

Comparative analysis of the performance of ChatGPT-5.4, Claude Sonnet 4.6, and Gemini 3.1 Pro large language models on theoretical orthodontic questions

Buket Erdem¹, Salih Baver Ataş², Meriç Arslan¹, Çağla Şar¹

¹ İstanbul Health and Technology University, Faculty of Dentistry, Department of Orthodontics, İstanbul, Türkiye

² Private Practice, Diyarbakır, Türkiye

Received: 6 September 2026

Revised: 30 September 2026

Accepted: 15 October 2026

Published: 19 October 2026

Correspondence:

Dr. Buket ERDEM

İstanbul Health and Technology
University, Faculty of Dentistry,
Department of Orthodontics, İstanbul,
Türkiye.

E-mail: buket.erdem@istun.edu.tr



How to cite this article:

Erdem B, Ataş SB, Arslan M, Şar Ç.
Comparative analysis of the
performance of ChatGPT-5.4, Claude
Sonnet 4.6, and Gemini 3.1 Pro large
language models on theoretical
orthodontic questions. Int Dent Res.
2026;16:e260777.
<https://doi.org/10.5577/intdentres.777>

Abstract

Aim: This study aimed to comparatively evaluate the performance of ChatGPT-5.4, Claude Sonnet 4.6, and Gemini 3.1 Pro large language models (LLMs) in answering theoretical orthodontic questions under web search-enabled conditions.

Methodology: A total of 30 open-ended questions covering the topics of growth and development, functional analysis, mixed dentition, extraoral appliances, malocclusion, and orthodontic treatment planning were prepared from Contemporary Orthodontics (7th edition, 2025). The questions were presented to all three models under standardized conditions in separate sessions in March 2026 with web search enabled. The responses were anonymized and independently evaluated in a blinded manner by two orthodontists using a 5-point Likert scale incorporating accuracy, scientific adequacy, completeness, and clarity into a single overall score. Data were analyzed using IBM SPSS Statistics (version 31.0); the Friedman test was used for model comparisons, and statistical significance was set at $p < 0.05$.

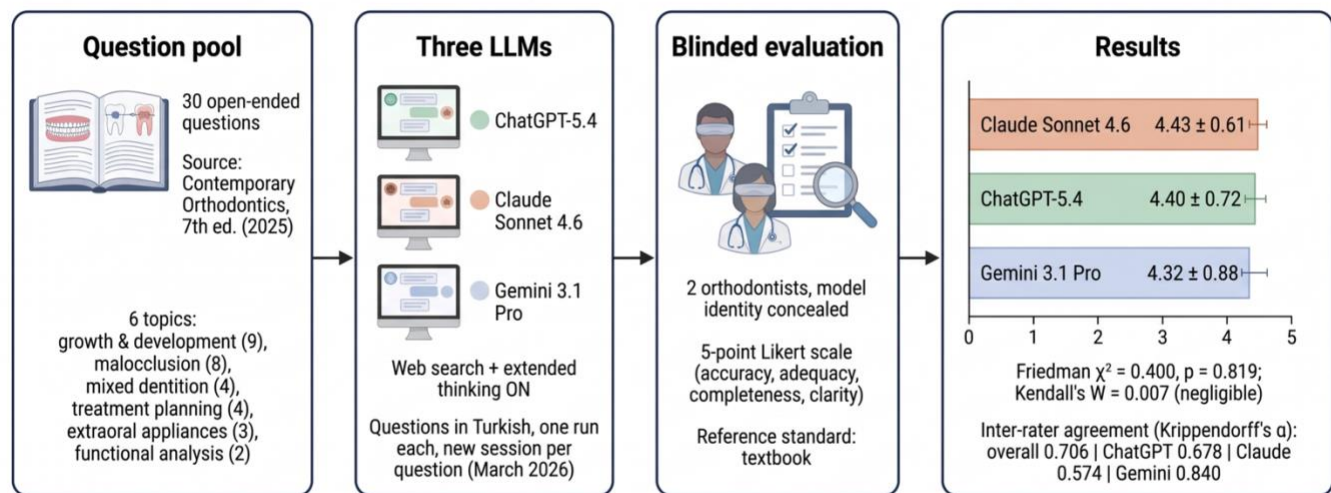
Results: Mean performance scores were 4.40 ± 0.72 for ChatGPT-5.4, 4.43 ± 0.61 for Claude Sonnet 4.6, and 4.32 ± 0.88 for Gemini 3.1 Pro. Median scores were 4.50, 4.50, and 4.75, respectively. No statistically significant difference in overall performance was found among the models ($\chi^2 = 0.400$; $df = 2$; $p = 0.819$; Kendall's $W = 0.007$). Subgroup analyses by topic category also revealed no significant differences in any category (all $p > 0.05$). Inter-rater reliability was limited ($\alpha = 0.706$).

Conclusion: All three models achieved high-performance scores, with no statistically significant difference in overall performance among them; this does not establish equivalence. With independent verification, these LLMs may serve as supportive tools in orthodontic education.

Keywords: Artificial intelligence, ChatGPT, dental education, knowledge assessment, large language models, orthodontics

Graphical abstract

LLM Performance on Theoretical Orthodontic Questions



All three models scored high with no significant difference – but ceiling effect, 30 questions and limited inter-rater reliability mean this is not proof of equivalence. LLMs may support orthodontic education, outputs need independent verification.

Introduction

The increasing adoption of artificial intelligence (AI) technologies in health sciences has fundamentally transformed information access and clinical decision-making (1). In dentistry, AI applications are being used in a wide range of areas, including diagnosis, radiographic image analysis, treatment planning, and patient education (2). Large language models (LLMs) are generative AI systems trained on large text datasets that can perform natural language processing tasks, producing human-like responses in question-answering, information summarization, and content generation (3).

Orthodontics is a comprehensive specialty of dentistry that integrates craniofacial growth and development, functional balance, aesthetics, and biomechanical principles (4). AI applications in orthodontics are used in cephalometric analysis, malocclusion classification, treatment planning, and patient information (5). The growing accessibility of LLM-based systems such as ChatGPT, Gemini, and Claude has prompted increasing interest in their potential applications in orthodontic education and clinical information retrieval. The increasing availability of orthodontic information through digital platforms has raised concerns regarding its accuracy, reliability, and educational value. Previous research has evaluated the

quality and reliability of online orthodontic information, highlighting the importance of critically assessing digital educational resources (6).

Although the number of studies evaluating LLM responses in orthodontics is growing, existing studies have examined applications, including patient education, treatment planning, emergency scenarios, and knowledge assessment (7-11). However, comparative studies that systematically cover the core theoretical topics of orthodontics based on a standardized reference source remain limited. Makrygiannakis et al. (7) compared ChatGPT, Google Bard, and Microsoft Bing in terms of orthodontic knowledge; Erdem et al. (9) examined the performance of four different LLMs in orthodontic emergency scenarios. Patil et al. (11) compared the reliability of four LLMs in orthodontic knowledge assessment and reported the highest accuracy for ChatGPT-4.0 and the highest reliability for Microsoft CoPilot, together with notable variability across repeated trials.

The aim of this study was to comparatively evaluate the performance of ChatGPT-5.4, Claude Sonnet 4.6, and Gemini 3.1 Pro on theoretical orthodontic questions drawn from Contemporary Orthodontics (7th edition, 2025), with the web search and extended-thinking features of each model enabled, and to provide evidence regarding the potential use of these models as supportive tools in orthodontic education.

Materials and Methods

This study is a comparative performance evaluation study. ChatGPT-5.4 (OpenAI, San Francisco, CA, USA), Claude Sonnet 4.6 (Anthropic, San Francisco, CA, USA), and Gemini 3.1 Pro (Google LLC, Mountain View, CA, USA) were evaluated. The model versions selected in each interface at the time of data collection were GPT-5.4 Thinking (ChatGPT), Claude Sonnet 4.6 with extended thinking (Claude), and Gemini 3.1 Pro (Gemini). Reporting followed the Chatbot Assessment Reporting Tool (CHART) statement (12).

Question pool development

The question pool was developed by two experienced orthodontists (B.E., Ç.Ş.) from the reference textbook Contemporary Orthodontics (7th edition, 2025) (4). Questions were selected based on the researchers' clinical and academic experience, with an emphasis on topics likely to appear in specialty examinations. This reference was preferred to represent core orthodontic topics through a standardized textbook, given that specialty examination questions may focus on specific subject areas. No pilot testing was performed. A total of

30 open-ended questions were prepared covering growth and development (n = 9), functional analysis (n = 2), mixed dentition (n = 4), extraoral appliances (n = 3), malocclusion (n = 8), and orthodontic treatment planning (n = 4) (Tables 1 and 2).

Table 1. Distribution of theoretical orthodontic knowledge questions by topic.

Topic	Number of Questions
Growth and development	9
Functional analysis	2
Mixed dentition	4
Extraoral appliances	3
Malocclusion	8
Orthodontic treatment planning	4
Total	30

Table 2. Theoretical orthodontic knowledge questions used in the study.

Topic	Question
Growth and Development	1. What is the cephalocaudal growth gradient?
	2. What do Scammon's curves indicate about the growth of different tissue systems?
	3. How does complete nasal obstruction affect head posture?
	4. How does growth of the cranial base differ from growth of the cranial vault?
	5. Why is the transverse dimension the first to cease in dentofacial growth?
	6. When does secondary palate fusion occur?
	7. What is the functional matrix theory?
	8. What type of ossification occurs without an intermediate cartilage stage?
	9. What are the stages of the adolescent growth period in males and females?
Functional Analysis	10. What three characteristics define infantile swallowing?
	11. What are the effects of digit sucking habits on dentofacial structures?
Mixed Dentition	12. What is a primate space?
	13. What is primary failure of eruption (PFE)?
	14. How much average leeway space is present in the maxillary and mandibular arches?
	15. If jaw growth does not occur anteriorly, where does the extra space come from?

Extraoral Appliances	16. How is anchorage defined in orthodontics?
	17. How do chin-cup appliances affect the restriction of mandibular growth?
	18. What are the effects of applying cervical headgear to a single tooth or the maxilla as a unit?
Malocclusion	19. What is anterior crossbite? What clinical findings distinguish dental from skeletal anterior crossbite?
	20. Define overbite and overjet. State their normal values.
	21. How is posterior crossbite defined?
	22. How is Class II malocclusion defined according to Angle?
	23. What is pseudo-Class III malocclusion?
	24. Why is it recommended to correct dental anterior crossbite in early mixed dentition?
	25. What are the typical characteristics of short-face (skeletal vertical deficiency) cases?
	26. Why is CBCT superior to periapical radiographs in diagnosing impacted maxillary canines?
Orthodontic Treatment Planning	27. What is the fundamental difference in timing between Class II and Class III treatments?
	28. What is the classic serial extraction sequence?
	29. When is early Class II treatment indicated?
	30. Describe the conditions requiring lifelong retention after orthodontic treatment.

Question administration procedure

All questions were administered to each of the three models in March 2026 (20-30 March 2026) via their respective web interfaces using newly created paid accounts (ChatGPT Plus, Claude Pro, and Google AI Pro). ChatGPT-5.4 (OpenAI), Claude Sonnet 4.6 (Anthropic), and Gemini 3.1 Pro (Google LLC) were accessed through their web interfaces with web search and extended thinking features enabled. To ensure standardization across models, all questions were presented in identical wording in Turkish. Each question was posed in an independent session; session history was cleared to eliminate any contextual carryover from previous questions. Models were asked only to answer the question, with no additional guidance or hints provided. Each question was submitted once to each model; no repeated runs were performed. As no human participants, patient data, or animals were involved, ethical committee approval was not required.

Response anonymization and evaluation method

All responses were recorded verbatim, and any information that could reveal the model's identity was

removed prior to evaluation. The three model responses to each question were randomly ordered and presented to the evaluators for scoring. Two orthodontic specialists (B.E., Ç.Ş.) who also developed the question pool independently evaluated the responses in a blinded manner, with the model identities concealed. The corresponding sections of the reference textbook (original English-language 7th edition) were used as the reference standard (4). Because response length, formatting, and citation style could not be fully standardized, complete concealment of the model identity could not be guaranteed; therefore, the word count of each response was calculated post hoc.

Each response was evaluated using a single overall score on a 5-point Likert scale incorporating accuracy, scientific adequacy, completeness, and clarity (1 = Very inadequate, 2 = Inadequate, 3 = Partially adequate, 4 = Adequate, 5 = Complete and correct). The mean of the two evaluators' scores for each question was taken as the final question score.

Statistical analysis

Data were analyzed using IBM SPSS Statistics (version 31.0; IBM Corp., Armonk, NY, USA). Model performance was summarized as mean, standard deviation, and

median. The non-parametric Friedman test was used to compare three dependent samples, applied in both overall and topic-based analyses. Pairwise comparisons with Bonferroni-corrected Wilcoxon signed-rank test were planned if significant. Effect size was assessed using Kendall's W coefficient. Statistical significance was set at $p < 0.05$. No a priori sample size calculation was performed. Post hoc, paired between-model differences were estimated as Hodges-Lehmann median differences and as mean differences with 95% percentile bootstrap confidence intervals (10,000 resamples), and response length was compared between models (Friedman test) and correlated with scores (Spearman's rank correlation); these post hoc analyses were computed in Python (Python Software Foundation, Wilmington, DE, USA; SciPy, version 1.18).

Inter-rater reliability was assessed using the ordinal Krippendorff's alpha coefficient, which was calculated for each model separately and overall across all evaluations. Following Krippendorff, $\alpha \geq 0.800$ was interpreted as reliable, $0.667 \leq \alpha < 0.800$ as permitting only tentative conclusions, and $\alpha < 0.667$ as unacceptable (13).

Results

Overall model performance

The responses of ChatGPT-5.4, Claude Sonnet 4.6, and Gemini 3.1 Pro to the theoretical orthodontic questions were scored by two orthodontic specialists. The mean of the two evaluators' scores for each question was used as the final score.

Descriptive statistics are presented in Table 3. Claude Sonnet 4.6 achieved the highest mean score (4.43 ± 0.61), followed by ChatGPT-5.4 (4.40 ± 0.72) and Gemini 3.1 Pro (4.32 ± 0.88). The median score was 4.50

for ChatGPT-5.4 and Claude Sonnet 4.6 and 4.75 for Gemini 3.1 Pro. Mean response length was 39.9 ± 16.7 words for ChatGPT-5.4, 40.5 ± 15.6 words for Claude Sonnet 4.6, and 40.5 ± 13.4 words for Gemini 3.1 Pro (Friedman $\chi^2 = 0.600$, $p = 0.741$), and response length was not correlated with scores (Spearman $\rho = 0.04$, $p = 0.704$). An example question with the responses of the three models is shown in Figure 1.

Table 3. Descriptive statistics of overall LLM performance.

Model	Mean $\pm$ SD	Median
ChatGPT-5.4	4.40 ± 0.72	4.50
Claude Sonnet 4.6	4.43 ± 0.61	4.50
Gemini 3.1 Pro	4.32 ± 0.88	4.75

SD: standard deviation

The Friedman test was applied to assess whether the performance differences were statistically significant. No statistically significant differences were found among the models ($\chi^2 = 0.400$; $df = 2$; $p = 0.819$). Kendall's W value of 0.007 indicated a negligible effect size (Table 4). As no significant differences were found, pairwise comparisons were not required. To quantify the magnitude of between-model differences, paired differences were also estimated post hoc: the Hodges-Lehmann median difference was 0.00 for all model pairs; the mean paired differences were 0.03 (95% CI -0.17 to 0.25) for Claude Sonnet 4.6 versus ChatGPT-5.4, 0.12 (95% CI -0.17 to 0.43) for Claude Sonnet 4.6 versus Gemini 3.1 Pro, and 0.08 (95% CI -0.10 to 0.30) for ChatGPT-5.4 versus Gemini 3.1 Pro.

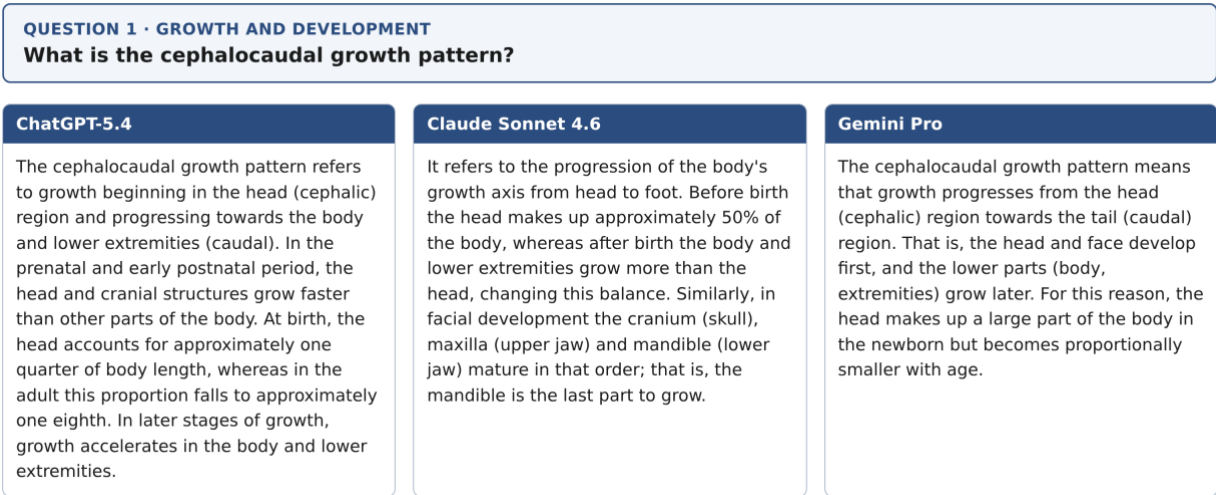


Figure 1. Example of the responses of the three models to Question 1 (growth and development). The question was posed and answered in Turkish; the responses are shown in English translation by the authors. Model names have been added to this figure; the evaluators assessed the responses with model identities concealed.

Table 4. Friedman test comparison of overall LLM performance.

Test	χ^2	df	p	Kendall's W
Friedman	0.400	2	0.819	0.007

χ^2 : chi-square; df: degrees of freedom; Kendall's W: effect size coefficient.

Inter-rater agreement was assessed using the ordinal Krippendorff's alpha coefficient. The overall reliability was $\alpha = 0.706$. The model-based values were $\alpha = 0.678$ for ChatGPT-5.4, $\alpha = 0.574$ for Claude Sonnet 4.6, and $\alpha = 0.840$ for Gemini 3.1 Pro. All pairwise score

differences between the evaluators were at most 1 point; no response showed a discrepancy of 2 or more points. According to the conventional thresholds (13), the overall agreement and agreement for ChatGPT-5.4 fell within the tentative range (0.667-0.800), the agreement for Claude Sonnet 4.6 was below the minimum acceptable level of 0.667, and only the agreement for Gemini 3.1 Pro exceeded 0.800.

Performance by topic category

The performance by topic category is presented in Table 5.

Table 5. Comparison of LLM performance by topic category.

Topic	ChatGPT-5.4 (Mean $\pm$ SD)	Claude Sonnet 4.6 (Mean $\pm$ SD)	Gemini 3.1 Pro (Mean $\pm$ SD)	χ^2	p
Growth and Development	4.22 $\pm$ 1.06	4.39 $\pm$ 0.65	4.33 $\pm$ 1.09	2.000	0.368
Functional Analysis	4.00 $\pm$ 0.00	3.75 $\pm$ 0.35	4.00 $\pm$ 0.00	2.000	0.368
Mixed Dentition	4.38 $\pm$ 0.75	4.63 $\pm$ 0.48	4.25 $\pm$ 0.96	0.286	0.867
Extraoral Appliances	4.33 $\pm$ 0.76	3.83 $\pm$ 0.76	3.83 $\pm$ 1.04	3.200	0.202
Malocclusion	4.44 $\pm$ 0.42	4.50 $\pm$ 0.53	4.50 $\pm$ 0.71	0.720	0.698
Orthodontic Treatment Planning	5.00 $\pm$ 0.00	5.00 $\pm$ 0.00	4.50 $\pm$ 1.00	2.000	0.368
Total	4.40 $\pm$ 0.72	4.43 $\pm$ 0.61	4.32 $\pm$ 0.88	0.400	0.819

SD: Standard deviation. Statistical significance was set at $p < 0.05$.

Discussion

In this study, the performance of ChatGPT-5.4, Claude Sonnet 4.6, and Gemini 3.1 Pro on theoretical orthodontic questions was comparatively evaluated. No statistically significant difference in overall performance was found among the models ($\chi^2 = 0.400$; $p = 0.819$). All three models achieved high mean scores, indicating that they were capable of providing generally high-quality responses to standardized theoretical orthodontic questions under the web search-enabled conditions of the present study. However, the absence of a statistically significant difference should not be interpreted as evidence of comparable or equivalent performance: scores clustered near the upper end of the scale (medians 4.50-4.75), no a priori sample size calculation was performed, the question set was small ($n = 30$), and no equivalence testing was undertaken. The 95% CIs of the paired mean differences extended to approximately 0.4 points on the 5-point scale; therefore,

between-model differences of this magnitude cannot be excluded.

The present findings are partly consistent with yet also meaningfully distinct from those of Makrygiannakis et al., whose study is among the most methodologically comparable investigations in the orthodontic domain (7). Both studies used open-ended questions, involved two blinded orthodontic specialists as evaluators, and applied Friedman's test for model comparisons. However, Makrygiannakis et al. benchmarked responses against evidence-based guidelines and systematic reviews and identified statistically significant differences among models, with Microsoft Bing Chat outperforming both ChatGPT-3.5 and Google Bard (7). In the present study, questions were grounded in a standardized textbook, and no statistically significant differences were observed. This divergence may reflect both the evolution of LLM capabilities between the two study periods and differences in the reference standard and scoring framework. Rubric-based scoring against expert-curated evidence and composite Likert scoring

applied to standardized, textbook-based theoretical questions are methodologically distinct approaches, and their relative sensitivity to inter-model differences cannot be directly compared without further study. These findings suggest that observed differences among LLMs in orthodontic evaluations may vary according to the reference standard, question characteristics, and scoring approach.

Patil et al. addressed the stability of LLM performance in orthodontic knowledge assessments. They compared four LLMs (ChatGPT-4.0, ChatGPT-4o, Google Gemini, and Microsoft CoPilot) using National Board of Dental Examiners orthodontic questions across three repeated testing sessions and reported the highest accuracy for ChatGPT-4.0 and the highest reliability for Microsoft CoPilot (11). They also documented variability across repeated administrations, with older model versions sometimes outperforming newer ones. This finding is relevant to the present study because each model was queried only once per question. The stochastic variability of LLM outputs means that the performance ordering observed in a single testing session cannot necessarily be assumed to remain stable across repeated sessions. Taşyürek et al. evaluated ChatGPT-5, Gemini 2.5 Flash, Grok 4, and DeepSeek-R1-0528 using open-ended questions related to cracked and fractured teeth (14). Although an overall significant difference was observed, post-hoc comparisons showed no statistically significant differences among ChatGPT-5, Gemini 2.5 Flash, and Grok 4; DeepSeek-R1-0528 was the only model that differed significantly and performed less favorably (14). This pattern, in which several models perform similarly while a single weaker model contributes to the overall difference, is broadly consistent with the high and closely clustered performance observed in the present study.

A more complex picture emerges when the present findings are compared with multiple-choice question (MCQ)-based evaluations. Kandemir and Sarıbaş compared four LLMs using periodontology questions from the Turkish Dental Specialty Examination (Diş Hekimliğinde Uzmanlık Sınavı, DUS) and reported statistically significant differences in accuracy, with ChatGPT-4 achieving the highest overall accuracy (73.3%) and Claude achieving the lowest (36.7%) (15). In the present study, by contrast, Claude Sonnet 4.6 achieved the highest mean score (4.43 ± 0.61). This difference may reflect several methodological factors. First, MCQ-based accuracy was assessed as a binary outcome (correct/incorrect), whereas the 5-point Likert scale used in the present study permitted partially adequate responses to receive partial credit. This difference in the scoring structure may affect the ability to distinguish the performance differences among the models. Second, the Claude version evaluated by Kandemir and Sarıbaş (Claude 3 Opus, accessed in July 2025) differs substantially from Claude Sonnet 4.6, which was evaluated in the present study. Although both belong to the same model family, different versions may differ in their capabilities and performance because of differences in model development, training,

optimization, and knowledge updating. Accordingly, comparisons across different model versions should be interpreted cautiously and not generalized to the model family as a whole. More broadly, these findings indicate that model rankings may vary across evaluation contexts, depending on the question format, model version, and scoring approach.

Turan Gökdoğan et al. similarly found comparable overall accuracy rates among ChatGPT-4o, Gemini Advanced, Copilot, and Claude (77.9-83.6%) on Turkish DUS endodontic questions, with no statistically significant between-model differences at the overall level (16). Where significant differences emerged, they were temporal rather than model-specific: Copilot and Claude performed significantly better on questions from 2012-2015 than on those from 2016-2021. The authors attributed this pattern to an increase in premise-based questions in later examination years, which require more multi-step logical inference rather than straightforward knowledge retrieval (16). This finding is relevant to the present study because our question pool was deliberately constructed to assess foundational theoretical orthodontic knowledge. The absence of significant between-model differences observed here may therefore reflect, at least in part, the relatively standardized nature of this type of knowledge assessment. Whether a similar pattern would be observed with more clinically complex or reasoning-intensive orthodontic questions warrants further investigation.

Some studies have reported more pronounced between-model differences. Karatekin found marked differences across LLM versions on Turkish DUS endodontics questions (17), while Taşyürek et al. found ChatGPT-5 to be significantly superior to Gemini 2.5 Flash and Claude Sonnet-4 on open-ended questions concerning endodontic iatrogenic events (18). These differences may reflect variation in the nature and complexity of the evaluated questions, as well as differences in question format, scoring structure, and model versions. Future studies comparing different question formats within the same clinical domain may help clarify the extent to which these methodological factors influence LLM performance.

The use of the reference textbook as the source of the questions provided a standardized, comprehensive, and well-established knowledge base for evaluating all models. This approach established a common reference framework for both the content of the questions and the expected answers. Danesh et al. and Kandemir and Sarıbaş used specialty examination questions (15, 19), whereas Taşyürek et al. similarly based their question pool on a specific textbook chapter (14). In theoretical orthodontic education and examination settings, established textbooks provide an important basis for defining foundational knowledge. Therefore, the use of a standardized textbook can be considered a methodological strength of the present study because it ensured that all models were evaluated within the same knowledge framework and under standardized content conditions.

Model responses were evaluated independently by two orthodontic specialists blinded to the model identity. Because the assessment incorporated accuracy, scientific adequacy, completeness, and clarity, the scoring process necessarily involved an element of subjective judgement. Inter-rater agreement varied across models, with Krippendorff's alpha values ranging from 0.574 for Claude Sonnet 4.6 to 0.840 for Gemini 3.1 Pro, yielding an overall coefficient of 0.706. Agreement for Claude Sonnet 4.6 fell below the minimum acceptable threshold of 0.667, and the overall coefficient reached only the range that permits tentative conclusions (13). Notably, the model with the numerically highest mean score had the least reliable scores. Because the mean differences between models were small relative to the disagreement between evaluators, the numerical ranking of the models should not be regarded as robust. Future studies could further strengthen the evaluation by using separate domain-specific scoring criteria and larger panels of independent orthodontic evaluators.

Although no statistically significant differences were found in any topic-based subgroup analysis, some numerical patterns were observed. Claude Sonnet 4.6 had the highest mean scores in growth and development and mixed dentition, and ChatGPT-5.4 in extraoral appliances. ChatGPT-5.4 and Claude Sonnet 4.6 shared the highest mean score in orthodontic treatment planning, Claude Sonnet 4.6 and Gemini 3.1 Pro in malocclusion, and ChatGPT-5.4 and Gemini 3.1 Pro in functional analysis. These differences did not reach statistical significance, and the subgroup results should be interpreted cautiously because some categories contained very few questions, particularly functional analysis ($n = 2$) and extraoral appliances ($n = 3$). The small subgroup sizes limit the statistical power of these comparisons. Simultaneously, the very small overall effect size observed in the study (Kendall's $W = 0.007$) indicates that the observed between-model differences were minimal at the overall level. Thus, the lack of statistical significance in the subgroup analyses should not be interpreted as definitive evidence of equivalence across the topics.

The high performance scores observed across the models indicate that current LLMs can generate high-quality responses to standardized theoretical orthodontic questions. However, the performance in this setting should not be assumed to extend directly to other applications. Studies involving patient education and orthodontic emergency scenarios have shown that LLM-generated responses may be incomplete or require verification (8, 9). Thus, strong performance on standardized textbook-based questions does not necessarily predict adequate performance in other educational or clinical contexts. In addition, the response variability documented by Patil et al. (11) indicates that a single-session evaluation cannot determine whether the performance order observed in the present study would remain stable across repeated tests. Taşyürek et al. also demonstrated that individual responses may contain errors or hallucinations, even

when aggregate accuracy is high (14). Taken together, these findings support the potential use of ChatGPT-5.4, Claude Sonnet 4.6, and Gemini 3.1 Pro as supportive tools for accessing and reinforcing theoretical knowledge of orthodontics. Nevertheless, their outputs should be independently verified, particularly when the information is used in clinically relevant contexts.

This study has several limitations. Only three LLMs were evaluated; given the rapid pace of AI development, different models or updated versions may perform differently. As the scope of training data and knowledge cutoff dates of LLMs cannot be fully verified, findings should be considered specific to a given time point. Each question was submitted only once to each model, which precluded assessment of stochastic variability in LLM responses; repeated runs (e.g., two to three per question) would be needed to establish whether the observed ordering is stable. Furthermore, the unequal distribution of questions across topics and the very low number of questions in some subgroups (functional analysis: $n = 2$, extraoral appliances: $n = 3$) limit the statistical power of topic-based comparisons. The questions were selected by two experienced orthodontists based on their clinical and academic experience, and no pilot testing or formal content-validation process was performed. Because the two evaluators had also developed the question pool, expectation bias in scoring cannot be excluded. Some questions were phrased in a leading manner (e.g., Q5 presupposes that transverse growth ceases first, and Q26 presupposes the superiority of CBCT), which may have cued the expected answer, and some topic assignments are debatable (e.g., Q16 on anchorage under extraoral appliances and Q26 on CBCT diagnosis under malocclusion). Although model names were removed, differences in response length, formatting, headings, and citation style may have allowed the evaluators to infer model identity, although response length was similar across models and was not correlated with scores. Inter-rater reliability was limited (overall $\alpha = 0.706$; $\alpha = 0.574$ for Claude Sonnet 4.6), which reduces confidence in the between-model comparisons. No a priori sample size calculation was performed, and the ceiling effect together with the small number of questions limited the ability to detect small differences; equivalence was not formally tested. The use of a single composite score rather than separate domain-specific scores may have reduced sensitivity to differences in specific response characteristics. Although the use of a single textbook strengthened standardization, restricting the question pool to one reference source and 30 questions may have limited the breadth of the knowledge domains represented. In addition, web search and extended-thinking features were enabled for all models; therefore, the findings reflect performance under these standardized tool-assisted settings and may not be directly generalizable to evaluations conducted without such features; accordingly, they represent search-assisted performance rather than isolated internal model knowledge.

Conclusion

In this study, the performances of ChatGPT-5.4, Claude Sonnet 4.6, and Gemini 3.1 Pro were comparatively evaluated using 30 open-ended theoretical orthodontic questions drawn from the reference textbook (Contemporary Orthodontics, 7th edition). With web search and extended-thinking features enabled, all three models achieved high mean scores, and no statistically significant difference in overall performance was detected among them. Given the ceiling effect, the limited number of questions, the limited inter-rater reliability, and the absence of equivalence testing, this finding should not be interpreted as evidence of equivalent performance.

Under these search-assisted conditions, LLMs may have potential as supportive tools in orthodontic education for information access, reinforcement of theoretical knowledge, and learning support for students, although their educational utility was not directly assessed in the present study. However, the outputs of these systems should be independently verified, particularly when used to inform clinically relevant decisions.

Future studies using larger question pools, different orthodontic sources, and clinical scenarios are expected to contribute to a more comprehensive understanding of the real potential of LLMs in orthodontics education.

Disclosures

Peer-review: Externally peer-reviewed.

Author Contributions: Conception – B.E., Ç.Ş.; Design – B.E., Ç.Ş.; Supervision – B.E., Ç.Ş.; Materials – Not applicable; Data Collection and/or Processing – S.B.A., M.A.; Analysis and/or Interpretation – B.E., S.B.A.; Literature Review – B.E., S.B.A.; Writing – B.E., S.B.A.; Critical Review – B.E., M.A., Ç.Ş.

Declaration of Generative AI and AI-Assisted Technologies: During the preparation of this work, FigureLabs was used to create the graphical abstract, and the AI figure-generation feature in BioRender was used for subsequent text edits to the figure. The author(s) then reviewed, edited, and verified all visual and scientific content for accuracy and take(s) full responsibility for the final published material.

Conflict of Interest: No conflict of interest relevant to this article was reported.

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data Availability: All model responses and the raw rater scores are available from the corresponding author on reasonable request.

References

1. Agrawal P, Nikhade P. Artificial intelligence in dentistry: past, present, and future. *Cureus*. 2022;14(8):e27405. <https://doi.org/10.7759/cureus.27405>
2. Khanagar SB, Al-Ehaideb A, Maganur PC, et al. Developments, application, and performance of artificial intelligence in dentistry – a systematic review. *J Dent Sci*. 2021;16(1):508-522. <https://doi.org/10.1016/j.jds.2020.06.019>
3. Chakraborty C, Pal S, Bhattacharya M, Dash S, Lee SS. Overview of chatbots with special emphasis on artificial intelligence-enabled ChatGPT in medical science. *Front Artif Intell*. 2023;6:1237704. <https://doi.org/10.3389/frai.2023.1237704>
4. Fields HW, Larson BE, Sarver DM, Proffit WR. *Contemporary Orthodontics*. 7th ed. St Louis: Elsevier; 2025.
5. 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(2):344. <https://doi.org/10.3390/jcm13020344>
6. Cesur E, Tuncer K, Sevgi D, Balaban BC, Arslan C. Evaluation of the quality and reliability of YouTube™ videos created by orthodontists as an information source for clear aligners. *Turk J Orthod*. 2024;37(1):44-49. <https://doi.org/10.4274/TurkJOrthod.2023.2022.127>
7. 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*. 2026;48(1):cjae017. <https://doi.org/10.1093/ejo/cjae017>
8. Vassis S, Powell H, Petersen E, et al. Large-language models in orthodontics: assessing reliability and validity of ChatGPT in pretreatment patient education. *Cureus*. 2024;16(9):e68085. <https://doi.org/10.7759/cureus.68085>
9. Erdem B, Özcan M, Şar Ç. Comparative analysis of artificial intelligence chatbots in orthodontic emergency scenarios: ChatGPT-3.5, ChatGPT-4.0, Copilot, and Gemini. *Angle Orthod*. 2026;96(1):100-105. <https://doi.org/10.2319/021825-146.1>
10. Sokmen T, Kar İ, Tuncer C, Baloş Tuncer B. Comparison of the information of generative artificial intelligence large language models and professional guidelines regarding nutritional advice for orthodontic patients. *BMC Oral Health*. 2025;25(1):1845. <https://doi.org/10.1186/s12903-025-07153-1>
11. Patil S, Eisenhuth G, El-Bialy T, Licari FW. Reliability and performance of four large language models in orthodontic knowledge assessment. *J Dent Educ*. 2026;90(4):462-465. <https://doi.org/10.1002/jdd.14002>
12. CHART Collaborative. Reporting guideline for chatbot health advice studies: the Chatbot Assessment Reporting Tool (CHART) statement. *BMJ Med*. 2025;4(1):e001632. <https://doi.org/10.1136/bmjmed-2025-001632>
13. Krippendorff K. *Content Analysis: An Introduction to Its Methodology*. 2nd ed. Thousand Oaks, CA: Sage; 2004.
14. Taşyürek M, Adıgüzel Ö, Cangül S, Ortaç H. Comparative evaluation of the accuracy and completeness of responses provided by ChatGPT-5, Gemini 2.5 Flash, Grok 4, and DeepSeek-R1 to open-ended questions on cracked and fractured teeth. *J Med Dent Invest*. 2025;6:e250161. <https://doi.org/10.5577/jomdi.e250161>

15. Kandemir M, Sarıbaş EE. Comparative performance of large language models in answering periodontology questions from the Turkish Dental Specialty Examination: a cross-sectional study on accuracy and coverage. *BMC Oral Health*. 2025;25(1):1804. <https://doi.org/10.1186/s12903-025-07241-2>
16. Turan Gökduman C, Arılı Öztürk E, Aktaş Ş, Çanakçı BC. Comparison of chatbots' accuracy in endodontics questions in dentistry specialization exam in Türkiye: ChatGPT-4o, Gemini Advanced, Copilot, and Claude. *BMC Oral Health*. 2026;26(1):28. <https://doi.org/10.1186/s12903-025-07346-8>
17. Karatekin AO. Comparison and review of different versions of OpenAI ChatGPT, Anthropic Claude and Google Gemini large language models' performance on endodontics questions in the Turkish Dentistry Specialization Exam. *Eur J Dent Educ*. 2026. <https://doi.org/10.1111/eje.70121>
18. Taşyürek M, Adıgüzel Ö, Ortaç H. Comparative evaluation of responses from ChatGPT-5, Gemini 2.5 Flash, Grok 4, and Claude Sonnet-4 chatbots to questions about endodontic iatrogenic events. *Healthcare (Basel)*. 2025;13(20):2615. <https://doi.org/10.3390/healthcare13202615>
19. Danesh A, Pazouki H, Danesh K, Danesh F, Danesh A. The performance of artificial intelligence language models in board-style dental knowledge assessment: a preliminary study on ChatGPT. *J Am Dent Assoc*. 2023;154(11):970-974. <https://doi.org/10.1016/j.adaj.2023.07.016>