BMC Oral Health

RESEARCH

Open Access

Performance of large language models in early orthodontic treatment: a comparative study of ChatGPT-5.4 and deepSeek-V3.2

Jialin Xu[1], Masahiro Nakamura[1*], Rui Chen[2], Janvier Habumugisha[1], Gohji Kozaki[1], Ju Zhou[1], Takashi Izawa[1] and Hiroshi Kamioka[1]

Abstract

Background Large language models (LLMs) are increasingly being used by patients and dental professionals to obtain orthodontic information. However, few studies have specifically compared the performance of LLMs in answering questions related to early orthodontic treatment, an area that requires complex clinical reasoning and high-quality patient education. This study aimed to evaluate the clinical applicability of ChatGPT-5.4 and DeepSeek-V3.2 in early orthodontic treatment using a multidimensional evaluation framework.

Methods Five orthodontists developed 20 questions across five domains (age of treatment, adverse oral habits, early treatment, complications, and retention and relapse). The standard answers were derived from clinical guidelines, authoritative textbooks, and articles published in Q1 journal. Each question was submitted once to each LLMs. Responses were anonymised and independently evaluated by five blinded orthodontists. Accuracy and comprehensiveness were assessed using a five-point Likert-type framework combined with visual analogue scale (VAS)-based continuous scoring. Readability was calculated using the SMOG formula. References were cross-checked against five databases and were considered authentic when their bibliographic details corresponded to an indexed publication in at least one database. Thinking time, generation time, and word count were recorded. Inter-rater agreement was assessed using the intraclass correlation coefficient (ICC). Depending on the distribution of paired differences, paired t-tests or Wilcoxon signed-rank tests were applied.

Results ChatGPT-5.4 showed significantly higher accuracy (9.03 ± 0.34 vs. 7.87 ± 0.95, $p < 0.001$) and comprehensiveness (4.49 ± 0.19 vs. 3.87 ± 0.54, $p < 0.001$) than DeepSeek-V3.2. DeepSeek-V3.2 was significantly faster (thinking time: 7.0 vs. 82.9 s; generation time: 39.4 vs. 133.4 s, both $p < 0.001$) and more readable (SMOG grade: 13.48 vs. 14.78, $p = 0.003$). No significant differences were observed for word count ($p = 0.306$) and reference authenticity ($p = 0.895$). In the domain analysis, ChatGPT-5.4 significantly outperformed DeepSeek-V3.2 in the domains of early treatment and retention and relapse ($p < 0.05$), whereas no significant differences were observed in the age of orthodontic treatment, adverse oral habits, and complications domains. Inter-rater agreement was good (ICC = 0.767–0.801).

*Correspondence:
Masahiro Nakamura
nakamurama@okayama-u.ac.jp

Full list of author information is available at the end of the article

© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

Conclusion ChatGPT-5.4 demonstrated greater accuracy and comprehensiveness, whereas DeepSeek-V3.2 generated responses more rapidly and produced more readable text. However, neither model achieved a readability level appropriate for the average patient. Neither model should replace professional clinical judgment.

Keywords Early orthodontic treatment, Large language models, Artificial intelligence, ChatGPT, DeepSeek

Background

The rapid advancement of generative artificial intelligence (Gen AI), particularly large language models (LLMs), has created new opportunities for healthcare communication and clinical decision support in dentistry [1, 2]. Trained on vast text corpora [3], LLMs can generate coherent and contextually appropriate responses to open-ended queries [4]. Their instant responsiveness has transformed information-seeking behaviours, with an increasing number of individuals relying on LLMs for oral health guidance [5, 6].

In orthodontics, LLMs have moved from theoretical exploration to initial clinical applications, mainly as supplementary tools for information retrieval and clinical decision support [7, 8]. Recent research of information retrieval has investigated the performance of LLMs in orthodontic emergencies [9], clear aligner therapy [10], functional appliances [11], early maxillary expansion [12], and general treatment risks [13]. LLMs have also been explored as decision-support tools for hand–wrist radiograph analysis [14], orthodontic aesthetic assessment [15], cervical vertebral maturation assessment [16], and cephalometric diagnosis and treatment planning [17]. These studies demonstrate the increasing role of LLMs in orthodontics, but most have focused on specific treatment modalities or specialised diagnostic tasks. However, there is limited evidence regarding questions in early orthodontic treatment. Moreover, previous studies have generally reported encouraging results in terms of response accuracy and comprehensiveness [18], while raising concerns regarding readability and reference authenticity. Although newer versions of ChatGPT have shown improved reliability, readability remains suboptimal, with Flesch–Kincaid grade levels at the college graduate level [19]. Similarly, only 72.7% of citations generated by LLMs in the natural sciences were reported to be authentic [20]. However, previous evaluation frameworks have typically assessed only a limited range of outcome measures, and few studies simultaneously examined response accuracy, comprehensiveness, readability, reference authenticity, response efficiency, and word count. This limited scope makes it difficult to obtain a comprehensive understanding of LLM performance in early orthodontic treatment.

Early orthodontic treatment is a specialised field within orthodontics. Unlike routine orthodontic care for adults, it requires clinicians to predict craniofacial growth, determine the optimal timing of intervention, and manage of developing malocclusions in growing patients [21]. Appropriate diagnosis and management of dental and maxillofacial morphological and functional abnormalities are therefore essential. These decisions also require patients and parents to have a clear understanding of the treatment and its clinical significance [22]. It is essential that patient-facing educational materials are accurate, reliable, and readily accessible. Although previous studies have evaluated LLM performance in selected treatment modalities within early orthodontic treatment, such as functional appliances and early maxillary expansion, these represent only individual components of clinical care [11, 12]. A comprehensive evaluation of LLM performance covering the broader spectrum of questions in early orthodontic treatment remains lacking, highlighting the need for assessment within this clinical context.

ChatGPT (OpenAI, San Francisco, CA, USA) and DeepSeek (DeepSeek AI, Hangzhou, Zhejiang, China) have attracted considerable research interests because of their different architectural characteristics. In March 2026, OpenAI released a next-generation model named "ChatGTP-5.4", which is pre-trained on larger datasets and fine-tuned with better reinforcement learning from human feedback (RLHF) techniques to improve instruction following, coherence, and alignment with user intent [23, 24]. The so-called "ChatGP-5" model recently reported showed remarkable accuracy and superiority in terms of content comprehensiveness in examinations based on the International Association for Dental Trauma (IADT) guidelines compared to other models such as Gemini 2.5 [25]. DeepSeek-V3.2 is released in December 2025 and adopts a mixture-of-experts (MoE) framework to improve computational efficiency while keeping large parameter capacity [26]. DeepSeek-V3 has shown better diagnostic accuracy on endodontic guideline-based assessments than the dental capacities of ChatGPT-5 with near-perfect repeatability in responses [27]. However, most previous work has been on older versions of LLMs, and there is still little evidence on the efficacy of ChatGPT-5.4 and DeepSeek-V3.2. The two models have different architectures and optimization techniques. These differences may focus on different aspects while generating responses, resulting in different performance profiles, and can be assessed in a complex system of assessment.

Thus, this study was designed to compare the performance of ChatGPT-5.4 and DeepSeek-V3.2 in answering questions related to early orthodontic treatment under

a multidimensional evaluation framework. Their clinical applicability was comprehensively evaluated with respect to response accuracy, comprehensiveness, readability, reference authenticity, generation time, thinking time, and word count. We hypothesised that ChatGPT-5.4 and DeepSeek-V3.2 would demonstrate different performance across multiple evaluation metrics. By characterising the comparative performance of these two LLMs in early orthodontic treatment, this study is expected to provide evidence to support the appropriate integration of these tools in orthodontic patient education and clinical communication.

Methods

Study design and ethical approval

This comparative study was conducted in the Department of Orthodontics, Graduate School of Medicine, Dentistry and Pharmaceutical Sciences, Okayama University, Japan. As the study did not involve human participants, animal, or patient data, ethical approval was no required. The study followed in a four-step workflow (Fig. 1).

Question development

An expert panel comprising five orthodontists from China, Japan, and Rwanda was convened. Open-ended questions included in this study were developed from the Oral Health Foundation website (orthodontic treatment section) [28] and from clinical encounters of the expert panel members. Duplicate, similar, or irrelevant questions were discarded. Through expert discussion and consensus, five predefined domains were established: (1) age of orthodontic treatment; (2) adverse oral habits; (3) early orthodontic treatment; (4) complications; and (5) retention and relapse. Within each domain, the expert panel selected four questions based on their collective clinical experience to represent common patient concerns. The final question set contained 20 questions (Table 1).

Standard answer creation

A standard answer was formulated for each question based on the latest clinical guidelines, expert consensus statements, authoritative textbooks, and articles published in Q1 journals or official journals of orthodontic associations. Each answer was reviewed and discussed by the expert panel until a consensus was reached.

AI response generation

On 8 April 2026, one author submitted each question to both LLMs using the same laptop (MacBook Pro, M1 chip, 8 GB RAM). The models were ChatGPT-5.4 (OpenAI, San Francisco, CA, USA) [29] and DeepSeek-V3.2 (DeepSeek AI, Hangzhou, Zhejiang, China) [30]. Both models were accessed through their official web interfaces (ChatGPT-5.4 via the ChatGPT Plus subscription and DeepSeek-V3.2 via the standard publicly available web interface). To reflect real life usage and reproducibility, the "search" option was disabled for both models to avoid them accessing the web in real time, whereas the default "think" option was enabled. No additional custom instructions, changes to reasoning parameters or

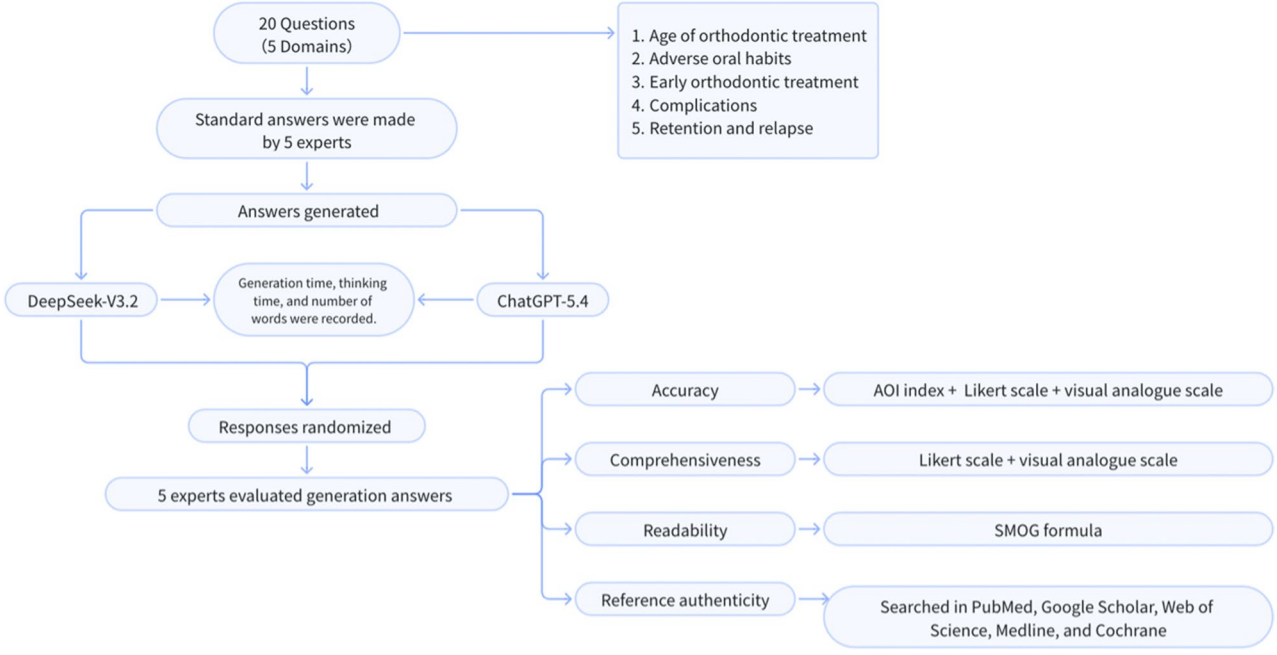

Fig. 1 Flowchart of the study

Table 1 Questions under domains related to early orthodontic treatment

Domains	Questions
Age of orthodontic treatment	1. What is the best age to have an orthodontic evaluation?
	2. What is the best age to treat crossbite in primary dentition?
	3. At what age should functional appliances be started?
	4. At what age can orthodontic sequence extractions be performed?
Adverse oral habits	1. What types of malocclusions can be caused by mouth breathing?
	2. What methods of orthodontic treatment should be used for children with mouth breathing?
	3. What types of malocclusions can be caused by tongue thrusting?
	4. What methods of orthodontic treatment should be used for children with tongue thrusting?
Early orthodontic treatment	1. Does temporary malocclusion in mixed dentition require emergency orthodontic treatment?
	2. What situations require space management?
	3. What are the common treatments for impacted maxillary canines?
	4. Can children or adolescents improve their breathing or airway problems after maxillary expansion?
Complications	1. Can orthodontists directly diagnose obstructive sleep apnoea (OSA) in children?
	2. How to maintain good oral hygiene during orthodontic treatment?
	3. Will orthodontic treatment cause caries?
	4. What are the most common intra-oral risks during early orthodontic treatment?
Retention and relapse	1. What are the reasons for retention after early orthodontic treatment?
	2. If treatment is done while my child has some baby teeth, does that mean s/he is done with treatment?
	3. What are the reasons for relapse after early orthodontic treatment?
	4. What type of retainer is most common after early orthodontic treatment?

different model settings were used; all remaining settings were kept at default values of the platform.

All questions were performed using a Role–Task–Format prompting strategy and the full prompting template was described in **Additional file 1**. The role was described as: "You are an orthodontic expert who has been working for more than 10 years." It was used to establish a consistent professional context for all queries [31]. The task was to "Please answer the following question. Be specific and include any relevant medical guidelines." The format was: "References go at the end." We asked each question only once, to avoid giving the models a chance to refine or alter their answers. The chat window was refreshed after each question to prevent any influence from previous interactions.

The questioning was recorded on video. All responses were presented in the same format, with all words identifying the model were removed. The responses were randomised and entered a Microsoft Excel (Microsoft Corp., Redmond, WA, USA) worksheet for blinded evaluation. To minimise potential biases introduced by human editing, no additional editing was made to the content or writing style of the responses.

Evaluation metrics

Five independent orthodontists evaluated the accuracy and comprehensiveness. All had expertise in clinical orthodontics and international educational background. None of the five evaluators had any prior experience with either of these models, thereby intending to reduce potential for familiarity bias during scoring. A calibration session was conducted prior to scoring to establish a common understanding of the scoring criteria.

Accuracy was assessed using a modified Accuracy of Information (AOI) index combined with a Likert-type scale [18, 32]. Five items of the AOI index were scored: factual accuracy, corroboration, consistency, clarity and specificity, and relevance, respectively (Table 2). Each item was scored from 0 to 2; the total score ranged from 0 to 10 (which represent standard answer). Comprehensiveness was evaluated using a five-point Likert-type scale. The scoring criteria were developed with reference to a previously published framework for LLM responses in dentistry [33] and were refined through expert panel discussion to suit the objectives of the current study (Table 3). Comprehensiveness scores were assigned according to the extent to which each response addressed the key points of the predefined standard answer.

Both accuracy and comprehensiveness were evaluated using a five-point Likert-type framework supplemented by visual analogue scale (VAS)-based continuous scoring. The Likert categories provided the overall scoring framework, whereas the VAS enabled evaluators to assign continuous scores between adjacent categories, thereby improving scoring precision.

Readability was examined using the Simple Measure of Gobbledygook (SMOG) index, a tool often recommended for evaluating health education materials and estimates the educational level necessary to understand written text [34, 35]. The SMOG formula is Reading grade level = 3 + $\sqrt{}$ [(number of words with 3 or more syllables) x (30 / number of sentences)] [36]. One author searched for each cited reference in PubMed, Google Scholar, Web of Science, Medline, and Cochrane databases. A reference was considered authentic if their bibliographic details (including title, authors, journal / publisher, and publication year) corresponded to an indexed publication

Table 2 AOI index with Likert scale for accuracy evaluation

Items	Definition	Parameter	Score
Factual accuracy	The answer is consistent with known facts, data or well-established knowledge on the subject.	Content is accurate with no errors.	2.0
		Core content is accurate with minor deviations.	1.5
		Some content is accurate, but key information exists with errors.	1.0
		Content is mostly inaccurate.	0.5
		Content is completely inaccurate.	0.0
Corroboration	The answer is based on studies, guidelines or textbooks evidence.	Content based on high-quality evidence with no deviation.	2.0
		Content based on moderate evidence with minor deviations.	1.5
		Content based on limited or low-quality evidence.	1.0
		Content lacks sufficient evidence to support.	0.5
		Content contradicts available evidence or professional consensus.	0.0
Consistency	The answer is internally consistent. There are no contradictory statements.	Fully consistent with no contradictory statements.	2.0
		Mostly consistent with minor contradictions.	1.5
		Partially consistent with some contradictions that do not fully undermine the answer.	1.0
		Largely inconsistent with major contradictions that affect the answer's coherence.	0.5
		Completely inconsistent and contradicts itself throughout.	0.0
Clarity and specificity	The answer is clear, specific and free from vague or ambiguous language.	Very clear and specific with no vague or ambiguous language.	2.0
		Clear and specific overall, with only minor ambiguities.	1.5
		Moderately clear, with some vague or ambiguous statements.	1.0
		Often unclear or vague, lacking specificity in key parts.	0.5
		unclear or completely ambiguous and confusing throughout.	0.0
Relevance	The answer directly addresses and adheres to the question asked.	Directly addresses the question and is fully relevant.	2.0
		Mostly addresses the question with only minor irrelevant content.	1.5
		Partially addresses the question with some irrelevant content.	1.0
		Much of the content is irrelevant to the question.	0.5
		Completely irrelevant.	0.0
Total Score			10.0

Table 3 Likert scale for comprehensiveness evaluation

Definition	Parameter	Score
The answer includes the points covered by the standard answer.	All points are mentioned.	5.0
	Most points are mentioned.	4.0
	Basic points are mentioned.	3.0
	Partial points are mentioned.	2.0
	Not many points are mentioned.	1.0

in at least one of these databases. The authenticity rate was calculated as (number of authentic references / total references generated) × 100%. Higher authenticity rates indicated greater citation reliability.

Generation time was measured from the video recordings. To minimise measurement error, the generation time for each response was measured independently three times, and the average was obtained. Thinking time and response word count were also recorded.

Statistical analysis

Statistical analyses were performed using IBM SPSS Statistics version 26.0 (IBM Corp., Armonk, NY, USA). Inter-rater agreement among the five evaluators for accuracy and comprehensiveness was assessed using the intraclass correlation coefficient (ICC), two-way random, absolute agreement. The primary objective of the study was to compare the overall performance of the two LLMs across seven performance metrics. As each performance metric represented a distinct aspect of model performance and was analysed independently. Analyses of individual clinical domains and correlation analyses were exploratory and were performed to identify potential domain-specific differences, cross-model agreement, and associations among performance metrics within each model. Normality of paired differences was assessed using the Shapiro–Wilk test. Normally distributed data were c analysed using paired t-tests; otherwise, Wilcoxon signed-rank tests were applied. Spearman's rank correlation was used to assess cross-model rank-order agreement for each metric and to explore internal correlations of metrics within each model. Descriptive statistics were presented as means and standard deviations. For paired comparisons, mean paired differences with corresponding 95% confidence intervals (CI) were reported. All statistical tests were two-sided, and P values < 0.05 were considered statistically significant.

Results

Inter-rater agreement for accuracy and comprehensiveness

The ICC for accuracy scores was 0.767 (95% CI: 0.664–0.853, $p < 0.001$), and for comprehensiveness scores was 0.801 (95% CI: 0.711–0.876, $p < 0.001$), indicating good inter-rater agreement, as shown in Table 4.

Table 4 Inter-rater agreement analysis for accuracy and comprehensiveness scores

Outcome	ICC (Single Measures)	95% CI	*p*-value
Accuracy	0.767	0.664–0.853	< 0.001**
Comprehensiveness	0.801	0.711–0.876	< 0.001**

ICC values were interpreted as: <0.5 poor, 0.5–0.75 moderate, 0.75–0.9 good, and > 0.9 excellent[37]. P-values were calculated using the intraclass correlation coefficient (two-way random, absolute agreement). *$P < 0.05$, **$P \leq 0.01$

Overall performance comparison

Overall comparisons across the performance metrics are presented in Table 5. ChatGPT-5.4 achieved significantly higher accuracy and comprehensiveness than DeepSeek-V3.2 (both $p < 0.001$). Thinking time and generation time were significantly shorter for DeepSeek-V3.2 (both $p < 0.001$). Regarding readability, higher scores indicate more difficult text that requires higher education levels, while lower scores indicate better readability. DeepSeek-V3.2 exhibited significantly better readability than ChatGPT-5.4 ($p = 0.003$). No significant differences were observed in word count or reference authenticity.

Domain-specific comparisons

Figure 2 showed that ChatGPT-5.4 scored higher than DeepSeek-V3.2 in accuracy scores across all five domains. Detailed statistics are presented in Additional file 2. Significant differences were observed for early orthodontic treatment (9.06 ± 0.20 vs. 7.12 ± 0.50, $p < 0.01$) and retention and relapse (9.00 ± 0.37 vs. 7.37 ± 0.73, $p < 0.05$). Differences in the domains of age, adverse oral habits, and complications were not statistically significant ($p > 0.05$). As shown in Fig. 3, ChatGPT-5.4 consistently achieved higher comprehensiveness scores across all domains. Corresponding mean paired differences and 95% CI are provided in Additional file 3. Significant differences were found for early orthodontic treatment (4.55 ± 0.09 vs. 3.47 ± 0.28, $p < 0.01$) and retention and relapse (4.43 ± 0.24 vs. 3.52 ± 0.46, $p < 0.05$). No significant differences were observed for the other three domains.

Cross-model correlations for individual metrics

Strong positive correlations between the two models were observed for generation time ($\rho = 0.579$, $p < 0.01$) and word count ($\rho = 0.749$, $p < 0.001$). Readability showed a moderate positive correlation ($\rho = 0.494$, $p < 0.05$). No significant correlations were found for accuracy, comprehensiveness, thinking time, or reference authenticity (Fig. 4).

Intra-model correlations of metrics

Within each model, significant positive correlations were observed between accuracy and comprehensiveness (ChatGPT-5.4: $\rho = 0.726$, $p < 0.001$; DeepSeek-V3.2: $\rho = 0.922$, $p < 0.001$) and between generation time and word count (ChatGPT-5.4: $\rho = 0.805$, $p < 0.001$; DeepSeek-V3.2: $\rho = 0.798$, $p < 0.001$), as shown in Fig. 5. Additional significant correlations were model-specific: for ChatGPT-5.4, thinking time versus generation time and thinking time versus word count; for DeepSeek-V3.2, thinking time versus comprehensiveness and generation time versus readability. No other correlations reached statistical significance.

Discussion

This study directly compared ChatGPT-5.4 and DeepSeek-V3.2 in answering open-ended questions about early orthodontic treatment. Overall, ChatGPT-5.4 was more accurate and comprehensiveness, while DeepSeek-V3.2 was faster and produced more readable responses. There were no significant differences in word count used and reference authenticity. These results indicate that the two LLMs exhibit distinct performance profiles, with

Table 5 Overall metrics of generated answers by the ChatGPT-5.4 and DeepSeek-V3.2

Variable	ChatGPT-5.4 mean ± SD (95% CI)	DeepSeek-V3.2 mean ± SD (95% CI)	Mean Difference (95% CI)	*P*-value
Accuracy	9.03 ± 0.34 (8.87–9.19)	7.87 ± 0.95 (7.42–8.31)	1.16 (0.71–1.62)	< 0.001**
Comprehensiveness	4.49 ± 0.19 (4.40–4.57)	3.87 ± 0.54 (3.61–4.12)	0.62 (0.37–0.88)	< 0.001**
Thinking time (s)	82.85 ± 41.64 (63.36–102.34)	7.00 ± 1.52 (6.29–7.71)	75.85 (56.34–95.36)	< 0.001**
Generation time (s)	133.42 ± 55.62 (107.39–159.46)	39.36 ± 8.70 (35.29–43.43)	94.06 (69.93–118.20)	< 0.001**
Readability	14.78 ± 2.14 (13.78–17.78)	13.48 ± 1.22 (12.91–14.05)	1.30 (0.51–2.09)	0.003**
No. of words	690.65 ± 318.97 (541.37–839.93)	741.85 ± 178.78 (658.18–825.52)	-51.20 (-153.11–50.71)	0.306
References (%)	76.97 ± 23.90 (70.08–83.86)	77.73 ± 23.90 (66.54–88.91)	0.76 (-11.06–12.57)	0.895

Values are presented as mean ± standard deviation (95% confidence interval). The normality of paired differences was assessed with the Shapiro-Wilk test. P-values were calculated using the paired t-test for paired comparisons. *$P < 0.05$, **$P \leq 0.01$

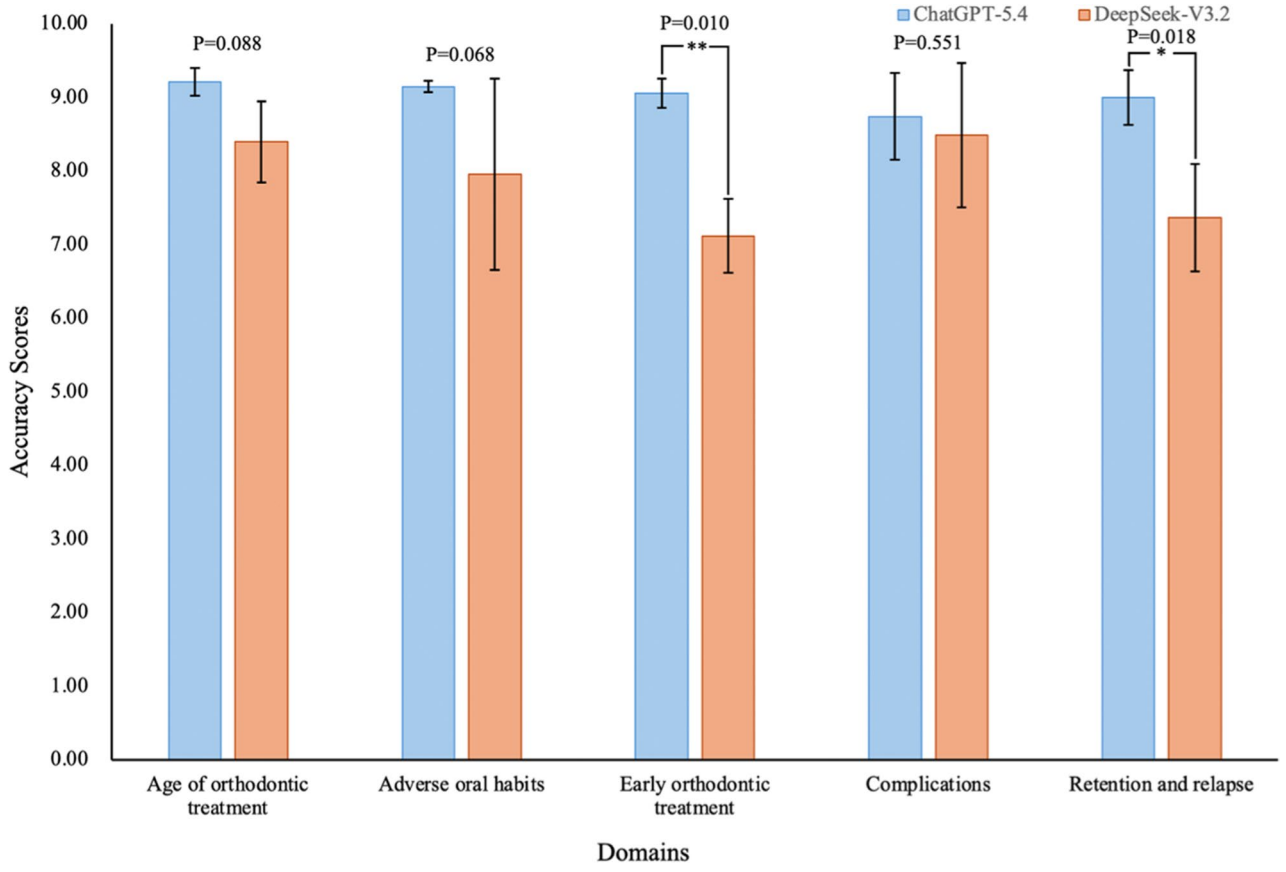

Fig. 2 Comparison of accuracy scores across clinical domains between the two models. These domain-specific comparisons were exploratory in nature and were intended to evaluate potential variations in model performance across clinical domains. *P*-values were calculated using the paired t-test or Wilcoxon signed-rank tests, as appropriate, according to the distribution of paired differences. The adverse oral habits domain was analyzed using the Wilcoxon signed-rank test, and all other domains were analyzed using the paired t-test. Detailed statistics are provided in Additional file 2. $^{*}P < 0.05$, $^{**}P \leq 0.01$

each demonstrating advantages in different aspects of clinical communication.

Our results are in line with previous studies demonstrating encouraging performance of ChatGPT across a range of orthodontic applications, including skeletal maturation assessment, orthodontic aesthetic evaluation, and clinical decision-making [14–17]. Metin and Goymen [38] reported that the ChatGPT-generated predictions were more accurate than those of students and general dentists. Similarly, Alqahtani [39] observed that ChatGPT-4o generated more accurate and comprehensive answers to dental occlusion-related questions. Here, ChatGPT-5.4 also achieved higher accuracy and comprehensiveness when answering questions on early orthodontic treatment, implying superior performance in clinically orientated queries. Such responses may better support patient education and provide dental students with reliable background information, although they should not replace professional judgement. Differences in knowledge cut-off dates have been proposed as one factor influencing LLM performance [40, 41]. In this study, as web search was disabled during testing, both models relied solely on their pre-trained knowledge, which differed between ChatGPT-5.4 and DeepSeek-V3.2 (August 2025 [42] vs. March 2025 [43]). Although this study was not designed to evaluate the impact of knowledge cut-off dates, this difference may have contributed to the observed performance differences and warrants further investigation.

However, a recent study in endodontics reported that DeepSeek-V3 had a higher diagnostic accuracy than ChatGPT-5 using multiple-choice and short-answer knowledge recall tasks [25]. Our task used open-ended questions requiring clinical reasoning. The different results highlight the task dependency of LLMs' performance. The absence of significant cross-model correlations for accuracy and comprehensiveness further suggests the task-dependent performance of LLMs, indicating that questions answered accurately or comprehensively by one model were not necessarily handled similarly by the other.

The overall advantage of ChatGPT-5.4 was not uniform across all clinical domains. Significant differences were observed only in the domains of early orthodontic

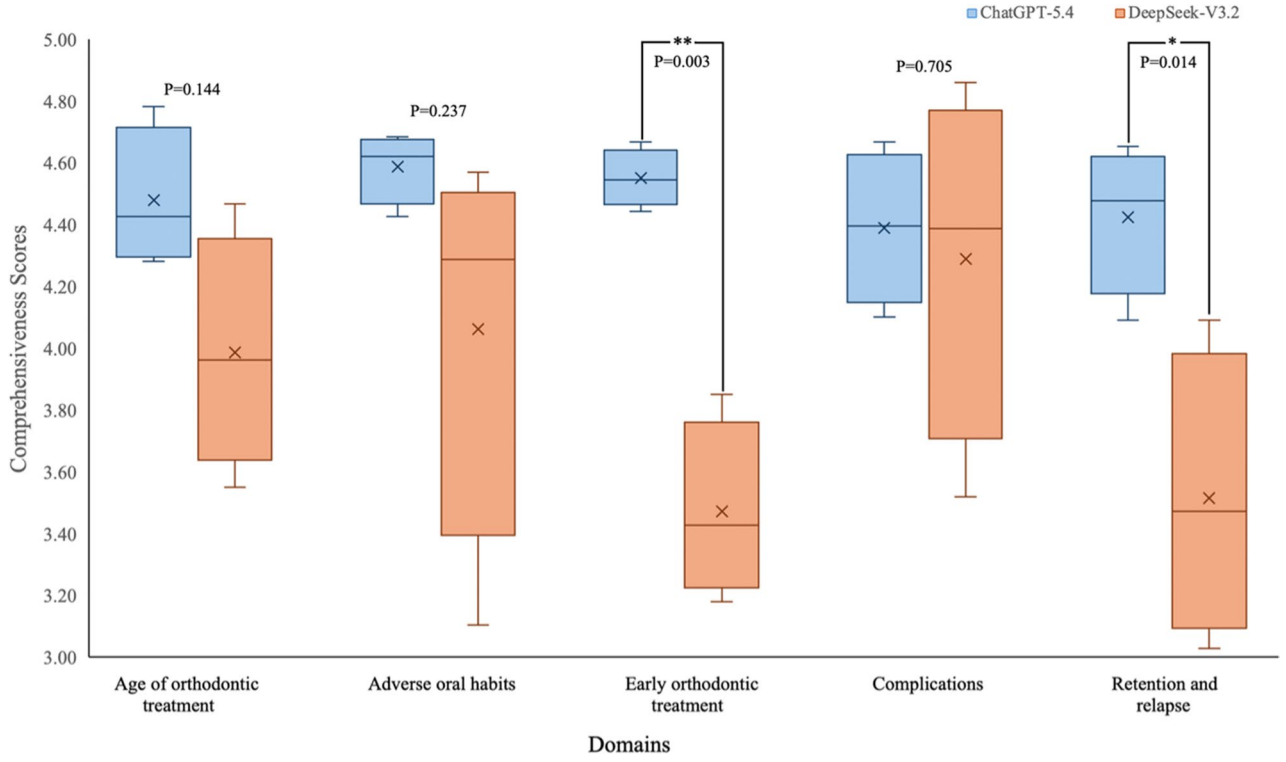

Fig. 3 Comparison of comprehensiveness scores across clinical domains between the two models.These domain-specific comparisons were exploratory in nature and were intended to evaluate potential variations in model performance across clinical domains. *P*-values were calculated using the paired t-test or Wilcoxon signed-rank tests, as appropriate, according to the distribution of paired differences. The age of orthodontic treatment domain was analyzed using the Wilcoxon signed-rank test, and all other domains were analyzed using the paired t-test. Detailed statistics are provided in Additional file 3. **P* < 0.05, ***P* ≤ 0.01

treatment and retention and relapse, which require nuanced judgement regarding skeletal growth prediction, optimal timing of intervention, and long-term stability [22, 44]. On the other hand, smaller differences were found in domains such as age of treatment and complications, which rely more heavily on factual or protocol-based knowledge. The domain-specific results provide useful insights into potential differences in model performance but should be interpreted with caution due to the exploratory nature of these analyses.

DeepSeek-V3.2 provided responses at a much faster rate than ChatGPT-5.4 even with a similar word count, supporting prior findings showing significant differences in response speed among LLMs used in dental applications [45]. Moderate correlations between models in generation time suggest that questions that took longer to generate responses were similarly long in both models. In practical terms, speed may be more important to improve the experience for users such as patients or their guardians seeking quick responses to general orthodontic questions, whereas the quality of responses may be more important in educational or clinical settings where reliability is key.

Overall, the two models showed inconsistent concordance across the performance indicators, with more agreement in response efficiency than response quality. This trend validates prior observations of a trade-off between efficiency and quality of responses in LLMs [46]. The present results are largely consistent with this suggested trade-off. However, our study did not explicitly address the factors that contribute to these performance differences and thus need further exploration.

The intra-model correlation analyses showed a strong positive correlation between accuracy and comprehensiveness within each model. This correlation indicates that more comprehensive responses tended to be associated with higher accuracy scores. Nevertheless correlation does not imply causation [47], and it might mean that accurate answers require more detail in the explanation or that more detailed answers provide more opportunities to include accurate information. Regardless of the underlying mechanism, clinicians and patients should remain aware that no LLM can replace professional judgement [48], especially in the selection of initial orthodontic choices that have long-term developmental consequences. Dental students should apply these models judiciously as educational tools, as the authoritative tone of generated responses may mask nuanced clinical or scholarly inaccuracies [49].

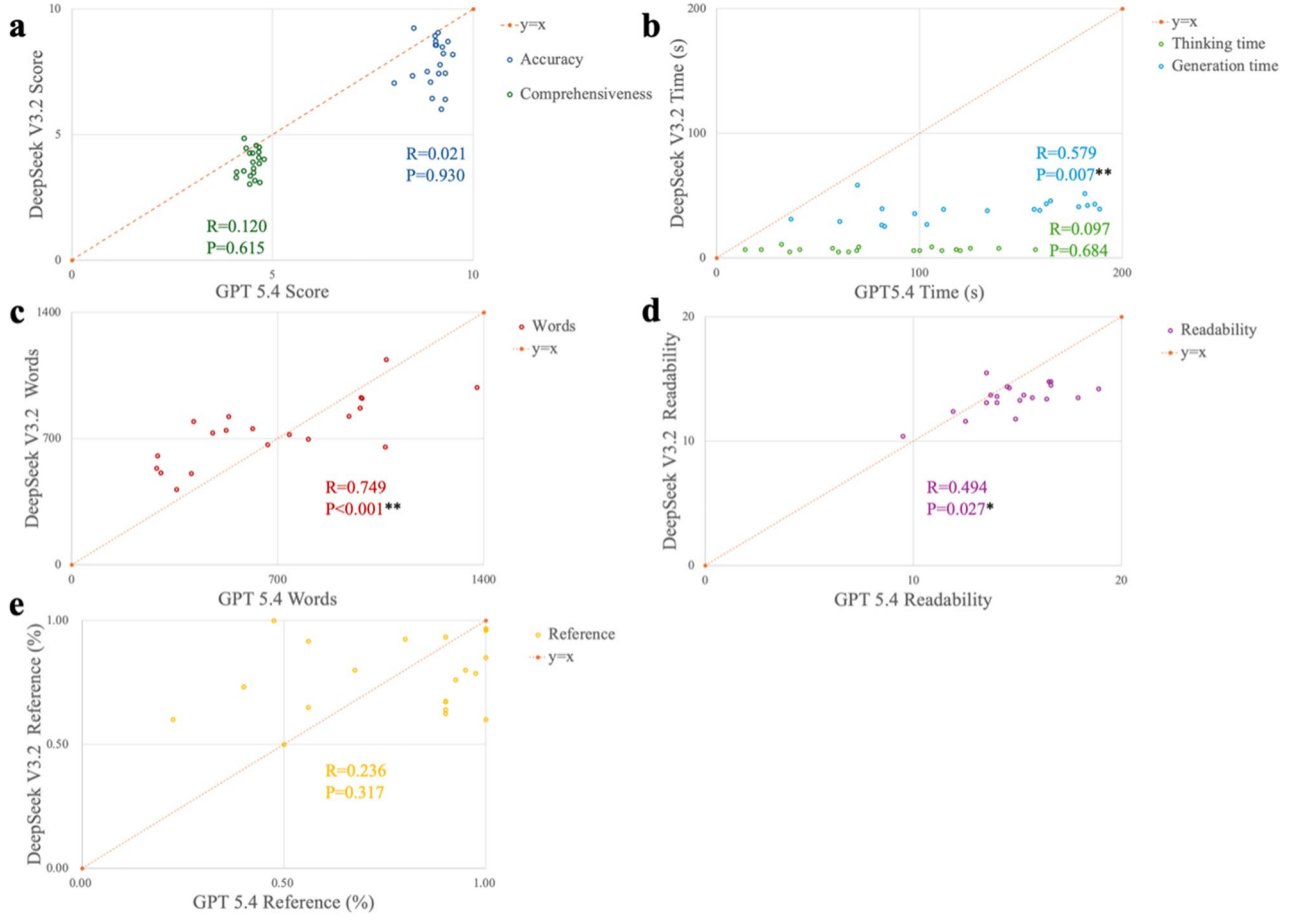

Fig. 4 Pairwise correlation of the same metrics between the two models. These correlation analyses were exploratory and were intended to evaluate cross-model associations. (**a**) Scatter plot of correlation in accuracy and comprehensiveness scores between ChatGPT-5.4 and DeepSeek-V3.2. (**b**) Scatter plot of correlation in thinking time and generation time between ChatGPT-5.4 and DeepSeek-V3.2. (**c**) Scatter plot of correlation between ChatGPT-5.4 and DeepSeek-V3.2 in the number of words generated per question. (**d**) Scatter plot of the correlation of readability between ChatGPT-5.4 and DeepSeek-V3.2. (**e**) Scatter plot of correlation of authenticity of references between ChatGPT-5.4 and DeepSeek-V3.2. Paired scatter plots for all metrics use y = x as a reference line, indicating that the two models have the same score on the same problem. P-values were calculated using Spearman's rank correlation to assess cross-model rank-order agreement for each metric. *$P < 0.05$, **$P \leq 0.01$

Both models generated text with reading grade levels far above the recommended standard for patient education materials [50], consistent with prior research in orthodontics [19]. A very recent study assessing various LLMs for patient education regarding early maxillary expansion also found readability to be an ongoing limitation, despite progress in the quality of responses [12]. This persistent readability gap indicates that LLMs are not yet suitable for direct patient use without further simplification. Similarly, neither model demonstrated fully reliable reference generation. Although the authenticity rate has improved compared to earlier reports of widespread citation fabrication [51–53], a notable proportion of references remain unverifiable, highlighting the persistent problem of citation hallucination in clinical contexts. Taken together, these findings showed that current LLMs still require careful human oversight before being used as patient-facing educational or clinical decision-support tools.

Unlike previous studies focusing primarily on image-based orthodontic tasks, such as hand–wrist radiographs, cervical vertebral maturation, or aesthetic assessment [14–16], this study evaluated open-ended clinical reasoning in early orthodontic treatment. Such questions require interpretation of growth patterns, treatment timing, and long-term developmental outcomes. The current study also provides a direct comparison between the open-source DeepSeek-V3.2 and the closed-source ChatGPT-5.4 in early orthodontic treatment, enabling evaluation of two widely used models with different architectures. Additionally, the inclusion of evaluators from three countries was intended to minimise the potential cultural and educational bias during the assessment. Predefined standard answers were derived from clinical protocols, reputable literature, and



Fig. 5 Exploratory correlation of performance metrics within one model. These intra-model correlation analyses were exploratory in nature and were intended to identify potential associations between performance metrics. (**a**) Correlation heatmap of internal metrics in ChatGPT-5.4. (**b**) Correlation heatmap of internal metrics in DeepSeek-V3.2. Darker shades of red indicate stronger positive correlations, whereas darker shades of blue indicate stronger negative correlations; white means no correlation. P-values were calculated using Spearman's rank correlation to explore internal metric correlations within each model. *$P < 0.05$, **$P \leq 0.01$

expert agreement to make the assessment more uniform and clearer. A further benefit of this work was the use of a standardised prompting strategy and identical model settings, which helped ensure methodological consistency and were aimed to minimise performance differences attributable to variations in prompting rather than model capability. Disabling web search focused on intrinsic knowledge of each model, rather than real-time information retrieval. Taken together, these methodological features set up a structured framework for the research.

There are several limitations to be considered. Initially, only 20 questions (four questions per domain) were employed, with each submitted only once to each LLM. The limited number of questions, together with the absence of a formal sample size calculation, may restrict the statistical power and generalisability of the findings. Meanwhile, due to the stochastic nature of LLMs, evaluating only one response per question may not sufficiently capture the consistency or variation in output of models. Secondly, all questions were presented in English. DeepSeek was trained on bilingual data and thus its performance in Chinese may vary, limiting the generalizability of the findings to multilingual clinical settings. Third, this study evaluated both LLMs at a single time point (April 2026). Given the rapid evolution of LLMs, performance may change over time, so the data may not accurately reflect future model performance. The final limitation is that the same expert panel developed both questions and standard answers, which may have introduced incorporation or review bias. Although accuracy and comprehensiveness were assessed using predefined scoring criteria and standard answers, some degree of subjective judgement was unavoidable despite the good inter-rater agreement. Moreover, the responses were not evaluated by lay users, such as patients or parents, limiting insight into how non-expert users interpret, understand, or apply LLM-generated information in real-world settings.

Future studies should include larger set of questions, repeated prompts, multilingual evaluations, and longitudinal monitoring as LLMs evolve. They should also incorporate independent evaluators and lay-user assessments,

in order to minimize subjectivity and evaluate the comprehensibility, perceived usefulness, and practical applicability of LLM-generated orthodontic information.

Conclusion

ChatGPT-5.4 demonstrated higher accuracy and comprehensiveness than DeepSeek-V3.2 in answering open-ended questions on early orthodontic treatment, whereas DeepSeek-V3.2 generated faster responses with better readability. Both LLMs exhibit different performance characteristics rather than one model being universally superior across all evaluated dimensions. These results emphasise the importance of evaluating multiple dimensions of LLM performance when assessing their application in early orthodontics. Although LLMs show considerable potential as supplementary information sources for early orthodontic treatment, limitations in readability and reference authenticity remains. Therefore, they should be used as supportive tools rather than standalone sources of clinical information, and their outputs interpreted critically under appropriate professional oversight.

Abbreviations

Gen AI	Generative Artificial Intelligence
LLMs	Large Language Models
RLHF	Reinforcement Learning from Human Feedback
IADT	International Association of Dental Traumatology
MoE	Mixture-of-Experts
AOI	Accuracy of Information
VAS	Visual analogue scale
SMOG	Simple Measure of Gobbledygook
ICC	Intraclass Correlation Coefficient
CI	Confidence Interval

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12903-026-09662-z.

Supplementary Material 1.

Supplementary Material 2.

Supplementary Material 3.

Acknowledgements

Not applicable.

Authors' contributions

J.X. conceived and designed the study, collected the question set, developed the standard answers, evaluated LLMs' responses, and drafted the manuscript. M.N. improved the question set, discussed and refined the standard answers, evaluated LLMs' responses, and contributed to the interpretation and revision of the manuscript. R.C. collected the LLMs responses, collected data, and analysed data. J.H. improved the question set, discussed and refined the standard answers, and evaluated LLMs' responses. G.K. evaluated LLMs' responses. J.Z. improved the question set, discussed and refined the standard answers, and evaluated LLMs' responses. T.I. improved the question set and discussed and refined the standard answers. H.K. provided design guidance and technical support. All authors reviewed the manuscript.

Funding

Not applicable.

Data availability

The datasets generated and/or analysed during the current study (files containing the standard answers and LLM responses) are not publicly available. However, they can be obtained from the corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Conflict of interest

The authors declare that they have no competing interests.

Competing interests

The authors declare no competing interests.

Author details

[1]Department of Orthodontics, Graduate School of Medicine, Dentistry and Pharmaceutical Sciences, Okayama University, Okayama 700-8525, Japan
[2]Department of Stomatology, General Hospital of Northern Theatre Command, Shenyang 110001, Liaoning, People's Republic of China

Received: 3 June 2026 / Accepted: 14 August 2026
Published online: 20 August 2026

References

1. Eggmann F, Weiger R, Zitzmann NU, Blatz MB. Implications of large language models such as ChatGPT for dental medicine. J Esthet Restor Dent. 2023;35(7):1098–102.
2. 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(1):29.
3. Kumar P. Large language models (LLMs): survey, technical frameworks, and future challenges. Artif Intell Rev. 2024;57(10):260.
4. Sarker IH, AI-Based Modeling. Techniques, Applications and Research Issues Towards Automation, Intelligent and Smart Systems. SN Comput Sci. 2022;3(2):158.
5. Shah C, Bender EM. Envisioning Information Access Systems: What Makes for Good Tools and a Healthy Web? ACM Trans Web. 2024;18(3):1–24.
6. Li H, Chen S, Chang B, Wang X, He Y, Xu B, Sun G, Yang C, Li G, Li S, et al. Application of artificial intelligence in oral health management: challenges and opportunities. Front Med (Lausanne). 2026;13:1700529.
7. Baxmann M, Kárpáti K, Baráth Z. The potentials and challenges of integrating generative artificial intelligence (AI) in dental and orthodontic education: a systematic review. BMC Oral Health. 2025;25(1):905.
8. Tuygunov N, Samaranayake L, Khurshid Z, Rewthamrongsris P, Schwendicke F, Osathanon T, Yahya NA. The transformative role of artificial intelligence in dentistry: a comprehensive overview part 2: the promise and perils, and the international dental federation communique. Int Dent J. 2025;75(2):397–404.
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. 2025;96(1):100–5.
10. Abu Arqub S, Al-Moghrabi D, Allareddy V, Upadhyay M, Vaid N, Yadav S. Content analysis of AI-generated (ChatGPT) responses concerning orthodontic clear aligners. Angle Orthod. 2024;94(3):263–72.
11. Badran SA, Pattanaik S, Jumaah S, Salmeh A, Al-Saadi M, Al-Mizban T, Al-Zaidi N, Hanifa A, Al-Omiri MK. Accuracy, readability, and content coverage of AI-generated responses to questions on functional appliances. BMC Oral Health 2026.
12. Cicek O, Yıldırım A. Evaluating the reliability of artificial intelligence in clinical decisions on early maxillary expansion: a comparative study of eight large language models. BMC Oral Health. 2026;26(1):715.
13. Fan Z, Lei J, Shi W, Lin Y, Wang Q, Bao L. Can AI chatbots accurately provide information on orthodontic risks? Angle Orthod. 2025;95(5):483–9.

14. Yıldırım A, Cicek O, Genç YS. Can AI-Based ChatGPT Models Accurately Analyze Hand-Wrist Radiographs? A Comparative Study. Diagnostics (Basel). 2025;15(12):1513.
15. Yıldırım A, Cicek O. Assessment of AI-Driven Large Language Models for Orthodontic Aesthetic Scoring Using the IOTN-AC. Diagnostics (Basel). 2025;15(23):3048.
16. Erdem R, Yıldırım A, Genç YS, Beşer Gül B, Naralan ME, Cicek O. Comparative performance analysis of AI-based large language models in assessing cervical vertebral maturation stages on lateral cephalometric radiographs. BMC Oral Health. 2026;26(1):979.
17. Yıldırım A, Cicek O, Genç YS, Erdem R. Comparative analysis of generative pre-trained transformers for text- and image-based cephalometric prompts using a novel Artificial Intelligence Based Diagnosis and Treatment Planning Index (AIDTI). BMC Med Inf Decis Mak 2026.
18. Daraqel B, Wafaie K, Mohammed H, Cao L, Mheissen S, Liu Y, Zheng L. The performance of artificial intelligence models in generating responses to general orthodontic questions: ChatGPT vs Google Bard. Am J Orthod Dentofac Orthop. 2024;165(6):652–62.
19. Kılınç DD, Mansız D. Examination of the reliability and readability of Chatbot Generative Pretrained Transformer's (ChatGPT) responses to questions about orthodontics and the evolution of these responses in an updated version. Am J Orthod Dentofac Orthop. 2024;165(5):546–55.
20. Mugaanyi J, Cai L, Cheng S, Lu C, Huang J. Evaluation of Large Language Model Performance and Reliability for Citations and References in Scholarly Writing: Cross-Disciplinary Study. J Med Internet Res. 2024;26:e52935.
21. Proffit WR, Fields H, Larson B, Sarver DM. Contemporary orthodontics. 6th ed. Philadelphia, PA: Elsevier; 2018.
22. Zhou C, Duan P, He H, Song J, Hu M, Liu Y, Liu Y, Guo J, Jin F, Cao Y, et al. Expert consensus on pediatric orthodontic therapies of malocclusions in children. Int J Oral Sci. 2024;16(1):32.
23. Dalvi PK, Digholkar KY. RLHF: Reinforcement Learning using Human Feedback for Optimization of ChatGPT. Grenze Int J Eng Technol (GIJET) 2024, 10.
24. GPT-5.4. makes a stunning debut. https://openai.com/zh-Hans-CN/index/introducing-gpt-5-4/. Accessed 5 March, 2026.
25. Er Öİ, Adıgüzel Ö, Taşyürek M, Ortaç H. Comparison of ChatGPT-5, Gemini 2.5 Pro, and DeepSeek-V3. 1 chatbot responses to dental avulsion injuries. Health Sci Monit. 2025;3(1):e251201–e.
26. Liu A, Mei A, Lin B, Xue B, Wang B, Xu B, Wu B, Zhang B, Lin C, Dong C. Deepseek-v3. 2: Pushing the frontier of open large language models. arXiv preprint arXiv:251202556 2025.
27. Demir Cicek B, Cicek O. Evaluating the Response of AI-Based Large Language Models to Common Patient Concerns About Endodontic Root Canal Treatment: A Comparative Performance Analysis. J Clin Med. 2025;14(21):7482.
28. Oral health information. https://www.dentalhealth.org/Listing/Category/all-oral-health-information?Take=92. Accessed 20 November, 2025.
29. ChatGPT. https://chatgpt.com/ Accessed 8 April, 2026.
30. DeepSeek. https://chat.deepseek.com/. Accessed 8 April, 2026.
31. White J, Fu Q, Hays S, Sandborn M, Olea C, Gilbert H, Elnashar A, Spencer-Smith J, Schmidt DC. A prompt pattern catalog to enhance prompt engineering with chatgpt. arXiv preprint arXiv:230211382 2023.
32. Erkan AEG, Arslan AB. Evaluation of the responses from different chatbots to frequently asked patient questions about impacted canines. Australasian Orthodontic J. 2025;41(1):288–300.
33. Dermata A, Arhakis A, Makrygiannakis MA, Giannakopoulos K, Kaklamanos EG. Evaluating the evidence-based potential of six large language models in paediatric dentistry: a comparative study on generative artificial intelligence. Eur Arch Paediatr Dent. 2025;26(3):527–35.
34. Wang LW, Miller MJ, Schmitt MR, Wen FK. Assessing readability formula differences with written health information materials: application, results, and recommendations. Res Social Adm Pharm. 2013;9(5):503–16.
35. Rudd RE, Anderson JE. The health literacy environment of hospitals and health centers. *National Center for the Study of Adult Learning and Literacy and Health And Adult Literacy and Learning Initiative, Harvard School of Public Health* http://wwwhsphharvardedu/wp-content/uploads/sites/135/2012/09healthliteracyenvironmentpdf,erişim:01042014) 2006.
36. Mc Laughlin GH. SMOG grading-a new readability formula. J Read. 1969;12(8):639–46.
37. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med. 2016;15(2):155–63.
38. Metin U, Goymen M. Information from digital and human sources: A comparison of chatbot and clinician responses to orthodontic questions. Am J Orthod Dentofac Orthop. 2025;168(3):348–57.
39. Alqahtani H. Assessment of artificial intelligence chatbots in responding to dental occlusion questions: a comparative study. BMC Oral Health. 2025;26(1):201.
40. Asai M, Lin A, Kishimoto Y, Obi T, Kosugi S, Funakoshi K, Okumura M. Can LLMs Be Constrained to the Past? Improving Knowledge Cutoff through Recall-Based Prompting. arXiv preprint arXiv:260605804 2026.
41. Chen Y, Zhao J, Wen Z, Li Z, Xiao Y, Temporalmed. Advancing medical dialogues with time-aware responses in large language models. In: *Proceedings of the 17th ACM International Conference on Web Search and Data Mining: 2024*; 2024: 116 – 24.
42. ChatGPT models. https://developers.openai.com/api/docs/models. Accessed 3 March, 2026.
43. DeepSeek, March, V3.2. https://docs.aws.amazon.com/bedrock/latest/userguide/model-card-deepseek-deepseek-v3-2.html. Accessed 2 2026.
44. Bahreman AA. Retention considerations in the assessment of long-term stability in early versus late orthodontic treatment. In: *Seminars in Orthodontics: 2017*: Elsevier; 2017: 123–48.
45. Independent analysis of AI. https://artificialanalysis.ai/. Accessed 20 July, 2026.
46. Van Der Veer SN, Riste L, Cheraghi-Sohi S, Phipps DL, Tully MP, Bozentko K, Atwood S, Hubbard A, Wiper C, Oswald M. Trading off accuracy and explainability in AI decision-making: findings from 2 citizens' juries. J Am Med Inform Assoc. 2021;28(10):2128–38.
47. Aslam A. Research ideas: Correlation does not imply causation. Br Dent J. 2015;219(2):49.
48. Genovese A, Hegstrom L, Prabha S, Gomez-Cabello CA, Haider SA, Collaco B, Wood NG, Forte AJ. Artificial Authority: The Promise and Perils of LLM Judges in Healthcare. Bioengineering. 2026;13(1):108.
49. Thurzo A, Strunga M, Urban R, Surovková J, Afrashtehfar KI. Impact of artificial intelligence on dental education: a review and guide for curriculum update. Educ Sci. 2023;13(2):150.
50. Rooney MK, Santiago G, Perni S, Horowitz DP, McCall AR, Einstein AJ, Jagsi R, Golden DW. Readability of patient education materials from high-impact medical journals: a 20-year analysis. J patient experience. 2021;8:2374373521998847.
51. Walters WH, Wilder EI. Fabrication and errors in the bibliographic citations generated by ChatGPT. Sci Rep. 2023;13(1):14045.
52. Lingard L. Writing with ChatGPT: An illustration of its capacity, limitations & implications for academic writers. Perspect Med Educ. 2023;12(1):261.
53. Emsley R. ChatGPT: these are not hallucinations–they're fabrications and falsifications. Schizophrenia. 2023;9(1):52.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.