Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

AI-based chatbots and orthodontic retention: A comparative analysis of response quality and readability

Yağmur Lena Sezici

yagmurlena.sezici@tinaztepe.edu.tr

Izmir Tınaztepe University

Genta Agani Sabah

Izmir Tınaztepe University

Research Article

Keywords: Artificial intelligence, Chatbots, Orthodontic retention, Information quality, Patient education

Posted Date: May 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9495119/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

To evaluate and compare the performance of three artificial intelligence (AI) chatbots in generating responses to orthodontic retention-related questions in terms of clinical quality and readability.

Methods

A cross-sectional content analysis was conducted on the responses to orthodontic retention- related questions by ChatGPT-5.4 (OpenAI LLC, San Francisco, Calif), Gemini 3.1 Pro (Alphabet Inc, Mountain View, Calif), and Copilot (Microsoft, Redmond, Wash). 25 clinically relevant orthodontic retention questions were selected by two experienced orthodontists from an initial pool of 50 questions and grouped into eight domains. Five questions were repeated for each chatbot to assess consistency, yielding 90 responses in total (30 per chatbot). Responses were anonymized and independently evaluated by three blinded orthodontists using a five-point Likert scale for clinical quality and completeness. Readability was assessed using the Flesch Reading Ease (FRE) and Flesch-Kincaid Grade Level (FKGL). Statistical analyses included weighted and Fleiss' kappa, Kruskal-Wallis H test and Spearman correlation.

Results

Intra-rater reliability ranged from substantial to almost perfect ($\kappa > 0.737$), and overall inter-rater reliability demonstrated substantial agreement ($\kappa = 0.781$) ($p < 0.001$). Overall chatbot performance differed significantly ($p = 0.019$), with ChatGPT outperforming Copilot (adjusted $p = 0.005$), while Gemini did not differ significantly from either model ($p > 0.017$). No differences were identified among chatbots within individual orthodontic retention domains ($p > 0.05$). No difference in FRE and FKGL scores was found among the chatbots ($p > 0.05$), with FRE scores indicating difficult readability, while FKGL scores corresponded to a high school- to college-level reading grade. Clinical quality of the chatbot responses was not correlated with either FRE ($\rho = -0.102$, $p = 0.386$) or FKGL ($\rho = 0.113$, $p = 0.336$).

Conclusion

AI chatbots demonstrated generally good performance in generating responses to orthodontic retention-related questions, with comparable performance across clinical domains. However, all responses were written at a readability level exceeding recommended thresholds for patient education materials, which may limit their practical usefulness. Although chatbots may serve as accessible sources of information, they should be used with caution and should not replace professional orthodontic guidance.

Introduction

Orthodontic retention represents a fundamental phase of treatment aimed at maintaining the stability of tooth alignment and occlusion achieved during active orthodontic treatment [1]. Following appliance removal, teeth exhibit a natural tendency to return toward their original positions, a phenomenon known as relapse, which can occur at any age [2]. Therefore, retention protocols are routinely implemented in nearly all orthodontic patients to maintain the results of orthodontic treatment [3, 4]. However, despite its universal application, there remains no consensus regarding the optimal retention strategy, including the type of retainer, duration of wear, or follow-up protocols [5]. The variability in retention approaches is influenced by multiple factors, including clinician preference, patient-specific characteristics, and the lack of strong evidence-based guidelines. Consequently, retention protocols often differ widely between practitioners and regions, and are frequently based on clinical experience rather than standardized recommendations [4].

In clinical practice, orthodontic treatment is typically characterized by regular and consistent clinical appointments during the active phase. In contrast, the retention phase is often associated with longer follow-up intervals, reduced supervision, and, in some cases, complete discontinuation of routine visits by patients [6]. This reduced level of professional monitoring places greater responsibility on patients to manage their retainers, adhere to instructions, and recognize potential complications. As a result, patients frequently seek additional sources of information outside the clinical setting, particularly through digital platforms. The increasing accessibility of online resources has transformed the way patients obtain health-related information, with a substantial proportion of individuals turning to the internet as their primary source of guidance [7]. In addition to the accuracy and completeness of information, readability is a critical determinant of patient comprehension and adherence, as health-related materials are often written above recommended literacy levels, potentially limiting their accessibility to the general population.

More recently, advances in artificial intelligence (AI), particularly in chatbot technology, have established AI-based chatbots as a new and rapidly growing source of medical information [8]. AI-driven technologies have become integral to modern dental practice, contributing to improved diagnostic precision, more effective treatment planning, and enhanced patient management, ultimately increasing the overall efficiency of care delivery across dental disciplines [9]. Chatbots operate through advanced human-computer interaction mechanisms and are specifically designed to generate conversational responses that mimic natural communication with users in an online environment [10–13]. Furthermore, these models are capable of generating instant, conversational responses to user queries and are increasingly being used by patients to obtain health-related advice. Their accessibility, immediacy, and ease of use make them especially appealing during periods when direct professional consultation is limited [14]. Although previous studies have demonstrated that chatbots can generate generally accurate responses to orthodontic queries [15–20], their performance remains inconsistent, and concerns persist regarding reliability, clinical validity, and completeness of the information provided. Furthermore, there is

still a limited body of evidence specifically focusing on the retention phase, which represents a critical period in which improper guidance may directly compromise long-term treatment outcomes.

Given that the retention phase represents the period of orthodontic treatment with the least frequent professional supervision, the availability of accurate and understandable information becomes particularly important for ensuring patient adherence and maintaining treatment stability. Previous studies have evaluated AI-based chatbot performance in orthodontic retention using structured scoring systems and patient-oriented metrics [24], and have also highlighted the importance of patient compliance and its relationship with long-term stability [31]. However, existing research has primarily focused on patient-directed outputs or predictive modeling approaches, with limited emphasis on clinically oriented evaluation frameworks that incorporate expert-based assessment, response consistency, and domain-specific performance analysis. Therefore, the aim of the present study was to evaluate and compare the clinical quality and readability of responses generated by three widely used chatbots, ChatGPT-5.4, Google Gemini, and Microsoft Copilot, to frequently asked questions related to orthodontic retention. The null hypothesis was that no significant differences would be observed among the chatbots in terms of response quality.

Materials and Methods

This study was designed as a cross-sectional content analysis aimed at evaluating the quality of responses generated by three widely used chatbots: ChatGPT (GPT-5.4; OpenAI, San Francisco, CA, USA), Gemini (3.1 Pro; Google, Mountain View, CA, USA), and Microsoft Copilot (Microsoft, Redmond, USA). Ethical approval was not required, as the study did not involve patient data.

In the initial stage, ChatGPT (GPT-5.4) was prompted with 50 frequently asked questions (FAQs) related to orthodontic retention. From this pool, 25 questions considered to be the most relevant and frequently encountered in clinical orthodontic practice were selected by two experienced orthodontists (Y.L.S and G.A.S). Afterwards, the questions were categorized into eight thematic domains: stability and relapse (2 questions), comparison of retainer types (5 questions), duration of retention (3 questions), fixed retainers (5 questions), removable retainers (3 questions), hygiene and maintenance (3 questions), retention according to treatment modality (2 questions), and current retainer materials (2 questions) (Table 1).

Table 1
Orthodontic retention questions submitted to the chatbots, organized by thematic domain

1. Stability and Relapse
What are the biological mechanisms underlying relapse after orthodontic treatment?
In which orthodontic malocclusions is the risk of relapse higher?
2. Comparison of Retainer Types
What are the advantages and disadvantages of fixed and removable retainers in terms of stability?
On what basis are different retention strategies determined for the maxillary and mandibular arches?
How do Hawley retainers and vacuum-formed retainers (VFRs) compare in terms of long-term stability?
What are the indications for fixed retainers?
What are the indications for removable retainers?
3. Duration of Retention
What are the main clinical factors that determine the duration of retention?
How are part-time and full-time wear protocols determined?
Is there a linear relationship between retention duration and relapse risk?
4. Fixed Retainers
Which type of wire should be preferred for fixed retainer applications?
What is the concept of an unintentionally active retainer?
What are the periodontal effects of fixed retainers?
How should fixed retainer failures or debonding be managed?
What are the risks associated with long-term use of fixed retainers?
5. Removable Retainers
What types of removable retainers can be used during the retention phase?
What are the causes of fracture and deformation in VFRs?
How should changes in retainer fit be interpreted?
6. Hygiene and Maintenance
What is the optimal oral hygiene protocol in the presence of a fixed retainer?
Does retainer material influence biofilm accumulation?

1. Stability and Relapse
What are the biological mechanisms underlying relapse after orthodontic treatment?
What is the risk of heat-related deformation in thermoplastic retainers?
7. Retention According to Treatment Modality
Does the retention approach differ after clear aligner therapy?
How should retention be planned in cases involving transverse expansion?
8. Current Retainer Materials
What are the material differences between Vivera retainers and conventional Essix-type retainers?
How do 3D-printed retainers influence stability?

In the second stage, a single investigator (Y. L. S) submitted each question to the three chatbots on March 30, 2026. Each question was entered only once per chatbot, and the initial response generated was recorded without any modification. To minimize technical and contextual variability, all queries were conducted on the same day using the same device (MacBook Air M2, 16 GB RAM; Apple, Cupertino, CA, USA), a stable internet connection and a virtual private network (VPN) server. The chat interface was refreshed after each question, and the chat history was deleted to prevent contextual bias from previous interactions. To assess response consistency and intra-rater reliability, five questions were randomly repeated for each chatbot. Consequently, a total of 90 responses (30 per chatbot) were obtained. Following data collection, all responses were anonymized and randomized using electronic randomization software (https://www.random.org/lists/). All references to the chatbots' identity were removed to ensure evaluator blinding. The anonymized responses were then distributed across three evaluation forms, each containing a set of responses.

In the third stage, three independent and blinded orthodontists (Y.L.S., G.A.S. and M.G.) with expertise in clinical orthodontics and research evaluated the responses. The orthodontists had different levels of clinical experience in orthodontics, which helped reduce potential selection and confirmation biases. The

flowchart of the study is given in Fig. 1.

The quality of the chatbot responses was assessed using a five-point Likert scale, as follows:

1. Very poor quality: Incomplete or unclear answer with minimal relevant information, lacking clinical justification and containing significant inaccuracies.
2. Poor quality: Major deficiencies with missing critical elements and limited support from clinical evidence.

3. Moderate quality: Answers meeting basic clinical requirements but containing noticeable gaps or inconsistencies.
4. Good quality: Well-structured answers with sound clinical reasoning and generally consistent with evidence-based orthodontic principles.
5. Excellent quality: Comprehensive, well-organized, and highly evidence-based responses addressing all clinically relevant aspects.

The readability of the chatbot-generated responses was assessed using the Flesch Reading Ease (FRE) index, which is a widely used readability metric that evaluates text complexity based on sentence length and syllable count per word, yielding scores on a scale from 0 (most difficult) to 100 (easiest to read). [21, 22] Conversion tables are used to convert scores to educational levels. FRE scores were calculated for each chatbot response using an online readability tool (https://goodcalculators.com).The resulting scores were classified according to the standard FRE interpretation scale: 0–29 (very difficult, college graduate level), 30–49 (difficult, college level), 50–59 (fairly difficult, high school level [10th to 12th grade]), 60–69 (standard/plain, 8th to 9th grade), 70–79 (fairly easy, 7th grade), 80–89 (easy, 5th to 6th grade), and 90–100 (very easy, 4th to 5th grade) [22, 23]. The Flesch-Kincaid Grade Level (FKGL) indicates what education level a person will need to understand a particular text, and is evaluated by examining how many words, sentences, and syllables a document contains. Higher FKGL scores indicate greater text complexity. All readability scores, including extreme values (e.g., FRE = 0), were retained in the analysis to reflect the original output of the chatbot models without modification.

All data were recorded and organized using Microsoft Excel (Microsoft Corp., Redmond, WA, USA) prior to statistical analysis.

Statistical Analysis

Descriptive statistics were reported as medians (minimum - maximum). Mean ranks from the Kruskal-Wallis test were additionally reported to facilitate interpretation of group differences. Intra-rater reliability was calculated using Weighted Kappa (κ). Inter-rater reliability among the three independent orthodontists was evaluated using Fleiss' κ for the combined overall dataset and individually for each chatbot.

The Kruskal-Wallis H test was utilized to analyze potential differences in the overall response quality scores, FRE and FKGL scores among the three chatbots, followed by Dunn's post-hoc pairwise comparisons with a Bonferroni correction (significant at $p < 0.017$). A subgroup analysis was performed using the Kruskal-Wallis test to compare the performance of the chatbots within each of the eight specific thematic orthodontic retention domains. Additionally, a Spearman rank-order correlation was performed to evaluate the relationship of FRE and FKGL scores with overall quality ratings.

Statistical analyses were performed using SPSS Statistics (version 26.0; IBM Corp., Armonk, NY, USA). All tests were two-sided, and the threshold for statistical significance was set at $p < 0.05$.

Results

Intra- and Inter-Rater Reliability

Regarding intrarater reliability, the evaluators demonstrated substantial to almost perfect reliability ($\kappa > 0.737$; $p < 0.05$).

The evaluation of the 75 distinct chatbot responses demonstrated substantial agreement among the three orthodontists ($\kappa = 0.781$; $p < 0.001$) (Table 2). When examining the agreement across specific quality categories for the combined dataset, the evaluators demonstrated significant consensus for categories representing "Poor Quality" ($\kappa = 1.000$), "Moderate Quality" ($\kappa = 0.847$), "Good Quality" ($\kappa = 0.733$), and "Excellent Quality" ($\kappa = 0.754$) ($p < 0.001$). Notably, no responses were scored as "Very poor quality" by any of the raters.

Regarding evaluator agreement for each specific chatbot, the inter-rater reliability remained high, though slight variations in consensus were observed depending on the chatbot being evaluated. Evaluators achieved the highest level of consensus when assessing responses generated by Gemini, demonstrating almost perfect agreement ($\kappa = 0.832$; $p < 0.001$). Responses generated by Copilot yielded substantial overall agreement among the raters ($\kappa = 0.792$; $p < 0.001$). While evaluators demonstrated perfect agreement ($\kappa = 1.000$) when identifying "Poor" quality responses from Copilot, there was no significant agreement when scoring Copilot's "Excellent" responses ($\kappa = 0.208$; $p = 0.072$). Evaluation of ChatGPT's responses also resulted in substantial overall agreement ($\kappa = 0.692$; $p < 0.001$). Evaluator consensus was statistically significant across "Moderate" ($\kappa = 0.702$) and "Good" ($\kappa = 0.627$) categories, with the highest agreement observed for responses deemed "Excellent" ($\kappa = 0.760$; $p < 0.001$).

Table 2
Inter-rater reliability for the evaluation of chatbot-generated orthodontic retention responses.

Chatbot	n	Fleiss' Kappa (κ)	95% CI Lower – Upper	p value
Overall Dataset	75	0.781	0.778–0.784	< 0.001***
ChatGPT	25	0.692	0.687–0.698	< 0.001***
Microsoft Copilot	25	0.792	0.786–0.797	< 0.001***
Gemini	25	0.832	0.827–0.838	< 0.001***

*CI: Confidence Interval. ***Significant at p < 0.001.*

Comparison of chatbot performance

A significant difference was observed in the overall performance of the chatbots ($p = 0.019$), with ChatGPT achieving the highest mean rank (45.54), followed by Gemini (38.46) and Copilot (30.00) (Table 3). Post-hoc pairwise comparisons revealed a significant difference in response quality between Microsoft Copilot and ChatGPT (Adjusted $p = 0.005$). However, the performance of Gemini did not differ significantly from either ChatGPT or Microsoft Copilot ($p > 0.017$). Regarding the comparative performance of the chatbots within each of the eight highly specific orthodontic retention domains, there were no significant differences between the models in any individual category ($p > 0.05$) (Fig. 2). The evaluation of questions pertaining to "Removable retainers" approached statistical significance ($p = 0.060$), but ultimately, the models performed comparably within isolated clinical categories.

Table 3
Chatbot performance across overall and specific orthodontic retention domain questions

Chatbot	Median	Min – Max	Test Statistic (H)	P value
Overall			7.934	0.019*
ChatGPT	4.0	3.0–5.0		
Gemini	4.0	3.0–5.0		
Copilot	4.0	2.0–5.0		
Stability and relapse			1.250	0.535
ChatGPT	4.5	4.0–5.0		
Gemini	4.5	4.0–5.0		
Copilot	4.0	4.0–4.0		
Comparison of retainer types			2.343	0.310
ChatGPT	4.0	3.0–4.0		
Gemini	4.0	3.0–5.0		
Copilot	4.0	2.0–4.0		
Duration of retention			1.143	0.565
ChatGPT	4.0	4.0–4.0		
Gemini	4.0	3.0–4.0		
Copilot	4.0	3.0–4.0		
Fixed retainers			3.274	0.195
ChatGPT	5.0	4.0–5.0		
Gemini	4.0	4.0–5.0		
Copilot	4.0	3.0–5.0		
Removable retainers			5.627	0.060
ChatGPT	4.0	4.0–5.0		
Gemini	3.0	3.0–3.0		
Copilot	4.0	3.0–4.0		
Hygiene and maintenance			3.111	0.211
ChatGPT	4.0	3.0–5.0		
Gemini	4.0	3.0–4.0		

Chatbot	Median	Min – Max	Test Statistic (H)	P value
Copilot	3.0	3.0–3.0		
Retention according to treatment modality			1.250	0.535
ChatGPT	4.5	4.0–5.0		
Gemini	4.0	3.0–5.0		
Copilot	3.5	3.0–4.0		
Current retainer materials			2.000	0.368
ChatGPT	3.5	3.0–4.0		
Gemini	4.0	4.0–4.0		
Copilot	4.0	4.0–4.0		

Min: Minimum; Max: Maximum. H: Kruskal-Wallis Test Statistic.

**Significant at p < 0.05.*

Readability Analysis

The median FRE scores were 20.9 (0.0–51.7) for ChatGPT, 23.4 (0.0–54.0) for Gemini, and 22.2 (0.0–62.4) for Microsoft Copilot. The median FKGL scores were 12.6 (10.2–16.3) for ChatGPT, 10.75 (7.2–12.7) for Gemini, and 12.5 (4.8–15.4) for Microsoft Copilot. There were no differences among the three chatbots in either FRE or FKGL scores ($p > 0.05$). Overall, the FRE scores across all chatbots fell within the very difficult to difficult readability range, corresponding to a college graduate or professional reading level, while FKGL scores indicated a high school- to college-level reading grade (Table 4). In addition, overall quality ratings were not significantly correlated with either FRE scores ($\rho = -0.102$, $p = 0.386$) or FKGL scores ($\rho = 0.113$, $p = 0.336$).

Table 4
Readability Analysis of Chatbot-Generated Responses

Chatbots	FRE score Median (Min-Max)	FKGL Median (Min-Max)
ChatGPT	20.9 (0.0–51.7)	12.6 (10.2–16.3)
Gemini	23.4 (0.0–54.0)	10.75 (7.2–12.7)
Microsoft Copilot	22.2 (0.0–62.4)	12.5 (4.8–15.4)

FRE: Flesch Reading Ease; FKGL: Flesch-Kincaid Grade Level.

Discussion

The present study evaluated the quality and clinical relevance of responses generated by three widely used AI-based chatbots, ChatGPT, Google Gemini, and Microsoft Copilot, in addressing orthodontic retention-related questions. Although all chatbot models demonstrated generally good overall performance, the findings revealed minor variations in response quality. However, no differences were observed among the chatbots within specific orthodontic retention domains. While higher scores were generally observed in broader conceptual domains such as “stability and relapse” and “retention duration”, comparatively lower or more variable scores were noted in clinically nuanced topics, including “removable retainers”, “hygiene and maintenance”, and “retainer materials”. This pattern suggests that chatbots may perform more effectively when addressing general theoretical concepts but may lack depth and precision in more practical, patient-specific, or material-based questions. In addition, readability analysis revealed that all chatbot responses were written at a level substantially exceeding recommended thresholds for patient education materials.

Questions related to removable retainers and appliance maintenance directly influence patient compliance and long-term treatment stability. Inadequate or incomplete information in these areas may lead to improper retainer use, poor hygiene practices, or delayed recognition of appliance-related complications. Given that patient adherence is a critical determinant of post-treatment stability, even minor inaccuracies in these domains may have disproportionately large clinical consequences.

These results are partially consistent with the findings of Aşkın and Manav [24], who reported that chatbot performance varied significantly depending on the evaluation metric, with healthcare-specific models providing more structured and reliable information, while general-purpose models demonstrated better readability and originality. However, their study primarily focused on patient-oriented outputs and included healthcare-specific AI systems evaluated using structured tools such as EQIP, SMOG, and similarity indices. In contrast, the present study adopts a more clinically oriented framework by incorporating blinded orthodontist evaluation, inter- and intra-rater reliability analyses, and domain-specific performance assessment. These methodological differences may account for the observed variation in findings, particularly regarding the consistency and clinical applicability of chatbot-generated responses.

Similarly, Erdem et al. [25] demonstrated that chatbot accuracy varies depending on the clinical context, with significant differences observed in orthodontic emergency scenarios. Notably, their study found that some categories showed significant differences while others did not, indicating that chatbot performance is context-dependent rather than uniform across topics. This aligns with the present findings, where no differences were observed in specific retention-related domains, yet variation was evident in the overall performance comparison. Although numerical differences were observed across domains, these did not reach statistical significance. Therefore, chatbot performance appeared broadly consistent across different orthodontic retention topics, with only minor variations that may reflect the complexity of specific clinical questions rather than true performance differences. From a clinical

perspective, however, not all retention-related topics carry equal importance for long-term treatment success. Domains such as retainer wear protocols, hygiene and maintenance, and appliance integrity are directly associated with patient compliance and the prevention of relapse. Inadequate or unclear information in these areas may lead to improper retainer use, increased risk of appliance failure, or compromised periodontal health. Similarly, information regarding contemporary retainer materials and treatment-specific retention strategies is clinically relevant, as it may influence both treatment stability and patient adherence. This consideration is particularly important given that the retention phase represents the least supervised stage of orthodontic treatment, during which patients are more likely to rely on external sources of information. Therefore, even in the absence of statistically significant differences, the clinical implications of the information provided in these domains remain highly relevant.

In addition, the findings of Kılınç and Mansız [23] highlight another important dimension of chatbot-generated responses, namely, the balance between readability and reliability. Their study demonstrated that although newer versions of ChatGPT showed improved reliability, readability decreased, resulting in more complex and potentially less accessible information for patients. This is particularly relevant in the context of orthodontic retention, where patient understanding plays a crucial role in compliance. Even when chatbot responses are clinically accurate, overly complex explanations may reduce their practical utility for patient education. This issue is particularly important considering that recommended health literacy levels for patient education materials are typically at or below the sixth-grade level [23]. The consistently higher FKGL scores observed across studies highlight a mismatch between the complexity of AI-generated content and the literacy level of the general population.

Health literacy refers to the ability of individuals to obtain, process, and understand basic health information in order to make appropriate health-related decisions. Readability is a fundamental component of health literacy, as the complexity of written content directly influences patient comprehension and engagement [26]. Previous studies have consistently demonstrated that health-related materials are often written above the recommended readability levels, thereby limiting their accessibility to the general population [27]. In the context of orthodontics, this issue becomes even more critical, as patients are particularly adolescents, are expected to understand and adhere to detailed instructions during treatment and retention phases. The United States Department of Health and Human Services recommends that patient education materials should not exceed a sixth-grade reading level to ensure adequate comprehension [28]. However, both the present findings and previous literature indicate that this threshold is rarely achieved in practice [22]. The results of this study demonstrated that all chatbot-generated responses corresponded to an upper high school to college-level reading grade, which is substantially higher than the recommended level for patient-directed materials. These findings are consistent with those reported by Kılınç and Mansız [23], who found that ChatGPT responses were written at a college graduate level for general orthodontic questions and at a high school level for treatment-related questions. Similarly, Sokmen et al. [22] reported that although AI-generated responses were generally understandable and clinically relevant, their readability remained difficult and exceeded recommended thresholds for patient comprehension.

One important consideration when interpreting readability indices is that they are primarily based on structural features such as sentence length and syllable count. While these metrics provide standardized and reproducible assessments, they may not fully capture true comprehension, as complex medical terminology may still be difficult for patients even when presented in shorter sentences. Conversely, oversimplification of content may compromise the accuracy and completeness of clinical information. Therefore, readability should not be evaluated in isolation, but rather in conjunction with content quality, clarity, and clinical relevance [23, 29, 30]. Taken together, these findings highlight a persistent gap between the linguistic complexity of AI-generated health information and the literacy level of the general population. Although chatbots are capable of producing clinically acceptable and structured responses, their limited readability may reduce their effectiveness as patient education tools. This mismatch underscores the need for optimization of AI-generated content to better align with established health literacy standards.

The importance of accurate patient guidance is further underscored by the well-established risk of orthodontic relapse, which remains a common outcome with rates influenced by factors such as retainer wear duration, treatment complexity, and patient compliance. This is further supported by the study of Agrawal et al. [31], which emphasizes that insufficient adherence to retention protocols is strongly associated with increased relapse risk. In this context, misinformation or incomplete guidance provided by chatbots, particularly regarding retainer wear, maintenance, or follow-up, may negatively affect patient behavior and compromise long-term orthodontic treatment outcomes.

Although no responses were classified as “very poor” in this study, variability within higher-quality categories indicates inconsistency in delivering comprehensive and clinically detailed information. This suggests that current chatbot systems may have reached a baseline level of competence but have not yet achieved consistent excellence in clinical communication. This is particularly relevant in orthodontic retention, where subtle differences in instructions, such as wear protocols, appliance care, or interpretation of retainer fit, can have significant clinical implications [6, 32].

The clinical context of retention further amplifies the importance of these findings. Unlike the active phase of orthodontic treatment, the retention phase is characterized by reduced frequency of professional supervision and increased patient autonomy. As a result, patients are more likely to rely on alternative information sources, including AI-based chatbots. While these systems offer immediate and accessible information, the variability observed in this study suggests that they should be considered adjunctive tools rather than primary sources of clinical guidance.

The present study has several limitations that should be acknowledged. The study evaluated only single-response outputs and did not assess the effect of iterative questioning, which may influence response quality. Additionally, although the selected questions were representative of common clinical scenarios, they may not fully encompass the complexity of orthodontic retention. The initial question pool was AI-assisted and then expert-filtered, which may have influenced phrasing or scope. Furthermore, as AI systems are continuously updated, the findings reflect a specific time point and may not be directly

applicable to different model versions. Future research should focus on domain-specific evaluation of chatbot performance, particularly in clinically sensitive areas such as appliance management, hygiene protocols, and long-term retention strategies. Additionally, integrating patient-centered outcomes, such as comprehension, adherence, and behavioral changes, may provide a more comprehensive understanding of the real-world impact of AI-generated information in orthodontics.

Conclusions

- AI-based chatbots, ChatGPT, Google Gemini, and Microsoft Copilot, demonstrated generally good performance in answering orthodontic retention-related questions.
- No differences were observed among chatbots across specific orthodontic retention domains, indicating comparable performance in addressing different clinical topics.
- Chatbots were more consistent in addressing general concepts, but showed limitations in providing precise, practice-oriented guidance relevant to patient compliance.
- Despite acceptable overall quality, all chatbot responses were written at a readability level substantially exceeding recommended thresholds for patient education materials.
- Given the critical role of patient adherence in preventing orthodontic relapse, incomplete or imprecise information in the retention phase may have clinically significant consequences.
- AI chatbots may serve as supportive tools for patient education, especially when clinical supervision is limited. However, they should not replace professional orthodontic guidance.

Declarations

Ethics approval and consent to participate

As this study did not involve human participants or animal subjects, ethical approval was not required.

Conflict of Interest

The authors declare that they have no conflict of interest.

Funding

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

Clinical trial number

Not applicable.

Consent to Publish

Not applicable.

Availability of supporting data

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

Acknowledgement

The authors would like to thank Dr.Meltem Gediz for her valuable contribution to this study.

Authors' contributions

YLS: Conceptualization, Investigation, Formal Analysis, Methodology, Project Administration, Supervision, Data Curation, Writing – Original Draft Preparation, Writing – Review & Editing. **GAS**: Formal Analysis, Resources, Software, Writing – Review & Editing, Validation, Visualization.

References

1. Melrose C, Millett DT. Toward a perspective on orthodontic retention? Am J Orthod Dentofac Orthop. 1998;113:507–14. https://doi.org/10.1016/S0889-5406(98)70261-6.
2. Little RM, Riedel RA, Artun J. An evaluation of changes in mandibular anterior alignment from 10 to 20 years postretention. Am J Orthod Dentofac Orthop. 1988;93:423–8. https://doi.org/10.1016/0889-5406(88)90102-3.
3. Thilandera B. Orthodontic relapse versus natural development. Am J Orthod Dentofac Orthop. 2000;117:562–3. https://doi.org/10.1016/S0889-5406(00)70200-9.
4. Padmos JAD, Fudalej PS, Renkema AM. Epidemiologic study of orthodontic retention procedures. Am J Orthod Dentofac Orthop. 2018;153:496–504. https://doi.org/10.1016/j.ajodo.2017.08.013.
5. Littlewood SJ, Millett DT, Doubleday B, Bearn DR, Worthington HV. Retention procedures for stabilising tooth position after treatment with orthodontic braces. Cochrane Database Syst Reviews 2016;2016. https://doi.org/10.1002/14651858.CD002283.pub4
6. Johnston CD, Littlewood SJ. Retention in orthodontics. Br Dent J. 2015;218:119–22. https://doi.org/10.1038/sj.bdj.2015.47.
7. Crispino R, Mannocci A, Dilena IA, Sides J, Forchini F, Asif Alherawi WM, et al. Orthodontic patients and the information found on the web: a cross-sectional study. BMC Oral Health. 2023;23:860. https://doi.org/10.1186/s12903-023-03609-4.
8. Eggmann F, Weiger R, Zitzmann NU, Blatz MB. Implications of large language models such as ChatGPT for dental medicine. J Esthetic Restor Dentistry. 2023;35:1098–102. https://doi.org/10.1111/jerd.13046.
9. Schwendicke F, Samek W, Krois J. Artificial Intelligence in Dentistry: Chances and Challenges. J Dent Res. 2020;99:769–74. https://doi.org/10.1177/0022034520915714.

10. Hu X, Zhao Y, Yang C. Evaluation of root position during orthodontic treatment via multiple intraoral scans with automated registration technology. Am J Orthod Dentofac Orthop. 2023;164:285–92. https://doi.org/10.1016/j.ajodo.2023.04.012.
11. Khanagar SB, Al-Ehaideb A, Vishwanathaiah S, Maganur PC, Patil S, Naik S, et al. Scope and performance of artificial intelligence technology in orthodontic diagnosis, treatment planning, and clinical decision-making - A systematic review. J Dent Sci. 2021;16:482–92. https://doi.org/10.1016/j.jds.2020.05.022.
12. Ferlito T, Hsiou D, Hargett K, Herzog C, Bachour P, Katebi N, et al. Assessment of artificial intelligence–based remote monitoring of clear aligner therapy: A prospective study. Am J Orthod Dentofac Orthop. 2023;164:194–200. https://doi.org/10.1016/j.ajodo.2022.11.020.
13. Mohammad-Rahimi H, Nadimi M, Rohban MH, Shamsoddin E, Lee VY, Motamedian SR. Machine learning and orthodontics, current trends and the future opportunities: A scoping review. Am J Orthod Dentofac Orthop. 2021;160:170–e1924. https://doi.org/10.1016/j.ajodo.2021.02.013.
14. Sharma D, Kaushal S, Kumar H, Gainder S. Chatbots in Healthcare: Challenges, Technologies and Applications., Technology S. (AIST), IEEE; 2022, pp. 1–6. https://doi.org/10.1109/AIST55798.2022.10065328
15. Dursun D, Bilici Geçer R. Can artificial intelligence models serve as patient information consultants in orthodontics? BMC Med Inf Decis Mak. 2024;24:211. https://doi.org/10.1186/s12911-024-02619-8.
16. Hatia A, Doldo T, Parrini S, Chisci E, Cipriani L, Montagna L, et al. Accuracy and Completeness of ChatGPT-Generated Information on Interceptive Orthodontics: A Multicenter Collaborative Study. J Clin Med. 2024;13:735. https://doi.org/10.3390/jcm13030735.
17. Arslan C, Kahya K, Cesur E, Cakan DG. An evaluation of orthodontic information quality regarding artificial intelligence (AI) chatbot technologies: A comparison of ChatGPT and google BARD. Australasian Orthodontic J. 2024;40:149–57. https://doi.org/10.2478/aoj-2024-0012.
18. Aziz AAA, Abdelrahman HH, Hassan MG. The use of ChatGPT and Google Gemini in responding to orthognathic surgery-related questions: A comparative study. J World Fed Orthod. 2025;14:20–6. https://doi.org/10.1016/j.ejwf.2024.09.004.
19. Daraqel B, Wafaie K, Mohammed H, Cao L, Mheissen S, Liu Y, et al. The performance of artificial intelligence models in generating responses to general orthodontic questions: ChatGPT vs Google Bard. Am J Orthod Dentofac Orthop. 2024;165:652–62. https://doi.org/10.1016/j.ajodo.2024.01.012.
20. Kurt Demirsoy K, Buyuk SK, Bicer T. How reliable is the artificial intelligence product large language model ChatGPT in orthodontics? Angle Orthod. 2024;94:602–7. https://doi.org/10.2319/031224-207.1.
21. Flesch R. A new readability yardstick. J Appl Psychol. 1948;32:221–33. https://doi.org/10.1037/h0057532.

22. 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;25:1845. https://doi.org/10.1186/s12903-025-07153-1.
23. Kılınç DD, Mansız D. Examination of the reliability and readability of Chatbot Generative Pretrained Transformer's (ChatGPT) responses to questions about orthodontics and the evolution of these responses in an updated version. Am J Orthod Dentofac Orthop. 2024;165:546–55. https://doi.org/10.1016/j.ajodo.2023.11.012.
24. Aşkın MB, Manav ÖC. Evaluation of AI-based chatbot responses to orthodontic retention-related patient questions: a comparative analysis. J Dent Sci Educ. 2025;3:74–9. https://doi.org/10.51271/JDSE-0059.
25. 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;96:100–5. https://doi.org/10.2319/021825-146.1.
26. Nutbeam D. Health promotion glossary. Health Promot Int. 1986;1:113–27. https://doi.org/10.1093/heapro/1.1.113.
27. Mcinnes N, Haglund BJA. Readability of online health information: implications for health literacy. Inf Health Soc Care. 2011;36:173–89. https://doi.org/10.3109/17538157.2010.542529.
28. Meade MJ, Dreyer CW. How readable are orthognathic surgery consent forms? Int Orthod. 2022;20:100689. https://doi.org/10.1016/j.ortho.2022.100689.
29. Seehra J, Cockerham L, Pandis N. A quality assessment of orthodontic patient information leaflets. Prog Orthod. 2016;17:15. https://doi.org/10.1186/s40510-016-0128-y.
30. Kenny T. A PIL for every ill? Patient information leaflets (PILs): a review of past, present and future use. Fam Pract. 1998;15:471–9. https://doi.org/10.1093/fampra/15.5.471.
31. Agrawal S, Chawla M, Khan R, Pandiyath SA, Das DB, Marwaha S. AI-driven predictive modelling of orthodontic relapse using retainer compliance and patient factors. Bioinformation. 2025;21:2022–6. https://doi.org/10.6026/973206300212022.
32. Quinzi V, Carli E, Mummolo A, De Benedictis F, Salvati SE, Mampieri G. Fixed and removable orthodontic retainers, effects on periodontal health compared: A systematic review. J Oral Biol Craniofac Res. 2023;13:337–46. https://doi.org/10.1016/j.jobcr.2023.02.015.

Figures

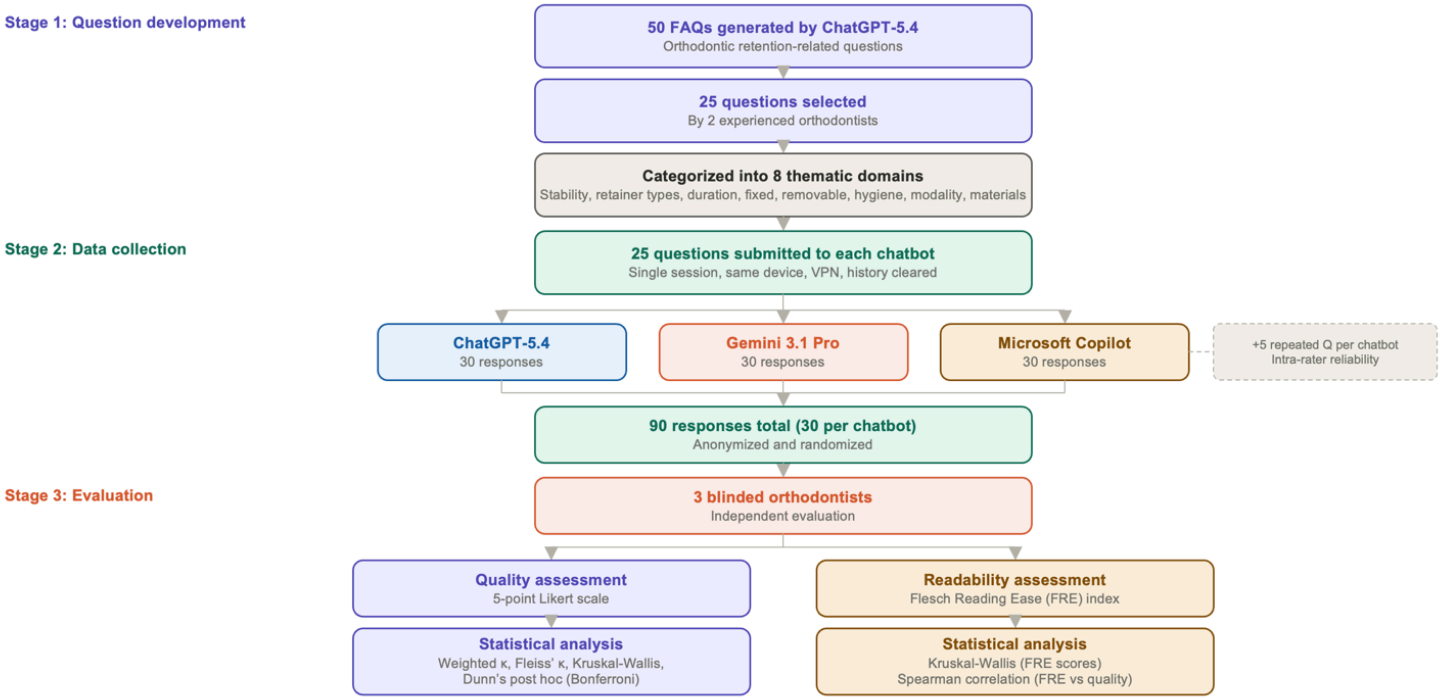


Figure 1

Flowchart of the study.

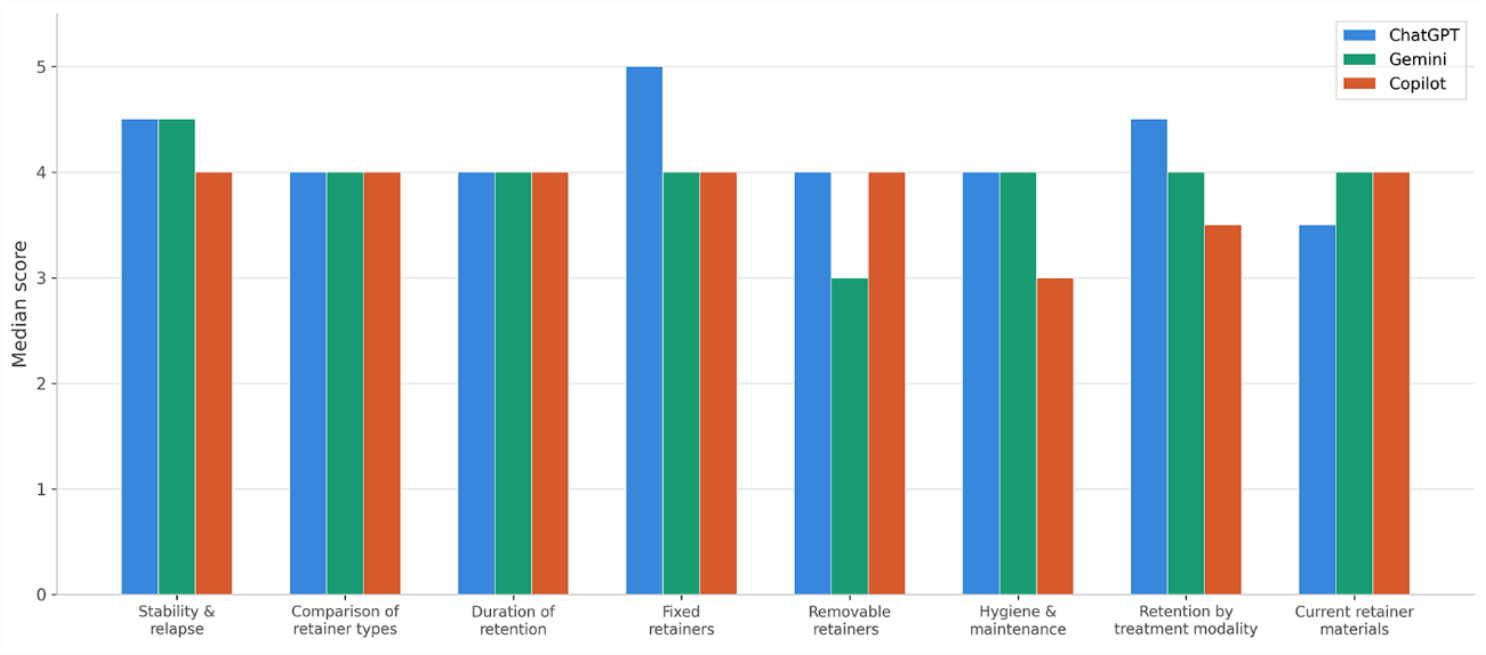


Figure 2

Chatbot performance across specific orthodontic retention domain questions