Original Article

Can AI chatbots accurately provide information on orthodontic risks?

Zeng Fan[a]; Jie Lei[b]; Wanwei Shi[a]; Yao Lin[a]; Qing Wang[a]; Lina Bao[c]

ABSTRACT

Objectives: To evaluate and compare the validity and reliability of different artificial intelligence (AI) chatbots in answering queries about potential orthodontic risks.

Materials and Methods: Answers to 20 frequently asked questions about the potential risks of orthodontics were derived from daily consultations with experienced orthodontists and AI chatbots (ChatGPT 4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro). The questions were repeated three times and submitted to the AI chatbots to assess the reliability of their answers. The answers from AI chatbots were scored using a modified Global Quality Scale (GQS). Low- and high-threshold validity tests were used to determine validity, and Cronbach's alpha was used to evaluate the consistency of the three responses to each of the 20 questions.

Results: In the low-threshold validity test, Gemini exhibited the highest overall performance. In the high-threshold validity test, Gemini also showed the highest overall effectiveness, but there was no significant difference observed among the three chatbots. All three chatbots demonstrated satisfactory levels of reliability, with Gemini having the highest consistency.

Conclusions: AI chatbots have some potential in providing orthodontic risk information, but they must be used cautiously and further optimized to improve their effectiveness in clinical practice. (*Angle Orthod*. 2025;00:000–000.)

KEY WORDS: Artificial intelligence; Large language models; Orthodontic risks; Patient information

INTRODUCTION

The rapid development of artificial intelligence (AI) technology, especially large language models (LLMs) such as ChatGPT developed by OpenAI, has resulted in revolutionary changes in several domains since its launch in November 2022.[1] These LLMs have profoundly changed various fields in many ways due to their ability to build on complex concepts and generate appropriate human-like text responses based on their extensive training using Internet text data.[2]

There has been a significant increase in the use of LLM tools and their applications in dentistry during the past 2 years, helping professionals deliver better oral treatment and healthcare. These tools have considerable potential for image recognition, data report generation, assisted diagnosis, and treatment planning.[3–5] LLMs, such as ChatGPT, Claude, and Gemini, are widely used by clinicians, researchers, and other professionals; patients can also obtain relevant information through these AI chatbots without time and space constraints.[2] With the increasing popularity of AI and the growing public demand for information about conditions and treatments, patients are turning to AI chatbots and online search engines as a convenient source of medical and dental information.[6,7]

Orthodontic treatment involves complex biomechanical and esthetic considerations, and patients need to fully understand the risks and challenges encountered during treatment.[8,9] However, if the risks of orthodontic treatment are misrepresented, whether by exaggeration or minimization, patients can easily make uninformed decisions, leading to heightened anxiety, overlooked complications, and dissatisfaction. Despite the excellent performance of AI chatbots in handling daily queries, their accuracy and reliability in providing orthodontic risk disclosure require further inestigation.[10] Therefore,

The first two authors contributed equally to this work.

[a] Orthodontic Resident, Department of Orthodontics, Stomatological Hospital, School of Stomatology, Southern Medical University, Guangzhou, China.

[b] Orthodontic Resident, Department of Orthodontics, Changsha Stomatological Hospital, Changsha, Hunan Province, China.

[c] Associate Professor and Program Director, Department of Orthodontics, Stomatological Hospital, School of Stomatology, Southern Medical University, Guangzhou, China.

Corresponding author: Lina Bao, S366 Jiangnan Boulevard, Haizhu District, Guangzhou City, Guangdong Province, China (e-mail: baolina1019@163.com)

Accepted: April 23, 2025. Submitted: December 14, 2024.
Published Online: June 20, 2025

Table 1. Frequently Asked Questions. (A) Questions Formulated By Orthodontists Through Their Daily Interactions With Patients; (B) Questions Provided by AI Chatbots as the Top "Frequently Asked Questions" Related to Potential Orthodontic Risk

	Question
A	
	1. Does orthodontic treatment lead to gingival recession?
	2. Does orthodontic treatment cause fenestration and dehiscence?
	3. Does orthodontic treatment result in the loosening of a tooth?
	4. Is there a risk of temporomandibular joint (TMJ) problems during or after orthodontic treatment?
	5. Do pregnancy and lactation affect orthodontic treatment?
	6. Is there a risk of relapse after orthodontic treatment, and how can it be prevented?
	7. Can orthodontic treatment lead to side effects in facial aesthetics?
	8. Is there a possibility of root resorption during orthodontic treatment?
	9. Can orthodontic treatment cause tooth decay or white spots on my teeth? How can that be prevented?
	10. Is it possible for oral ulcers to occur during orthodontic treatment? What if this happens?
B	
	11. Are there any risks associated with tooth extractions for orthodontic purposes?
	12. Is orthodontic treatment painful, and what kind of discomfort should I expect?
	13. Are there any risks of orthodontic treatment affecting my speech?
	14. Are there any risks associated with orthodontic X-rays?
	15. What happens if a bracket or wire comes loose during orthodontic treatment? Is it dangerous?
	16. Does orthodontic treatment lead to tooth sensitivity?
	17. Are there any age-related risks during orthodontic treatment that I should know?
	18. How long does orthodontic treatment usually take, and what happens if it takes longer than expected?
	19. What can I do if I'm not satisfied with the progress or outcome of the orthodontic treatment?
	20. Are there any risks specific to ceramic braces or clear aligners compared to metal braces for orthodontic treatment?

the reliability of AI chatbots in answering queries about potential risks associated with orthodontics needs to be assessed.

This study was conducted to evaluate and compare the validity and reliability of different AI chatbots in answering queries about potential orthodontic risks. The null hypothesis was that different AI chatbots would not differ significantly in effectiveness and reliability when answering queries about potential orthodontic risks.

MATERIALS AND METHODS

Data Collection

ChatGPT 4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro were chosen for this study, mainly based on their unique performance advantages, to more fully reflect the advanced nature of current AI models. These models are currently the most advanced versions in their respective series and, therefore, reflect the current state of AI technology.

A list of 20 frequently asked questions about the potential risks of orthodontic treatment was carefully compiled, covering a wide range of topics of general concern to patients. The issues were derived from the following two sources:

(A) First, 10 questions were selected based on practical questions frequently encountered in daily consultations by five full-time orthodontists with many years of clinical work experience.

(B) Using the prompt "20 frequently asked questions about orthodontic risks," the research team asked each of the three AI chatbots questions and collected their responses, totaling 60 frequently asked questions about orthodontic risks. The responses that closely resembled those provided by the orthodontic team were then removed. Finally, the research team selected an additional 10 questions based on everyday clinical practice.

Twenty of the most representative issues were identified through this comprehensive selection process, as shown in Table 1. Then, the questions were all presented to the three AI chatbots (ChatGPT 4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro) on December 9, 2024. Each question was asked three times to evaluate the reliability of the responses, with each query initiating a new chat conversation. The main application programming interfaces of each chatbot were used to simulate real-world interaction.

Scoring

All words related to the identity of each AI chatbot were removed from responses to ensure blinding of the evaluators. Responses were evaluated by a team of five full-time orthodontists familiar with the current literature and experienced in clinical practice, and scored using the modified Global Quality Scale (GQS), a tool used to evaluate the quality of written materials in the medical field. The scoring criteria for GQS were:

Score 1: Poor quality; most information is missing, of no help to the patient.

Score 2: Overall poor quality; numerous key subjects are absent, significantly restricting patient utility.

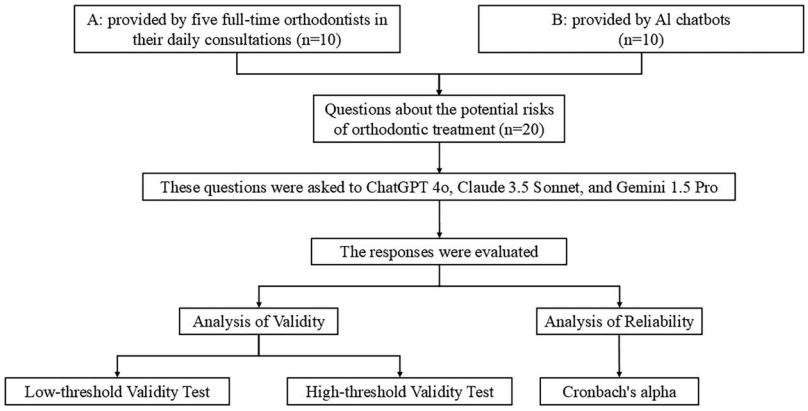

Figure 1. Flowchart of the study.

Score 3: Medium quality; certain crucial details are thoroughly examined, but other information is discussed insufficiently.
Score 4: Good quality; most of the relevant information listed is beneficial for the patient.
Score 5: Excellent quality; the process is excellent and extremely useful to the patient.

Differences in scoring among team members were addressed through evidence-based discussion regarding context and contents. The final score was established based on consensus achieved within the team. The process is shown in Figure 1.

Statistical Analyses

Analysis of validity. Low-threshold validity test: A chatbot was considered valid if all three responses to a question received a score of 4 or above. Any answer with a score lower than 4 was considered invalid.

High-threshold validity test: A chatbot was considered valid only if all three answers to a question received a score of 5. If any answer had a score less than 5, it was considered invalid.

The chi-square test was used to compare the validity of responses from various chatbot responses, with a significance level of less than 0.05. All statistical analyses were conducted using SPSS version 26 (IBM SPSS Statistics, NY, USA).

Analysis of reliability.

- Reliability was defined by analyzing the GQS score of chatbot responses to questions when repeated under consistent conditions (three repetitions).
- Cronbach's alpha was used to assess the consistency of all three responses to 20 questions. Cronbach's alpha values ranged from 0 to 1, where 0 indicated no consistency and 1 implied perfect consistency. A Cronbach's alpha value of $\geq$ 0.70 was regarded as an acceptable level of reliability as was considered in previous medical and health-related studies.[11]

RESULTS

Each of the three chatbots answered 20 questions, for a total number of 180 responses. The chatbots were evaluated based on the GQS score. The average GQS scores are shown in Figure 2. Additionally, low- and high-threshold validity tests were conducted.

Low-Threshold Validity

Of the three chatbots, Gemini showed the highest overall validity, with 19 of 20 responses (95%) classified as valid. ChatGPT 4o and Claude AI both demonstrated an overall validity of 75%, with 15 of 20 valid responses (Figure 3). However, there were no significant differences among ChatGPT 4o, Claude AI, and Gemini (Table 2).

High-Threshold Validity

As shown in Figure 4, Gemini demonstrated the highest overall validity, with 12 of 20 responses (60%) classified as valid, followed by Claude AI (55%) and ChatGPT 4o (50%). The high-threshold validity test revealed no significant differences among the chatbots (Table 3).

Analysis of Reliability

All three chatbots exhibited satisfactory levels of reliability. Gemini stood out with the highest overall consistency, with a Cronbach's alpha value of 0.95, followed closely by ChatGPT 4o with a Cronbach's alpha value of 0.94, and Claude AI with a Cronbach's alpha value of 0.92.

DISCUSSION

Orthodontic treatment carries potential risks, including tooth demineralization, gingival recession, and root

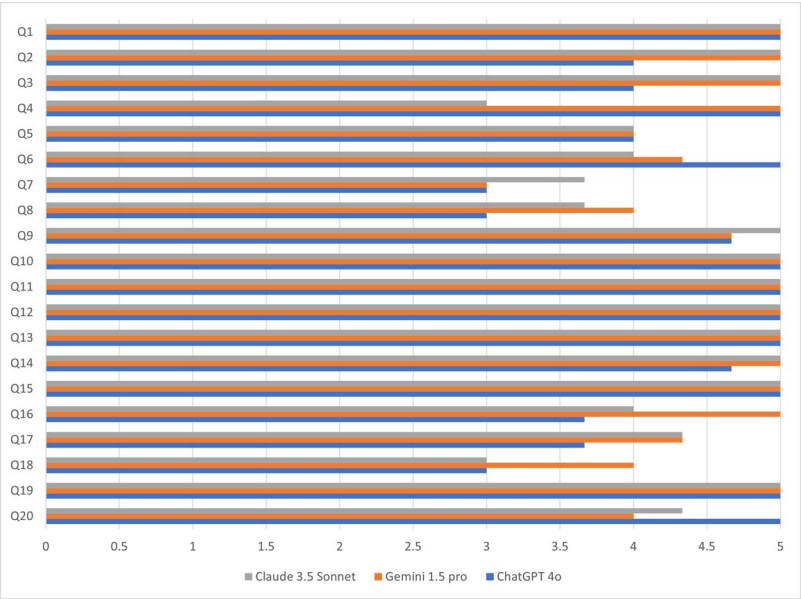

Figure 2. Mean scores for responses of chatbots with scoring on *x*-axis and questions on *y*-axis.

resorption.[12] With growing awareness of oral health, more patients are concerned about these treatment risks and are seeking more comprehensive information to make informed treatment decisions. Exaggerating or downplaying the potential risks of orthodontic treatment can lead to patients making decisions based on inaccurate information, increasing unnecessary anxiety or ignoring potential complications, affecting treatment outcomes and patient satisfaction, and potentially even leading to legal liability issues. Therefore, ensuring that patients receive accurate and objective risk-related information is essential.

In the 21st century, patients often seek orthodontic-related information from online resources, including social media; however, the quality and reliability of this information may often be questionable.[13–15] Use of AI in recent years has led to the development of AI chatbots that are revolutionizing healthcare. These chatbots

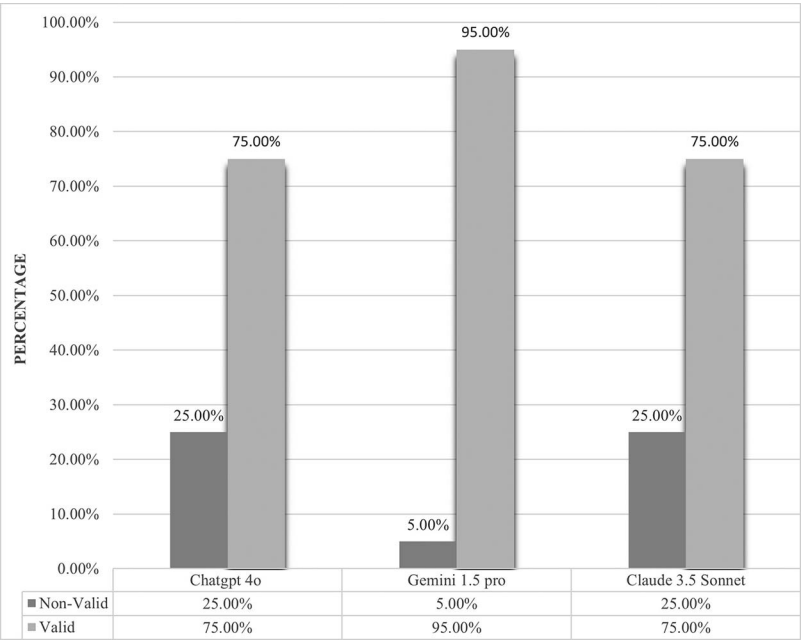

Figure 3. Analysis of valid and invalid responses in a low-threshold validity assessment.

Table 2. Comparison Between Chatbots for Low-Threshold Validity Test[a]

Comparison	ChatGPT 4o	Gemini 1.5 Pro	Claude 3.5 Sonnet
ChatGPT 4o	-	0.077	1.000
Gemini 1.5 Pro	0.077	-	0.077
Claude 3.5 Sonnet	1	0.077	-

[a] Chi-square test.

Table 3. Comparison Between Chatbots for High-Threshold Validity Test[a]

Comparison	ChatGPT 4o	Gemini 1.5 Pro	Claude 3.5 Sonnet
ChatGPT 4o	-	0.525	0.752
Gemini 1.5 Pro	0.525	-	0.749
Claude 3.5 Sonnet	0.752	0.749	-

[a] Chi-square test.

provide personalized medical support and education based on individual needs and preferences.[16] They are powered by natural language processing and machine learning algorithms. They can learn from patient interactions and adjust responses accordingly, making the user experience more natural and engaging.[17] However, assessment of the accuracy of the LLM responses to questions about the potential risks of orthodontics is lacking.

The source of the questions is a crucial factor in assessing the effectiveness of chatbot responses. The questions in this study were designed to reflect public concerns regarding risks related to orthodontic treatment. Some questions were developed through interactions with five full-time orthodontists who contributed based on their daily experiences with patients. Others were derived using insights from three large data models that identified common concerns about orthodontic risks as expressed by the general public. These two sources were integrated to obtain a more comprehensive perspective regarding the information needs and potential anxieties of individuals seeking orthodontic treatment.

For evaluating the effectiveness of AI chatbot responses, two different thresholds (low and high thresholds) were used to assess various levels of needs and risk management within the field. The low threshold applied to basic, low-risk interactions, such as providing general oral health advice or scheduling appointments where efficiency and convenience are more important than precision. The high threshold was applied to medical consultations requiring a high degree of professionalism and precision, such as providing specific treatment recommendations or analysis of complex cases. In these situations, the accuracy and reliability of chatbot responses are crucial to ensuring patient safety and treatment effectiveness.

In this study, Gemini performed best on the low-threshold validity test, whereas Claude AI and ChatGPT 4o performed equally. In the high-threshold validity test, Claude AI also exhibited the highest validity, but there was no statistically significant difference observed among the three AI chatbots. A previous study compared the performance between two AI chatbots: ChatGPT and Google Bard (now called Gemini), in answering general

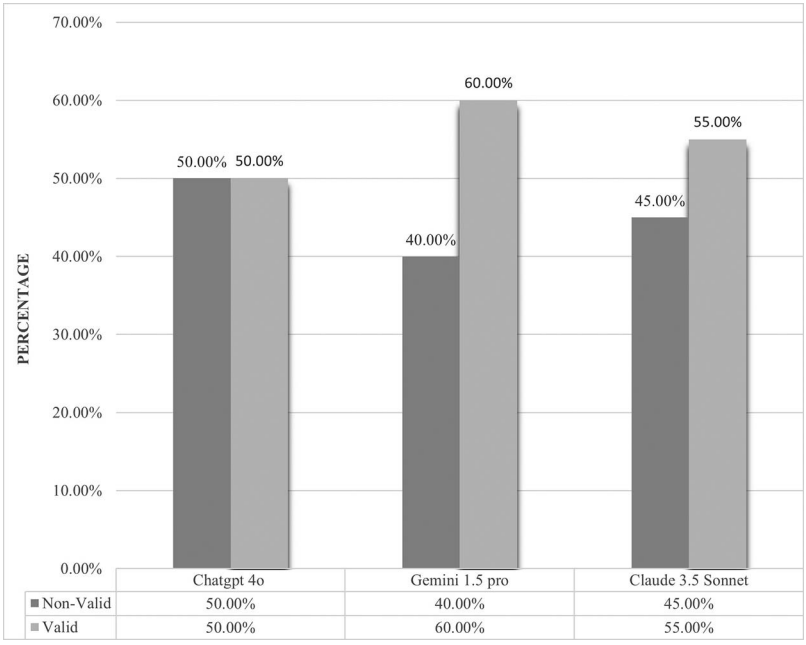

Figure 4. Analysis of valid and invalid responses in a high-threshold validity assessment.

orthodontic questions; both provided accurate and complete responses.[18] In other dental fields, a study on endodontic information found no significant differences in validity among Bing, ChatGPT 3.5, and Google Bard during low-threshold validity assessments.[19] However, significant differences were noted in the validity of ChatGPT 3.5 compared with Bing and Google Bard in high-threshold validity tests. The results of this study differed, likely due to two potential reasons. First, the evolution of LLMs led to further development of chatbots and their ability to generate responses,[20] likely leading to changes in responses compared with previous studies. Also, chatbots may behave differently across various fields due to different databases and algorithms.

In the low-threshold validity test, the responses of the three chatbots showed high validity. However, there was a significant decline in the effectiveness of their responses when assessed using high-threshold criteria. This finding highlighted how the quality of chatbot responses might vary significantly under criteria of varying rigor, especially in scenarios requiring more precise and specialized medical information. The chatbots had a high overall efficiency score. However, they may still make serious errors in certain specific responses, potentially misleading the public. For example, when addressing a question about orthodontics and temporomandibular joint (TMJ) disease, Claude AI did not emphasize that the relationship between orthodontic treatment and TMJ disorders is still not fully understood.[21] Instead, it agreed that orthodontic treatment could cause TMJ disorders, giving patients false preconceptions. Therefore, extra caution is required in medical fields, especially in specialized areas such as orthodontics. As each patient has specific conditions and risk factors, developing individualized treatment plans is essential. Patients must be discerning when relying on AI chatbots for medical decisions and work closely with healthcare professionals to create treatment plans that consider their unique individual needs.

Reliability, as a measure of consistency, is a key criterion for evaluating chatbot performance. The chatbots were built using deep learning models and they inherently exhibit a certain degree of randomness, implying that their responses may be unpredictable.[19] One study revealed that ChatGPT provided a different and faster response when the same question was asked again or at a different point in time.[22] Therefore, the present study assessed the reliability of three different chatbots in terms of consistency among responses to questions repeated at three different times. Results showed that all three AI chatbots demonstrated an acceptable level of reliability, which was greater than 0.7. Despite room for improvement, performance of these chatbots was satisfactory for providing consistent information about potential orthodontic risks. The current study was designed to evaluate the model multiple times (three repetitions) over a relatively short period of time (one day). The main purpose was to evaluate the consistency and reliability of the model over a short period of time to reflect its immediate performance in real-world applications.

Overall, all three chatbots achieved satisfactory levels of effectiveness and reliability. The null hypothesis was rejected. This emphasizes the need for cooperation between regulatory bodies and chatbot developers to ensure accuracy of information and prevent the spread of errors or misleading content to the public. In the future, orthodontic specialty associations may choose actively to form a team of experts to regularly evaluate the quality of orthodontic information provided by chatbots and widely disseminate these evaluation results to the public. This could lead to significant improvements in the accuracy and reliability of the information.

This study also had some limitations. First, only 20 questions were used, which was not enough to cover all orthodontic risks. Future studies should increase the sample size of questions to better assess LLM performance in orthodontics. Second, the team evaluating GQS scores was not independent from the question formulation group, which may have introduced some bias. Future research should consider using separate teams to enhance objectivity. In addition, this study evaluated consistency of performance over a relatively short period of time (one day). Future studies should consider longer-term evaluations (3 months) to better assess stability and reliability of the model.

CONCLUSIONS

- Gemini was the most effective and reliable AI chatbot in answering questions about potential orthodontic risks, followed by Claude AI and ChatGPT 4o.
- Although chatbots demonstrated reasonable reliability in providing orthodontic information, continuous improvement and customization are vital to optimize their effectiveness in clinical practice.
- Collaborative effort is essential for addressing ethical issues and guaranteeing the accuracy and credibility of the information provided by AI platforms.

ACKNOWLEDGMENTS

The data underlying this study are available upon request from the corresponding author. This study was supported by the China Oral Health Foundation Smile Teenagers Research Fund (Grant No. A2023-03) and the science research cultivation program of Stomatological Hospital, Southern Medical University (Grant No. PY2024030). The authors declare no competing or financial interests.

REFERENCES

1. Elkarmi R, Abu-Ghazaleh S, Sonbol H, Haha O, Al-Haddad A, Hassona Y. ChatGPT for parents' education about early childhood caries: a friend or foe? *Int J Paediatr Dent.* 2024. doi: 10.1111/ipd.13283. Online ahead of print
2. Umer F, Batool I, Naved N. Innovation and application of large language models (LLMs) in dentistry – a scoping review. *BDJ Open.* 2024;10(1):90.
3. Diniz-Freitas M, Lago-Méndez L, Limeres-Posse J, Diz-Dios P. Challenging ChatGPT-4V for the diagnosis of oral diseases and conditions. *Oral Dis.* 2025;31(2):701–706.
4. Huang H, Zheng O, Wang D, 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.
5. Suárez A, Jiménez J, Llorente de Pedro M, et al. Beyond the scalpel: assessing ChatGPT's potential as an auxiliary intelligent virtual assistant in oral surgery. *Comput Struct Biotech J.* 2024;24:46–52.
6. Shen Y, Heacock L, Elias J, et al. ChatGPT and other large language models are double-edged swords. *Radiology.* 2023; 307(2):e230163.
7. Demir GB, Süküt Y, Duran GS, Topsakal KG, Görgülü S. Enhancing systematic reviews in orthodontics: a comparative examination of GPT-3.5 and GPT-4 for generating PICO-based queries with tailored prompts and configurations. *Eur J Orthod.* 2024;46(2):cjae011.
8. Tanaka OM, Gasparello GG, Hartmann GC, Casagrande FA, Pithon MM. Assessing the reliability of ChatGPT: a content analysis of self-generated and self-answered questions on clear aligners, TADs and digital imaging. *Dent Press J Orthod.* 2023;28(5):e2323183.
9. Kurt Demirsoy K, Buyuk SK, Bicer T. How reliable is the artificial intelligence product large language model ChatGPT in orthodontics? *Angle Orthod.* 2024;94(6):602–607.
10. Kresevic S, Giuffrè M, Ajcevic M, Accardo A, Crocè LS, Shung DL. Optimization of hepatological clinical guidelines interpretation by large language models: a retrieval augmented generation-based framework. *npj Digital Medicine.* 2024; 7(1):102.
11. Bland JM, Altman DG. Cronbach's alpha. *BMJ.* 1997; 314(7080):572.
12. Perry J, Popat H, Johnson I, Farnell D, Morgan MZ. Professional consensus on orthodontic risks: what orthodontists should tell their patients. *Am J Orthod Dentofacial Orthop.* 2021;159(1):41–52.
13. Ustdal G, Guney AU. YouTube as a source of information about orthodontic clear aligners. *Angle Orthod.* 2020;90(3):419–424.
14. Tamošiūnaitė I, Vasiliauskas A, Dindaroğlu F. Does YouTube provide adequate information about orthodontic pain? *Angle Orthod.* 2023;93(4):403–408.
15. Dursun D, Bilici Geçer R. Can artificial intelligence models serve as patient information consultants in orthodontics? *BMC Med Inform Decis Mak.* 2024;24(1):211.
16. Kurt Demirsoy K, Buyuk SK, Bicer T. How reliable is the artificial intelligence product large language model ChatGPT in orthodontics? *Angle Orthod.* 2024;94(6):602–607.
17. Perez-Pino A, Yadav S, Upadhyay M, Cardarelli L, Tadinada A. The accuracy of artificial intelligence-based virtual assistants in responding to routinely asked questions about orthodontics. *Angle Orthod.* 2023;93(4):427–432.
18. Daraqel B, Wafaie K, Mohammed H, et al. The performance of artificial intelligence models in generating responses to general orthodontic questions: ChatGPT vs Google Bard. *Am J Orthod Dentofacial Orthop.*2024;165(6):652–662.
19. Mohammad-Rahimi H, Ourang SA, Pourhoseingholi MA, Dianat O, Dummer PMH, Nosrat A. Validity and reliability of artificial intelligence chatbots as public sources of information on endodontics. *Int Endod J.* 2024;57(3):305–314.
20. Kılınç DD, Mansız D. Examination of the reliability and readability of Chatbot Generative Pretrained Transformer's (ChatGPT) responses to questions about orthodontics and the evolution of these responses in an updated version. *Am J Orthod Dentofacial Orthop.* 2024;165(5):546–555.
21. Mohlin B, Axelsson S, Paulin G, et al. TMD in relation to malocclusion and orthodontic treatment. *Angle Orthod.* 2007; 77(3):542–548.
22. Vaishya R, Misra A, Vaish A. ChatGPT: Is this version good for healthcare and research? *Diabetes Metab Syndr.* 2023; 17(4):102744.