for the respective item. Linear-weighted Cohen's κ was used for ordinal items (O2 and P1), whereas unweighted Cohen's κ was used for nominal items (O1a, P2a, and P2c). For each AI model, P2c was evaluated only when that model's P2a response indicated that surgical intervention was necessary; consequently, the number of evaluable cases varied across models. NE indicates that κ was not estimable because the model and expert-based reference responses were identical and invariant across the evaluable cases. κ, Cohen's kappa; CI, confidence interval; NE, not estimable.

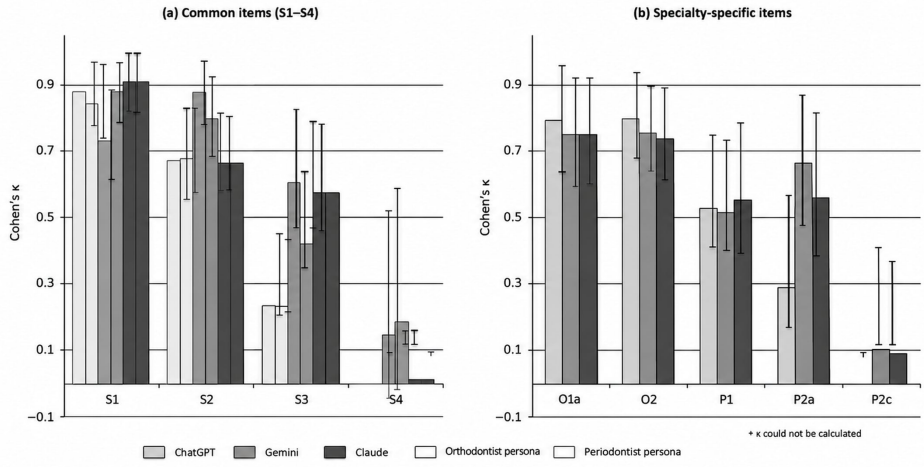


Figure 3. Cohen's κ values for agreement between the AI models and the expert-based reference decisions (95% CI).

Table 4. Selection frequencies for individual options in the multiple-response items O1b and P2b based on applicable observations.

Item	Option	Human Experts, *n* (%)	ChatGPT-5.6 (OpenAI) *n* (%)	Gemini Pro (Google) *n* (%)	Claude Opus 4.8 High (Anthropic) *n* (%)
O1b	Denominator (applicable observations, n) †	*73*	*24*	*25*	*21*
	1—Limitation of the amount of tooth movement	47 (64.4)	20 (83.3)	18 (72.0)	5 (23.8)
	2—Modification of the direction of tooth movement	11 (15.1)	5 (20.8)	7 (28.0)	3 (14.3)
	3—Avoidance of specific tooth movement	73 (100.0)	24 (100.0)	25 (100.0)	2 (9.5)
	4—Modification of force/biomechanics	23 (31.5)	0 (0.0)	13 (52.0)	16 (76.2)
	5—Modification of anchorage	8 (11.0)	1 (4.2)	5 (20.0)	13 (61.9)
	6—Limitation of treatment objectives	3 (4.1)	1 (4.2)	5 (20.0)	7 (33.3)
	7—Modification of the extraction/tooth-preservation plan	0 (0.0)	0 (0.0)	0 (0.0)	1 (4.8)
	8—Other	0 (0.0)	0 (0.0)	1 (4.0)	0 (0.0)
P2b	Denominator (applicable observations, n) †	*57*	*5*	*13*	*11*
	1—Free gingival graft (FGG)	20 (35.1)	0 (0.0)	2 (15.4)	1 (9.1)
	2—Connective tissue grafting/augmentation	36 (63.2)	5 (100.0)	9 (69.2)	8 (72.7)
	3—Root coverage	29 (50.9)	3 (60.0)	7 (53.8)	7 (63.6)
	4—Frenotomy/frenectomy	13 (22.8)	0 (0.0)	4 (30.8)	2 (18.2)
	5—Vestibuloplasty	7 (12.3)	0 (0.0)	1 (7.7)	0 (0.0)
	6—Other	2 (3.5)	0 (0.0)	0 (0.0)	0 (0.0)

Because multiple options could be selected for a single case in the multiple-response items, the sum of the category percentages may exceed 100%. The human expert column represents pooled evaluator × case observations derived from the individual assessments of the three orthodontists for O1b and the three periodontists for P2b. The AI models evaluated O1b under the orthodontist persona and P2b under the periodontist persona. † For calculation of the percentages, the denominator was defined as the number of observations for which the respective item was applicable; O1b was considered applicable only for observations with O1a = 1, whereas P2b was considered applicable only for observations with P2a = 1. Because the O1a and P2a decisions could differ across AI models, the number of applicable observations, and consequently the denominators for O1b and P2b, could also vary across models. Because the denominators and applicable case sets differed between human experts and AI models, these percentages are descriptive and should not be interpreted as direct comparisons of AI–expert agreement or model performance.

Responses coded as "Unable to assess" (code 9) were infrequent throughout the study and occurred only for periodontal items. Among the human experts, this response was recorded in 2/144 (1.4%) observations for P2a and 2/63 (3.2%) for P2b. Among the AI models, "Unable to assess" responses were observed only for ChatGPT-5.6 (OpenAI), occurring in 1/48 (2.1%) observations for P2a, 1/6 (16.7%) for P2b, and 1/5 (20.0%) for P2c. Neither Gemini nor Claude Opus 4.8 High (Anthropic) generated an "Unable to assess" response for any item. Item-level response-category distributions, the handling of responses coded as "Unable to assess", and model–reference disagreement patterns are presented in Supplementary Table S3.

3.4. Test–Retest Consistency of AI Responses

An additional test–retest analysis was conducted using a purposively selected subset of 12 clinical cases to assess within-model response stability. The results demonstrated variability in consistency across AI models, specialty-role sessions, and clinical decision items. In the periodontist-role sessions, pooled exact agreement across single-response

items was 73.7% for ChatGPT-5.6, 85.5% for Gemini Pro, and 81.6% for Claude Opus 4.8 High. These pooled estimates were reported descriptively without confidence intervals because multiple decision items originated from the same clinical cases. Item-level agreement varied substantially; for example, ChatGPT-5.6 showed 91.7% agreement for P2a but only 41.7% for S4. For multiple-response items, exact response-set agreement was generally lower, although Jaccard similarity indicated partial overlap between the original and repeated selections. These findings suggest that within-model response stability was dependent on the clinical decision item and assessment context. Detailed test–retest results, including exact agreement rates, Cohen's κ coefficients, and Jaccard similarity measures, are provided in Supplementary Table S4.

4. Discussion

The primary aim of this study was to evaluate the agreement between AI-generated decisions and those of expert clinicians in structured cases reflecting multidisciplinary clinical decision-making between orthodontics and periodontology, rather than to assess AI performance within a single dental specialty. A previous study evaluating AI performance across different stages of clinical decision-making reported that AI performed significantly worse than expert clinicians not only in diagnosis but also in selecting appropriate additional investigations and recommending treatment, with some recommendations being clinically relevant but incomplete [40]. These findings highlight the importance of assessing whether AI models can integrate complex clinical findings and generate appropriate specialty-specific decisions in interdisciplinary clinical scenarios.

Synthetic clinical cases enable LLMs to be compared under consistent clinical conditions while reducing privacy concerns associated with the use of real patient data [41]. However, the synthetic images generated using ChatGPT Images may not fully reproduce the subtle grayscale variations, anatomical distortions, imaging artifacts, and complex diagnostic features encountered in real clinical and radiographic images. These limitations are particularly relevant to imaging modalities such as cone-beam computed tomography (CBCT), periapical radiography, and cephalometric radiography, in which subtle anatomical and radiographic details may influence clinical interpretation. Consequently, the standardized visual representations used in this study may not adequately reflect the diagnostic complexity and uncertainty associated with interpreting real patient images. Nevertheless, standardized case content alone may not ensure comparability. Buldur et al. demonstrated that presenting the same clinical content using different prompt structures could influence AI–expert agreement [15]. Accordingly, in the present study, both case presentation and prompt structure were kept consistent across models.

The relevance of LLMs in the context of the present study lies not merely in their ability to recognize individual findings but in their potential to integrate periodontal and orthodontic variables within the same clinical context and translate them into clinical decisions. However, accurate recognition of a clinical risk does not necessarily imply that an appropriate clinical decision will follow [42]. Moreover, LLM–expert reference agreement has been reported to vary according to the clinical domain and case complexity [43]. Evaluating model performance across distinct stages of clinical decision-making, rather than relying on a single overall performance measure, may therefore provide a more informative characterization of model behavior.

In the present study, assessing inter-expert agreement alongside the comparison of LLM decisions with expert-based references allowed model–expert reference discrepancies to be interpreted within their clinical context. A model–expert discrepancy in a decision domain characterized by high agreement among experts may not have the same clinical implications as a discrepancy in a domain in which expert opinions themselves are more

variable. Accordingly, a more balanced assessment of LLM performance may require the consideration of not only concordance with the reference decision but also the degree of inter-expert variability within the corresponding decision domain. The sensitivity analysis further highlighted the influence of reference construction and case selection on AI–expert agreement estimates. The higher agreement observed in the primary analysis may partly reflect the exclusion of cases without sufficient multidisciplinary consensus. Because both the reference definition and case set differed between analyses, their individual effects could not be distinguished. Furthermore, when multiple treatment or management strategies are clinically acceptable, differences in clinical priorities, professional judgment, and the relative weighting of treatment options may contribute to disagreement among experts and between experts and LLMs. Therefore, disagreement with an expert-based reference decision should not automatically be interpreted as a clinically inappropriate response, and statistical agreement measures should be considered alongside the clinical context and the potential acceptability of alternative management approaches. The test–retest analysis revealed lower exact response-set agreement for certain multiple-response items despite partial overlap between response sets. This suggests that the reproducibility of specific treatment recommendations may be limited across repeated evaluations.

Cases requiring the integrated assessment of orthodontic and periodontal considerations involve the simultaneous evaluation of multiple clinical factors, including disease activity, periodontal support, anatomical risk factors, and the planned orthodontic tooth movement [11,13]. The variation in inter-expert agreement across decision domains observed in this study suggests that some clinical decisions may be influenced by specialty perspective. This pattern was particularly evident for S4: although agreement was higher within each specialty group, it decreased when all experts were considered together ($\alpha = 0.218$), which may indicate that the same clinical findings can be interpreted differently across disciplines.

In determining periodontal readiness to initiate orthodontic treatment, disease activity and periodontal stability should be considered alongside the extent of periodontal support loss. In patients with periodontitis, control of inflammation and achievement of periodontal stability before the initiation of orthodontic treatment are recommended [23]. Within this context, the high level of agreement observed for S1 between the models and the expert-based references from both specialties may suggest that the LLMs were able, to some extent, to integrate the clinical variables relevant to this decision domain.

Periodontal risk assessment requires the integration of multiple variables, including periodontal phenotype, alveolar bone morphology, root position, and the direction of the planned orthodontic tooth movement [11]. Although proclination has been reported as a potential risk factor for gingival recession, the development of recession cannot be attributed to a single orthodontic variable and should instead be evaluated in conjunction with periodontal and mucogingival characteristics [3]. The differences in agreement observed across models within this decision domain may suggest that the models integrated or weighted the same clinical variables differently. The high level of inter-expert agreement further suggests that these differences are unlikely to be explained solely by variability in the expert-based reference decisions.

When translating periodontal risk into the orthodontic treatment approach, it is important to consider the existing periodontal support together with the characteristics of the planned orthodontic tooth movement [11]. For S3, the variation in model–expert reference agreement across both models and specialty-specific expert references may indicate that identifying a risk and modifying treatment in response to that risk represent distinct stages of clinical decision-making. Differences in agreement between orthodontists and periodontists further suggest that this decision may be influenced by specialty perspective.

Previous findings indicating that recognition of a clinical risk does not always translate into an equally appropriate clinical action further support the relevance of this distinction [42].

The variability in model–expert agreement for decision confidence suggests that expressed confidence may not directly correspond to clinical decision performance. Similarly, although agreement was more consistent within each specialty group, the decrease in agreement when all experts were considered together ($\alpha = 0.218$) may indicate that decision confidence is influenced by specialty perspective. Therefore, a high level of expressed confidence should not, by itself, be interpreted as an indicator of clinical appropriateness. The findings of Mayer et al., who reported that clinically inadequate or potentially harmful responses could occur in periodontal clinical scenarios despite high overall performance, further support this cautious interpretation [13].

In the presence of periodontal risk, orthodontic treatment may be adapted in several ways, including limiting the planned tooth movement, modifying biomechanics or anchorage, and reconsidering treatment objectives [10,12]. Periodontal support loss may also alter the center of resistance and the moment-to-force relationship, potentially requiring the biomechanical approach to be adjusted to the existing periodontal conditions [44]. In the present study, although the models showed high agreement with the expert-based reference in determining whether modification of the orthodontic treatment plan was required (O1a), marked differences emerged regarding the specific modifications selected (O1b). This divergence may be related to the fact that O1b requires clinical prioritization among multiple potential interventions rather than the recognition of a single correct option. Indeed, the same periodontal risk may be managed through several approaches, such as limiting tooth movement, modifying biomechanics, or reconsidering treatment objectives. Thus, the differences observed for O1b may reflect not only variability in model performance, but also differences in the relative emphasis placed by the models on alternative management strategies for the same clinical risk. The findings of Tanaka et al., who reported differences between AI-generated orthodontic treatment plans and recommendations made by expert orthodontists, further support the view that treatment-plan formulation should be considered as a distinct stage of clinical decision-making beyond the recognition of clinical findings [45].

A similar distinction is relevant to treatment timing. The need for periodontal intervention does not necessarily imply that orthodontic treatment must be completely deferred in every case. Studies examining concurrent or sequential periodontal and orthodontic treatment have reported that under appropriate periodontal control, different treatment sequences may lead to different outcomes in terms of clinical results and overall treatment duration [20,46]. Within this context, the relatively high model–expert agreement observed for O2 suggests that the models may have evaluated periodontal treatment needs not only in terms of their presence, but also in relation to timing with orthodontic treatment in a manner broadly consistent with the expert-based reference.

Achieving periodontal stability before orthodontic treatment is important; however, the required approach may range from preventive and oral hygiene-focused measures to active periodontal treatment [28]. Accordingly, P1 requires not only the recognition of periodontal disease but also interpretation of the clinical significance of the findings and the level of intervention required. The relatively limited agreement observed for P1, both among experts and between the models and the expert-based reference, suggests that this decision may be more susceptible to differences in interpretation depending on case-specific characteristics.

Mucogingival decisions involve sequential but clinically distinct questions: P2a assesses whether surgical intervention is required, P2b addresses which intervention is appropriate, and P2c evaluates the timing of that intervention. Determining the need for surgery requires

the integrated assessment of multiple factors, including the amount of keratinized tissue, vestibular depth, frenulum attachment, alveolar bone morphology, root position, and the planned orthodontic tooth movement [11,35]. Although a thin periodontal phenotype or tooth movements considered at risk in relation to the alveolar boundaries may warrant the consideration of phenotype modification in selected cases, the presence of periodontal risk or the need for periodontal consultation does not, by itself, constitute an indication for surgery [25,38]. Accordingly, determining the need for surgical intervention (P2a) does not automatically determine the type of intervention (P2b) or its timing (P2c); in selected cases, grafting may also be deferred until after orthodontic treatment [38]. Particularly for P2b, which allowed multiple responses, more than one approach may be clinically reasonable, and experts may differ in their preferred intervention according to their clinical experience and priorities. Therefore, not all disagreement at this decision stage should necessarily be interpreted as clinical decision error. Similarly, the timing decision in P2c may be influenced not only by periodontal conditions, but also by clinical considerations related to the planned orthodontic tooth movement and overall treatment sequencing.

The variation in model–expert agreement across P2a, P2b, and P2c suggests that LLM performance may not be uniform across the different stages of mucogingival decision-making. However, because P2b and P2c were conditional decisions dependent on P2a, the number and composition of evaluable cases differed across models. This limitation was particularly pronounced for P2c, for which only 4–12 cases were evaluable per model, resulting in wide 95% confidence intervals and substantial uncertainty in the agreement estimates. Notably, the 100.0% exact agreement observed for ChatGPT-5.6 was based on only four evaluable cases and was accompanied by a wide 95% confidence interval (51.0–100.0%). Consequently, apparent differences in agreement rates should be interpreted cautiously and do not support definitive conclusions regarding the superiority of any model in determining the timing of mucogingival surgical intervention. Furthermore, because the models were evaluated on different conditional subsets, these findings should not be interpreted as direct comparisons of model performance. Decisions regarding the need for surgical intervention, the choice of treatment, and its timing involve different clinical considerations. Evaluating these decisions separately may provide a more detailed assessment of LLM performance. This pattern is consistent with previous studies suggesting that the ability of LLMs to recognize clinical findings may not necessarily translate to the same degree into case-specific and clinically actionable treatment decisions [47,48]. The variation in disagreement patterns across items suggests that alignment with the expert-based reference depended on the type of clinical judgment required. Decisions involving risk interpretation, treatment intensity, and the indication for surgery may require different forms of clinical reasoning, which could help explain why agreement was not uniform across items. These observations should be interpreted as exploratory, particularly for conditional decisions such as P2c, where the small number of evaluable cases limits the strength of any inference.

Limitations

This study has several limitations. Although an additional test–retest analysis was conducted using 12 selected cases, this limited subset may not fully represent response variability across all 48 cases. Therefore, the findings regarding within-model consistency should be interpreted cautiously. The study was based on a limited pool of investigator-developed synthetic cases accompanied by AI-generated synthetic images. Although the cases were designed to represent a range of clinical conditions and to provide a standardized assessment environment, they may not fully capture the clinical and imaging variability encountered in real-world patient populations. In particular, AI-generated synthetic images may not faithfully reproduce the subtle grayscale differences, anatomical distortions, imag-

ing artifacts, or complex diagnostic details present in real patient radiographs, including CBCT, periapical, and cephalometric images. Such differences may limit the extent to which the observed model–expert agreement can be generalized to clinical settings requiring the interpretation of authentic diagnostic images. In addition, structured scenarios cannot fully reproduce incomplete or conflicting information, the need for additional information or assessment, or the dynamic patient–clinician interactions encountered in routine clinical practice. This study also did not evaluate the ability of LLMs to independently interpret real clinical or radiographic images. Accordingly, the findings should be interpreted within the context of a standardized synthetic case environment and should not be regarded as direct evidence of independent diagnostic or treatment-planning performance in real-world clinical settings. Future studies using multimodal case sets incorporating real clinical and radiographic data may help assess the external validity and clinical applicability of these findings.

The primary analysis was restricted to cases in which at least four of the six experts selected the same category. Therefore, its agreement estimates represent cases with multidisciplinary consensus and may not reflect cases in which the two specialties or individual experts disagreed. The sensitivity analysis using specialty-specific references yielded different agreement estimates, particularly for the periodontist-perspective session, indicating that the results depend in part on how the expert reference is defined.

Two investigators involved in developing the synthetic cases also participated in the expert evaluations, which may have introduced bias related to prior familiarity with the cases. All experts assessed the cases independently, had no access to one another's responses, and did not discuss or revise their assessments after evaluation. These safeguards reduced the potential for influence between evaluators but could not eliminate the possibility that prior familiarity affected the judgments of the two investigators. Future validation studies should use independent teams for case development and expert evaluation, with evaluators blinded to the case-development process. The sensitivity analysis yielded substantially different agreement estimates, particularly for the periodontist-role session. Because both reference construction and case inclusion differed between analyses, their individual effects could not be determined, and agreement estimates should be interpreted in relation to these methodological choices.

5. Conclusions

This study showed that the agreement between LLMs and expert-based reference decisions in multidisciplinary orthodontic–periodontal decision-making varied across clinical decision domains. Agreement was generally higher for fundamental periodontal risk assessments and orthodontic treatment-planning decisions, whereas greater variability was observed in more complex and interpretive multidisciplinary and periodontal decisions. These findings suggest that LLMs may have a supportive role in structured clinical assessments but should not be considered independent substitutes for expert clinical judgment. Further studies involving real patient data and diverse clinical settings are needed to determine the clinical applicability and generalizability of these models.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/diagnostics16234013/s1, Supplementary Material S1: Synthetic orthodontic–periodontal case scenarios and clinical data; Supplementary Material S2: Structured Case-Based Clinical Decision Assessment Form for 48 synthetic orthodontic–periodontal cases; Supplementary Table S1: Agreement of AI models with specialty-specific expert-based references for shared items (S1–S4) in the sensitivity analysis; Supplementary Table S2: Option-specific exact agreement between AI models and specialty-specific expert-based reference decisions in common evaluable cases; Supplementary Table S3: Response-category distributions and model–expert reference agreement

patterns, including the handling of responses coded 9 ("Unable to assess"); Supplementary Table S4: Test–retest consistency of AI responses across specialty-role sessions.

Author Contributions: Conceptualization, Z.B.Ç.E. and S.S.; methodology, Z.B.Ç.E. and S.S.; validation, Z.B.Ç.E., S.S., H.A., E.Ş., Ş.A. and G.D.; formal analysis, Z.B.Ç.E.; investigation, Z.B.Ç.E. and S.S.; resources, Z.B.Ç.E. and S.S.; data curation, Z.B.Ç.E., S.S., H.A. and Ş.A.; writing—original draft preparation, Z.B.Ç.E. and S.S.; writing—review and editing, Z.B.Ç.E. and S.S.; visualization, Z.B.Ç.E. and S.S.; supervision, Z.B.Ç.E.; project administration, Z.B.Ç.E. and S.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data supporting the findings of this study, including the data generated from the evaluation of the synthetic clinical cases, are available from the corresponding author upon reasonable request.

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

Abbreviations

LLM	Large language model
AI	Artificial intelligence
S1	Periodontal readiness to initiate orthodontic treatment
S2	Risk of adverse periodontal effects during orthodontic treatment
S3	Need for modification of the orthodontic approach based on periodontal conditions
S4	Level of confidence in the clinical assessment
O1a	Need to modify the orthodontic treatment plan
O1b	Nature of the required modification to the orthodontic treatment plan
O2	Timing and sequencing of orthodontic treatment
P1	Need for periodontal disease control
P2a	Need for mucogingival surgical intervention
P2b	Type of mucogingival surgical intervention
P2c	Timing of mucogingival surgical intervention
CI	Confidence interval
κ	Cohen's kappa

References

1. Amar, S.A.O.; Ennibi, O.; Erraji, S. Periodontal treatment, a key of successful orthodontic treatment. *Int. Dent. J.* **2024**, *74*, S294. [CrossRef]
2. Luchian, I.; Surlari, Z.; Goriuc, A.; Ioanid, N.; Zetu, I.; Butnaru, O.; Scutariu, M.-M.; Tatarciuc, M.; Budala, D.-G. The Influence of Orthodontic Treatment on Periodontal Health between Challenge and Synergy: A Narrative Review. *Dent. J.* **2024**, *12*, 112. [CrossRef] [PubMed]
3. Cadenas de Llano-Pérula, M.; Castro, A.B.; Danneels, M.; Schelfhout, A.; Teughels, W.; Willems, G. Risk factors for gingival recessions after orthodontic treatment: A systematic review. *Eur. J. Orthod.* **2023**, *45*, 528–544. [CrossRef] [PubMed]
4. Górski, B.; Kalina, E.; Derwich, M. Mechanobiology and Molecular Regulation of Periodontal Tissue Remodeling During Orthodontic Tooth Movement: From Periodontal Phenotype to Precision Orthodontics—A Narrative Review. *Int. J. Mol. Sci.* **2026**, *27*, 7698. [CrossRef] [PubMed]
5. Sahoo, H.S.; Kumar, A.R.P.; Durvasulu, A.; Ravichandran, A. Comparative analysis of diagnostic prediction and clinical reasoning of large language models in complex endodontic case scenarios. *J. Dent.* **2026**, *166*, 106309. [CrossRef] [PubMed]
6. Lechien, J.R.; Georgescu, B.M.; Hans, S.; Chiesa-Estomba, C.M. ChatGPT performance in laryngology and head and neck surgery: A clinical case-series. *Eur. Arch. Otorhinolaryngol.* **2024**, *281*, 319–333. [CrossRef] [PubMed]
7. Alsayed, N.A. The Role of Generative Artificial Intelligence in the Diagnosis and Treatment of Malocclusion. *Cureus* **2026**, *18*, e105050. [CrossRef] [PubMed]
8. Kaplan, T.T.; Akyol, G.E. The role of large language models in dental diagnosis, decision-making, and communication: A systematic review. *Jpn. Dent. Sci. Rev.* **2026**, *62*, 57–67. [CrossRef] [PubMed]

9. Erbe, C.; Heger, S.; Kasaj, A.; Berres, M.; Wehrbein, H. Orthodontic treatment in periodontally compromised patients: A systematic review. *Clin. Oral Investig.* **2023**, *27*, 79–89. [CrossRef] [PubMed]
10. Han, S.H.; Ko, Y.; Ham, L.K.; Park, J.H.; Kim, Y. Precautions and possibilities in orthodontic treatment of periodontally compromised patients: Current recommendations. *J. Esthet. Restor. Dent.* **2024**, *36*, 595–605. [CrossRef] [PubMed]
11. Zhong, W.; Zhou, C.; Yin, Y.; Feng, G.; Zhao, Z.; Pan, Y.; Bai, Y.; Jin, Z.; Xu, Y.; Fang, B.; et al. Expert consensus on orthodontic treatment of patients with periodontal disease. *Int. J. Oral Sci.* **2025**, *17*, 27. [CrossRef] [PubMed]
12. Jepsen, K.; Sculean, A.; Jepsen, S. Complications and treatment errors involving periodontal tissues related to orthodontic therapy. *Periodontol. 2000* **2023**, *92*, 135–158. [CrossRef] [PubMed]
13. Mayer, Y.; Demetriou, B.; Rasperini, G.; Gabay, E.; Gašperšič, R.; Božić, D.; Calatrava, J.; Ginesin, O.; Giladi, H.Z.; Almeida, R.F. Large Language Models and Retrieval-Augmented Platforms for the Diagnosis and Management of Periodontal Diseases: A Blinded Expert-Rated Comparative Study of 11 Systems. *J. Clin. Periodontol.* **2026**, *53*, 1546–1556. [CrossRef] [PubMed]
14. Fanelli, F.; Saleh, M.; Santamaria, P.; Zhurakivska, K.; Nibali, L.; Troiano, G. Development and Comparative Evaluation of a Re-instructed GPT-4o Model Specialized in Periodontology. *J. Clin. Periodontol.* **2025**, *52*, 707–716. [CrossRef] [PubMed]
15. Buldur, M.; Ayan, G.; Misilli, T. Prognostic Prediction of Avulsed Permanent Teeth Using Conversational AI Models Versus Expert Dentists: Influence of Prompt Structure and Temporal Stability. *Dent. Traumatol.* **2026**, *42*, 747–758. [CrossRef] [PubMed]
16. Wu, Y.; Zhang, Y.; Xu, M.; Jinzhi, C.; Xue, Y.; Zheng, Y. Effectiveness of various general large language models in clinical consensus and case analysis in dental implantology: A comparative study. *BMC Med. Inform. Decis. Mak.* **2025**, *25*, 147. [CrossRef] [PubMed]
17. Londono, J.; Ghasemi, S.; Moghaddasi, N.; Baninajarian, H.; Fahimipour, A.; Hashemi, S.; Fathi, A.; Dashti, M. Prevalence of malocclusion in Turkish children and adolescents: A systematic review and meta-analysis. *Clin. Exp. Dent. Res.* **2023**, *9*, 689–700. [CrossRef] [PubMed]
18. Wishney, M. Potential risks of orthodontic therapy: A critical review and conceptual framework. *Aust. Dent. J.* **2017**, *62*, 86–96. [CrossRef] [PubMed]
19. Fleming, P.S.; Andrews, J. Periodontitis: Orthodontic implications and management. *Br. Dent. J.* **2024**, *237*, 334–340. [CrossRef] [PubMed]
20. Tu, K.W.; Kuo, C.H.; Hung, C.C.; Yan, D.Y.; Mau, J.L.P. Strategic sequencing of orthodontic treatment and periodontal regenerative surgery: A literature review. *J. Dent. Sci.* **2025**, *20*, 1391–1397. [CrossRef] [PubMed]
21. Abasi, M.; Goodarzi, A.; Solhmirzaei, R.; Fazel, F.; Moharrami, A.; Ghasemi, H.; Milani, H.A. The Interaction Between Orthodontics and Periodontal Tissue Remodeling. *Galen Med. J.* **2025**, *14*, e4005. [CrossRef] [PubMed]
22. Chackartchi, T.; Polak, D.; Stabholz, A.; Chaushu, S. Orthodontic treatment in periodontitis patients. *Periodontol. 2000* **2025**, 1–10. [CrossRef] [PubMed]
23. Herrera, D.; Sanz, M.; Kebschull, M.; Jepsen, S.; Sculean, A.; Berglundh, T.; Papapanou, P.N.; Chapple, I.; Tonetti, M.S. EFP Workshop Participants and Methodological Consultant. Treatment of stage IV periodontitis: The EFP S3 level clinical practice guideline. *J. Clin. Periodontol.* **2022**, *49*, 4–71. [CrossRef] [PubMed]
24. Stähli, A.; Orishko, A.; Imber, J.C.; Roccuzzo, A.; Sculean, A.; Kloukos, D.; Salvi, G.E. Supportive periodontal care during active orthodontic therapy in patients with history of stage IV periodontitis: A narrative review. *Periodontol. 2000* **2025**, 1–16. [CrossRef] [PubMed]
25. Park, W.B.; Park, W.; Lim, S.W.; Han, J.Y. Periodontal Phenotype Modification Using Subepithelial Connective Tissue Graft and Bone Graft in the Mandibular Anterior Teeth with Mucogingival Problems Following Orthodontic Treatment. *Medicina* **2023**, *59*, 584. [CrossRef] [PubMed]
26. Laursen, M.G.; Rylev, M.; Melsen, B. The role of orthodontics in the repair of gingival recessions. *Am. J. Orthod. Dentofac. Orthop.* **2020**, *157*, 29–34. [CrossRef] [PubMed]
27. Aiyer, S.; Higham, H.; Yeung, N. Confidence and certainty in medical diagnoses within acute healthcare: A scoping review. *BMJ Qual. Saf.* **2026**, *35*, 188–200. [CrossRef] [PubMed]
28. Kwon, T.H.; Salem, D.M.; Levin, L. Periodontal considerations in orthodontic treatment: A review of the literature and recommended protocols. *Semin. Orthod.* **2024**, *30*, 80–88. [CrossRef]
29. Guo, R.; Zhang, L.; Hu, M.; Huang, Y.; Li, W. Alveolar bone changes in maxillary and mandibular anterior teeth during orthodontic treatment: A systematic review and meta-analysis. *Orthod. Craniofac. Res.* **2021**, *24*, 165–179. [CrossRef] [PubMed]
30. Wang, C.W.; Yu, S.H.; Mandelaris, G.A.; Wang, H.L. Is periodontal phenotype modification therapy beneficial for patients receiving orthodontic treatment? An American Academy of Periodontology best evidence review. *J. Periodontol.* **2020**, *91*, 299–310. [CrossRef] [PubMed]
31. Hashim, N.T.; Dasnadi, S.P.; Ziada, H.; Rahman, M.M.; Ahmed, A.; Mohammed, R.; Gismalla, B.G.; Abubakr, N.H. Orthodontic Management of Different Stages and Grades of Periodontitis According to the 2017 Classification of Periodontal Diseases. *Dent. J.* **2025**, *13*, 59. [CrossRef] [PubMed]

32. Carbone, A.C.; Joly, J.C.; Botelho, J.; Machado, V.; Avila-Ortiz, G.; Cairo, F.; Chambrone, L. Long-term stability of gingival margin and periodontal soft-tissue phenotype achieved after mucogingival therapy: A systematic review. *J. Clin. Periodontol.* **2024**, *51*, 177–195. [CrossRef] [PubMed]
33. Cortellini, P.; Bissada, N.F. Mucogingival conditions in the natural dentition: Narrative review, case definitions, and diagnostic considerations. *J. Periodontol.* **2018**, *89*, S204–S213. [CrossRef] [PubMed]
34. Skhiri, M.; Belgacem, O.; Abid, R.; Belhajkacem, S.; Nasri, W.; Mbarek, R. Frenectomy: Optimizing periodontal health and function. *Int. Dent. J.* **2024**, *74*, S249. [CrossRef]
35. Moreno Rodríguez, J.A.; Ortiz Ruiz, A.J. A Surgical Approach to Root Coverage and Correction of Mucogingival Conditions and Deformities in Mandibular Incisors with Isolated Gingival Recession: Free Mucogingival Graft. A Pilot Prospective Cohort Study. *Int. J. Periodontics Restor. Dent.* **2024**, *44*, 167–175. [CrossRef] [PubMed]
36. Agusto, M.; Salman, A.; Parker, D.; Choi, D.; Schincaglia, G.P. Root Coverage Predictability in the Treatment of Gingival Recessions on Mandibular Anterior Teeth. *JDR Clin. Trans. Res.* **2022**, *7*, 224–233. [CrossRef] [PubMed]
37. Clementini, M.; Signorini, L.; Lione, R. The impact of orthodontic treatment on gingival recession: Prevention and therapeutic strategies. *Oral Implantol.* **2025**, *17*, 193–195. [CrossRef]
38. Wang, T.; Zhao, R.; Yang, R.; Li, Y.; Lien, H.L.; Mei, L.; Nogueira, G. Perceptions of NZ orthodontists and periodontists on the management of gingival recession in orthodontic patients. *Aust. Dent. J.* **2022**, *67*, S41–S49. [CrossRef] [PubMed]
39. Park, S.Y. Mucogingival surgery for patients under orthodontic treatment. *Taehan Chikkwa Uisa Hyophoe Chi* **2017**, *55*, 249–256. [CrossRef]
40. Peters, M.; Le Clercq, M.; Yanni, A.; Vanden Eynden, X.; Martin, L.; Vanden Haute, N.; Tancredi, S.; De Passe, C.; Boutremans, E.; Lechien, J.; et al. ChatGPT and trainee performances in the management of maxillofacial patients. *J. Stomatol. Oral Maxillofac. Surg.* **2025**, *126*, 102090. [CrossRef] [PubMed]
41. Huang, H.; Zheng, O.; Wang, D.; Yin, J.; Wang, Z.; Ding, S.; Yin, H.; Xu, C.; Yang, R.; Zheng, Q.; et al. ChatGPT for shaping the future of dentistry: The potential of multi-modal large language model. *Int. J. Oral Sci.* **2023**, *15*, 29. [CrossRef] [PubMed]
42. Karaoğlan, N.; Karaoğlan, H. Patient-Facing AI Chatbot Treatment-Direction Advice in Orthodontic Health Communication: A Scenario-Based Comparison with Expert Consensus. *Healthcare* **2026**, *14*, 2565. [CrossRef] [PubMed]
43. Chiesa-Estomba, C.M.; Lechien, J.R.; Vaira, L.A.; Brunet, A.; Cammaroto, G.; Mayo-Yanez, M.; Sanchez-Barrueco, A.; Saga-Gutierrez, C. Exploring the potential of ChatGPT as a supportive tool for sialendoscopy clinical decision making and patient information support. *Eur. Arch. Otorhinolaryngol.* **2024**, *281*, 2081–2086. [CrossRef] [PubMed]
44. Choi, S.H.; Kim, Y.H.; Lee, K.J.; Hwang, C.J. Effect of labiolingual inclination of a maxillary central incisor and surrounding alveolar bone loss on periodontal stress: A finite element analysis. *Korean J. Orthod.* **2016**, *46*, 155–162. [CrossRef] [PubMed]
45. Tanaka, O.M.; Gasparello, G.G.; Mota-Júnior, S.L.; Bark, M.J.; Rozyscki, J.D.A.A.; Wolanski, R.B. Effectiveness of AI-generated orthodontic treatment plans compared to expert orthodontist recommendations: A cross-sectional pilot study. *Dent. Press J. Orthod.* **2025**, *30*, e2524186. [CrossRef] [PubMed]
46. Razali, N.S.; Adam, F.A.; Noor, E. Multidisciplinary relationship between Periodontology and orthodontic department: A clinical audit study. *Compend. Oral Sci.* **2024**, *11*, 60–67. [CrossRef]
47. Issa, J.; Sohrabniya, F.; Brinz, J.; Dyszkiewicz-Konwińska, M.; Chaurasia, A. The role of large language models in dental diagnosis, treatment planning, and prognosis. *Czas. Stomatol.* **2025**, *78*, 298–303. [CrossRef]
48. Akdeniz, C.; Doğru, A.G. Comparison of artificial intelligence-based chatbots and expert periodontists in responding to patient questions: A multi-dimensional analysis. *BMC Oral Health* **2026**, *26*, 1278. [CrossRef] [PubMed]

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