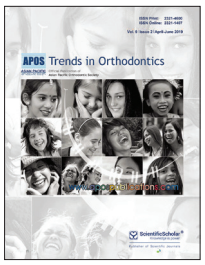
ScientificScholar®
Knowledge is power
Publisher of Scientific Journals

www.apospublications.com

APOS Trends in Orthodontics



Original Article

Are artificial intelligence based chatbots reliable sources for patients regarding orthodontics?

Tuğba Haliloğlu Özkan[1], Ahmet Hüseyin Acar[2],Enes Özkan[2], Mustafa Düzyol[3],Elif Aybüke Öztürk[4]

Departments of [1]Orthodontics, [2]Oral and Maxillofacial Surgery, [3]Restorative Dentistry, Faculty of Dentistry, Istanbul Medeniyet University, [4]Department of Orthodontics, Faculty of Dentistry, Marmara University, Istanbul, Turkey.

***Corresponding author:**
Tuğba Haliloğlu Özkan,
Department of Orthodontics,
Istanbul Medeniyet University,
Istanbul, Turkey.

dttuuba@gmail.com

Received: 08 August 2024
Accepted: 06 December 2024
EPub Ahead of Print: 06 January 2025
Published: 09 May 2025

DOI
10.25259/APOS_203_2024

Quick Response Code:

ABSTRACT

Objectives: The objective of this study was to conduct a comprehensive and patient-centered evaluation of chatbot responses within the field of orthodontics, comparing three prominent chatbot platforms: ChatGPT-4, Microsoft Copilot, and Google Gemini.

Material and Methods: Twenty orthodontic-related queries were presented to ChatGPT-4, Microsoft Copilot, and Google Gemini by ten orthodontic experts. To assess the accuracy and completeness of responses, a Likert scale (LS) was employed, while the clarity of responses was evaluated using a Global Quality Scale (GQS). Statistical analyses included One-way analysis of variance and *post-hoc* Tukey tests to assess the data, and a Pearson correlation test was used to determine the relationship between variables.

Results: The results indicated that ChatGPT-4 (1.69 ± 0.10) and Microsoft Copilot (1.68 ± 0.10) achieved significantly higher LS scores compared to Google Gemini (2.27 ± 0.53) ($P < 0.05$). However, the GQS scores, which were 4.01 ± 0.31 for ChatGPT-4, 3.92 ± 0.60 for Google Gemini, and 4.09 ± 0.15 for Microsoft Copilot, showed no significant differences among the three chatbots ($P > 0.05$).

Conclusion: While these chatbots generally handle basic orthodontic queries well, they show significant differences in responses to complex scenarios. ChatGPT-4 and Microsoft Copilot outperform Google Gemini in accurately addressing scenario-based questions, highlighting the importance of strong language comprehension, knowledge access, and advanced algorithms. This underscores the need for continued improvements in chatbot technology.

Keywords: Artificial intelligence, Chatbot, Orthodontics

INTRODUCTION

Traditional search engines (TSE), conversational intelligence virtual assistants (CIVA), and artificial ıntelligence-(AI-) based chatbots are all technological tools designed to provide information and assistance.[1,2] Each of these tools serves different purposes and meets various user needs in the area of information access and interaction. However, the utilization of AI-based chatbots for delivering health-related information to patients can offer specific advantages over TSE and CIVA, particularly in certain scenarios.[3,4]

AI-based chatbots possess several advantages, such as 24/7 availability, enabling patients to seek information at any time, providing instant responses to patient queries, and reducing the necessity to wait for appointments or responses from healthcare professionals.[5] In the context

of orthodontic practice, chatbots can educate patients about orthodontic procedures, assisting them in making informed decisions regarding their care. In addition, patients can use chatbots for quick clarifications or answers to minor concerns without the need for a formal appointment, thus saving time and potentially reducing healthcare costs. These devices also offer a private setting, avoiding potential embarrassment in face-to-face discussions. [2,3,6-8]

Although AI chatbots hold promise, their effectiveness must be supported by rigorous research and validation. Such research can provide a comprehensive perspective on the capabilities and limitations of AI-based chatbot technology when it comes to delivering sensitive orthodontic information, ultimately contributing to a more comprehensive understanding of their performance in healthcare contexts.

While there is a growing body of scientific studies investigating how accurately chatbots provide information to patients in various healthcare fields, to the best of our knowledge, there is currently no research examining the accuracy and effectiveness of AI-powered chatbots in responding to frequently asked questions by patients in the field of orthodontics.[9-14] Therefore, the aim of this study was to evaluate the accuracy and efficiency of three popular AI-powered chatbots, ChatGPT-4 (1), Microsoft Copilot (2), and Google Gemini (3), in responding to common inquiries posed by orthodontic patients.

MATERIAL AND METHODS

This study did not require ethical approval due to its publicly available nature. We identified a total of 20 questions that patients commonly asked about orthodontic treatment and posed them to three AI-based chatbots: ChatGPT-4, Microsoft Copilot, and Google Gemini [Table 1]. When formulating the queries, we deliberately avoided simple models, opting instead for more scenario-based questions to evaluate the capabilities of AI systems. We accomplished this by identifying the most frequently asked orthodontic queries on the "Quora Digest" platform, facilitating a comprehensive evaluation of chatbot performance in addressing real-world orthodontic problems. This platform is widely used for sharing and discussing trending topics, making it an ideal source for gathering questions that are representative of the types of inquiries users typically seek answers to in everyday situations.

The responses provided by the three platforms were recorded and assessed for accuracy, comprehensiveness, and clarity. Ten orthodontic experts, all of whom have extensive clinical experience and are regularly exposed to similar cases in their professional practice, reviewed the information to assess its clinical validity. To evaluate the accuracy and completeness level of each response, we employed a modified Likert scale (LS), customized for this study. This modified LS included five response options: (1) the chatbot provided accurate information, and the response covered all relevant aspects of the question; (2) the chatbot provided accurate information, but the response did not cover all relevant aspects of the question; (3) the chatbot provided inaccurate information, but the response covered all relevant aspects of the question; (4) the chatbot provided inaccurate information, and the response did not cover all relevant aspects of the question; and (5) the chatbot couldn't provide an answer.

To assess the clarity and flow of the responses, we employed the Global Quality Scale (GQS) by Bernard *et al*.[15] [Table 2]. The score assigned to each device was calculated by averaging the ratings provided by the experts for each response.

RESULTS

Statistical Package for the Social Sciences (SPSS) for Windows, version 23.0 (SPSS Inc., Chicago, IL, USA), was utilized for all statistical analyses, with the level of significance set at $P < 0.05$. The normality of the data was assessed through the Q-Q plot diagram and Shapiro–Wilk test, revealing that all variables were normally distributed. Consequently, we conducted a one-way analysis of variance and *post-hoc* analyses.

The mean LS and GQS scores for each query of the three chatbots are presented in [Table 1]. The results indicated that ChatGPT-4 (1.69 ± 0.10) and Microsoft Copilot (1.68 ± 0.10) achieved significantly higher Likert scores than Google Gemini (2.27 ± 0.53) ($P < 0.05$). However, the GQS scores were as follows: 4.01 ± 0.31 for ChatGPT-4, 3.92 ± 0.60 for Google Gemini, and 4.09 ± 0.15 for Microsoft Copilot. Notably, no significant differences were observed among the three chatbots in terms of GQS ($P > 0.05$) [Tables 3 and 4].

Furthermore, we found moderate negative correlations between LS and GQS scores for ChatGPT-4 (r = −0.250), mild negative correlations for Microsoft Copilot (r = −0.088), and moderate positive correlations for Google Gemini (r = 0.466).

DISCUSSION

Over the years, AI has been a driving force in advancing digital healthcare.[16,17] AI-powered tools in the field of dentistry have proven to be valuable for analyzing medical images, aiding in the diagnosis of conditions such as dental caries, periodontitis, and implants, as well as assisting in the planning of oral and maxillofacial surgeries.[18-22]

In the field of orthodontics, neural networks can help diagnose and plan treatments, mark cephalometric landmarks, analyze anatomy, assess growth and development, and evaluate treatment outcomes.

Table 1: The Likert scale and Global Quality Scale scores of the queries that were posed to ChatGPT-4, Microsoft Copilot, and Google Gemini.

	ChatGPT-4	Google Gemini	Copilot	*P*-value	ChatGPT-4	Google Gemini	Copilot	*P*-value
1. My braces caused a sore inside my cheek. What can I do for it?	1.70	1.70	1.70	1.00	3.20	4.40	3.20	0.000*
2. What happens if I does not wear my retainer for a month?	1.00	1.30	1.70	0.002*	4.60	4.60	4.60	0.728
3. My bracket has broken and I have swallowed it. What can I do for it?	1.30	1.60	1.40	0.413	3.80	4.20	4.60	0.058
4. I still have pain for 10 days since my first bracket placament visit. Is it normal?	1.70	3.00	2.00	0.000*	4.60	3.20	4.20	0.002*
5. I have been wearing braces for 1.5 years and my upper front 4 teeth are still not aligned. What can I do?	3.00	2.00	1.70	0.001*	3.40	3.80	3.00	0.085
6. My bracket has loosen. What happens if I does not visit my orthodontist?	1.60	1.30	1.70	0.189	4.40	4.40	5.00	0.039*
7. If I does not have my wisdom teeth extracted, will my teeth get stuck again after my orthodontic treatment?	2.00	2.40	1.40	0.000*	4.60	4.20	4.80	0.091
8. I did not wear my clear aligners for 3 days and it does not fit now. Should I force to fit it?	2.40	3.00	2.00	0.002*	3.40	3.20	4.60	0.004*
9. I got braces about a week ago and my tooth to the left of my two front teeth feels loose .Is this a problem or will it go away?	1.40	2.80	1.70	0.003*	4.80	3.80	4.20	0.112
10. I-have-braces-to-straighten-my-front-teeth-but-after-about-a-month-I-start-to-notice-that-my-molars-are-slanting-inwards-Is-this-normal	1.00	3.10	1.00	0.000*	4.60	3.00	3.80	0.000*
11. I have had braces for 8 months. Do I still have to wear braces to complete the given duration even though the teeth look fine and aligned?	3.00	2.40	1.30	0.000*	4.00	3.60	4.80	0.006*
12. I have been wearing braces for 6 months. Before treatment, my teeth were very crooked, and now they are straight. But now my two front teeth are not aligned with the center. This was not a problem before. Is this normal?	1.60	3.10	1.70	0.000*	4.00	4.40	4.60	0.074
13. My jaw is too far back. Can I change that by just moving my jaw with my muscles and trying to let them get used to it? Or should I get surgery?	2.00	3.00	1.60	0.000*	3.40	4.20	4.20	0.088
14. Is it true that braces will make my bite, smile and jaw narrower, and pull my lower jaw inward and cause sleep apnea?	2.40	2.80	3.00	0.002*	4.60	3.40	3.80	0.010
15. Can I smoke once in a while after having jaw surgery?	1.60	3.10	1.30	0.000*	3.40	4.40	4.80	0.002*
16. Does it get normal to deal with the permanent numbness after double jaw surgery?	1.30	2.40	1.70	0.005*	3.60	3.20	3.40	0.623

(*Contd...*)

Table 1: (*Continued*)

	ChatGPT-4	Google Gemini	Copilot	P-value	ChatGPT-4	Google Gemini	Copilot	*P*-value
17. After jaw advancement surgery 5 years ago to reduce sleep apnea, I had much facial swelling. I still have cheeks that are puffy and asymmetrical (the right side is worse). How can this be explained? Can any treatment correct it (surgically or otherwise)?	1.30	2.50	2.60	0.004*	4.00	3.40	2.20	0.000*
18. Is it possible to get my teeth fixed with just braces without having to go through an underbite surgery?	1.30	1.70	2.30	0.000*	5.00	4.40	3.40	0.000*
19. What orthognatic surgery would I need If both my upper and lower jaws are too narrow due to which I have sleep apnea and breathing problems and also my jawline is not visible at all?	1.30	1.00	1.30	0.165	3.40	4.00	4.60	0.002*
20. I have underbite and I needed double jaw surgery when I was 15 but my parents refused and instead, I got braces. Now, I am older and I can choose for myself. Is it possible to get it reversed and to get surgery?	1.00	1.30	2.30	0.000*	3.40	4.60	4.20	0.010

*Statistically significant

Table 2: The Global Quality Scale score description.

Score 1	Poor quality, poor flow of the site, most information missing, not at all useful for patients
Score 2	Generally poor quality and poor flow, some information listed but many important topics missing, of very limited use to patients
Score 3	Moderate quality, suboptimal flow, some important information is adequately discussed but others poorly discussed, somewhat useful for patients
Score 4	Good quality and generally good flow, most of the relevant information is listed, but some topics are not covered, useful for patients 5 Excellent quality and excellent flow, very useful for patients
Score 5	Excellent quality and excellent flow, very useful for patients

Table 3: Mean, standard deviation, and range of the Likert scale scores.

	Min	Max	Mean±Standard deviation	*P*-value
ChatGPT-4[a]	1.60	1.85	1.6950±0.10916	0.000*
Google Gemini[a,b]	1.50	2.65	2.2750±0.53658	
Copilot[b]	1.55	1.80	1.6850±0.10288	

Superscripts indicate a statistically significant difference between the groups (*post-hoc* Tukey test); *Statistically significant

Table 4: Mean, standard deviation, and range of the Global Quality Scale scores.

Global quality scores				
	Min	Max	Mean±Standard deviation	*P*-value
ChatGPT-4	3.50	4.40	4.0100±0.31163	0.645
Google Gemini	3.00	4.50	3.9200±0.60516	
Copilot	3.85	4.30	4.0900±0.15420	

P<0.05 Statistically significant

In the present study, we examined the performance of three individual AI-powered chatbot systems – ChatGPT-4, Microsoft Copilot, and Google Gemini – in providing responses to orthodontic-related queries from 10 orthodontic experts. Our evaluation criteria included LS to assess the accuracy and completeness of responses and GQS to evaluate the clarity of the responses.

Microsoft Copilot, Google, Google Gemini, and ChatGPT-4 were selected due to their current prominence and widespread use in the field of conversational AI. These models represent the most advanced and commonly accessed systems at this time, making them highly relevant for the purpose of our study. We have also highlighted the importance of evaluating the leading AI systems in real-world contexts, as these are the

models that users are most likely to engage with when seeking information on specialized topics such as orthodontics. This adjustment aims to reinforce the reasoning behind our selection and to make the decision clearer to the reader.

ChatGPT-4 is based on a large pre-trained language model that can create fluent and varied texts using a deep neural network. Microsoft Copilot is a web search engine that uses different natural language processing (NLP) techniques, such as understanding queries, retrieving documents, and summarizing them to give clear and informative answers. However, Google Gemini is still in the early stages of development and is mainly designed to create creative content such as poetry and lyrics.[10,12]

The results of our study indicate that there were no significant differences between the responses provided by the three chatbots, ChatGPT-4, Google Gemini, and Microsoft Copilot, for the three straightforward, patterned questions posed (N:1, 3, 6) ($P > 0.05$) [Table 1]. These questions revolved around common orthodontic issues, including discomfort caused by braces, broken brackets, and loose brackets. This consistency suggests that these chatbots, despite differences in their underlying algorithms or training data, are able to provide uniform and reliable information for straightforward orthodontic queries. This finding is promising as it implies that users can expect consistent and accurate responses to their simple patterned questions regardless of the specific chatbot they interact with.

When the LS scores of 3 chatbots are examined, ChatGPT-4 (1.69 ± 0.10) and Microsoft Copilot (1.68 ± 0.10) received mean scores falling within the range of categories 1–2. This indicates that these chatbots generally provided accurate information along with comprehensive responses. However, each platform exhibited strengths in different areas: ChatGPT-4 excelled in generating comprehensive responses, while Microsoft Copilot offered concise and clear explanations. In contrast, Google Gemini (2.27 ± 0.53) received mean scores falling in categories 1–3, signifying a wider variability in its performance compared to the other two chatbots. This variability indicates that users should be cautious when using this platform, as Google Gemini occasionally offers inaccurate information and less comprehensive answers.

ChatGPT-4 has demonstrated a good ability to provide coherent and comprehensive answers to orthodontic questions, regardless of their complexity. Microsoft Copilot, too demonstrated clarity in its responses, providing comprehensive informative information. The superior performance of ChatGPT-4 and Microsoft Copilot could be attributed to their advanced NLP capabilities, which enable them to understand complex queries and generate relevant and coherent responses. However, Google Gemini, while user-friendly in its interface, is a chatbot that uses a simpler approach of matching keywords and phrases from the query to a predefined database of answers. This method may limit the ability of Google Gemini to handle queries that require more reasoning and synthesis. In addition, the responses provided by Google Gemini sometimes seemed to veer slightly off-topic or include extraneous information, which could potentially confuse users. These results are consistent with previous research, which has shown that large language models such as ChatGPT-4 and Microsoft Copilot are capable of generating accurate and informative responses to a wide range of queries.[22-25]

Regarding response flow and clarity, as evaluated through GQS scores, the orthodontic experts perceived all three chatbots in a similar manner. This observation implies that, despite variations in response accuracy and comprehensiveness, the clarity and flow of explanations and information presentation remain consistently uniform among the platforms examined. It is important to note that these models are still under development, and their performance can vary depending on the specific task and the quality of the training data. As AI continues to advance, these platforms have the potential to play a pivotal role in enhancing patient education and information dissemination in the field of orthodontics and healthcare more broadly. The ongoing refinement of AI algorithms will likely lead to even more accurate, comprehensive, and user-friendly responses, further bridging the gap between patients and reliable medical information.

It is important to remember that the field of AI is dynamic, and the accuracy of chatbot responses may evolve over time as models are updated. This study is limited by the specific sample of queries used, the potential for evolving AI capabilities, and the subjectivity of expert review.

CONCLUSION

In summary, while AI-based chatbots can provide valuable information and guidance, their accuracy depends on the quality of their training data, the involvement of medical experts, continuous updates, and clear communication of limitations. Therefore, users should always verify medical information with qualified professionals before making decisions about their orthodontic treatment.

Acknowledgements: The authors would like to express their gratitude to the orthodontic experts who provided valuable insights during the development of the orthodontic queries used in this study.

Ethical approval: Institutional Review Board approval is not required.
Declaration of patient consent: Patient's consent is not required as there are no patients in this study.
Financial support and sponsorship: None.
Conflicts of interest: There are no conflicts of Interest.
Use of artificial intelligence (AI)-assisted technology for manuscript preparation: The authors confirm that there was no use of artificial intelligence (AI)-assisted technology for assisting

in the writing or editing of the manuscript and no images were manipulated using AI.

REFERENCES

1. Priya B, Bhanu V, Sharma V. Exploring users' adoption intentions of intelligent virtual assistants in financial services: An anthropomorphic perspectives and socio-psychological perspectives. Comput Human Behav 2023;148:107912.
2. Milne-Ives M, de Cock C, Lim E, Shehadeh MH, de Pennington N, Mole G, *et al.* The effectiveness of artificial intelligence conversational agents in health care: Systematic review. J Med Internet Res 2020;22:e20346.
3. De Cock C, Milne-Ives M, van Velthoven MH, Alturkistani A, Lam C, Meinert E. Effectiveness of conversational agents (virtual assistants) in health care: Protocol for a systematic review. JMIR Res Protoc 2020;9:e16934.
4. Agarwal S, Agarwal B, Gupta R. Chatbots and virtual assistants: A bibliometric analysis. Library Hi Tech 2022;40:1013-30.
5. Gao M, Liu X, Xu A, Akkiraju R. Chat-XAI: A new chatbot to explain artificial intelligence. In: Intelligent systems and applications: Proceedings of the 2021 Intelligent systems conference (IntelliSys). Vol. 3. Germany: Springer International Publishing; 2022.
6. Pham KT, Nabizadeh A, Selek S. Artificial intelligence and chatbots in psychiatry. Psychiatr Q 2022;93:249-53.
7. Aggarwal A, Tam CC, Wu D, Li X, Qiao S. Artificial intelligence-based chatbots for promoting health behavioral changes: Systematic review. J Med Internet Res 2023;25:e40789.
8. Chew HS. The Use of artificial intelligence-based conversational agents (chatbots) for weight loss: Scoping review and practical recommendations. JMIR Med Inform 2022;10:e32578.
9. Kumari A, Kumari A, Singh A, Singh SK, Juhi A, Dhanvijay AD, *et al.* Large language models in hematology case solving: A comparative study of ChatGPT-4-3.5, Google Gemini, and Microsoft copilot. Cureus 2023;15:e43861.
10. Zúñiga Salazar G, Zúñiga D, Vindel CL, Yoong AM, Hincapie S. Efficacy of AI chats to determine an emergency: A comparison between OpenAI's ChatGPT-4, Google Gemini, and Microsoft Copilot AI Chat. Cureus 2023;15:e45473.
11. Al-Ashwal FY, Zawiah M, Gharaibeh L, Abu-Farha R, Bitar AN. Evaluating the sensitivity, specificity, and accuracy of ChatGPT-4-3.5, ChatGPT-4-4, Copilot AI, and Google Gemini against conventional drug-drug interactions clinical tools. Drug Healthc Patient Saf 2023;15:137-47.
12. Dhanvijay AK, Pinjar MJ, Dhokane N, Sorte SR, Kumari A, Mondal H. Performance of large language models (ChatGPT-4, Copilot Search, and Google Gemini) in solving case vignettes in physiology. Cureus 2023;15:e42972.
13. Oh YJ, Zhang J, Fang ML, Fukuoka Y. A systematic review of artificial intelligence chatbots for promoting physical activity, healthy diet, and weight loss. Int J Behav Nutr Phys Act 2021;18:160.
14. Rao A, Pang M, Kim J, Kamineni M, Lie W, Prasad AK, *et al.* Assessing the utility of ChatGPT throughout the entire clinical workflow: Development and usability study. J Med Internet Res 2023;25:e48659.
15. Bernard A, Langille M, Hughes S, Rose C, Leddin D, Veldhuyzen van Zanten S. A systematic review of patient inflammatory bowel disease information resources on the World Wide Web. Am J Gastroenterol 2007;102:2070-7.
16. Kurian N, Cherian JM, Sudharson NA, Varghese KG, Wadhwa S. AI is now everywhere. Br Dent J 2023;234:72.
17. Johnson SB, King AJ, Warner EL, Aneja S, Kann BH, Bylund CL. Using ChatGPT-4 to evaluate cancer myths and misconceptions: Artificial intelligence and cancer information. JNCI Cancer Spectr 2023;7:pkad015.
18. Mohammad-Rahimi H, Motamedian SR, Rohban MH, Krois J, Uribe SE, Mahmoudinia E, *et al.* Deep learning for caries detection: A systematic review. J Dent 2022;122:104115.
19. Urban R, Haluzová S, Strunga M, Surovková J, Lifková M, Tomášik J, *et al.* AI-assisted CBCT data management in modern dental practice: Benefits, limitations and innovations. Electronics 2023;12:1710.
20. Revilla-León M, Gómez-Polo M, Barmak AB, Inam W, Kan JY, Kois JC, *et al.* Artificial intelligence models for diagnosing gingivitis and periodontal disease: A systematic review. J Prosthet Dent 2023;130:816-24.
21. Mohammad-Rahimi H, Motamedian SR, Pirayesh Z, Haiat A, Zahedrozegar S, Mahmoudinia E. *et al.* Deep learning in periodontology and oral implantology: A scoping review. J Periodontal Res 2022;57:942-51.
22. Minnema J, Ernst A, van Eijnatten M, Pauwels R, Forouzanfar T, Batenburg KJ, *et al.* A review on the application of deep learning for CT reconstruction, bone segmentation and surgical planning in oral and maxillofacial surgery. Dentomaxillofac Radiol 2022;51:20210437.
23. Kühnisch J, Meyer O, Hesenius M, Hickel R, Gruhn V. Caries detection on intraoral images using artificial intelligence. J Dent Res 2022;101:158-65.
24. Roos J, Kasapovic A, Jansen T, Kaczmarczyk R. Artificial intelligence in medical education: Comparative analysis of ChatGPT-4, copilot, and medical students in Germany. JMIR Med Educ 2023;9:46482.
25. Dursun D, Bilici Geçer R. Can artificial intelligence models serve as patient information consultants in orthodontics? BMC Med Inform Decis Mak 2024;24:211.

How to cite this article: Özkan Haliloğlu T, Acar AH, Özkan E, Düzyol M,Öztürk EA. Are artificial intelligence-based chatbots reliable sources for patients regarding orthodontics? APOS Trends Orthod. 2025;15:141-6. doi: 10.25259/APOS_203_2024