Journal of Radiology Medicine in

Original Article

Doi: 10.51271/JRM-0051

Comparison of chatbots for patient information on head and neck imaging techniques in terms of scientific quality and readability

Meltem Horoz*[1], Ayşegül Cavnar[2]

[1]Department of Oral and Maxillofacial Radiology, Faculty of Dentistry, Kırıkkale University, Kırıkkale, Turkiye
[2]Department of Prosthodontics, Faculty of Dentistry, Gazi University, Ankara, Turkiye

Received: 15.06.2026 ◆ **Accepted**: 17.08.2026 ◆ **Published**: 19.09.2026

Cite this article: Horoz M, Cavnar A. Comparison of chatbots for patient information on head and neck imaging techniques in terms of scientific quality and readability. *J Radiol Med.* 2026;3(3):54-58. doi:10.51271/JRM-0051

***Corresponding Author:** Meltem Horoz, meeltemhoroz@hotmail.com

ABSTRACT

Aims: This study aimed to evaluate the scientific quality and readability of the responses of ChatGPT and Microsoft Copilot to patient questions about head and neck imaging methods in dentistry.

Methods: Twenty patient-focused questions were directed to both chatbots. Responses were evaluated independently by two experts using a 7-point Likert scale, and readability levels were analyzed using the Flesch Reading Ease Score and Flesch-Kincaid Grade Level (FKGL) indices. Inter-expert agreement was assessed using Cohen's Kappa, and differences between chatbots were assessed using an independent-samples t-test.

Results: Statistically significant but weak agreement was observed among experts (κ=0.399, $p<0.001$). No significant difference was found between ChatGPT and Microsoft Copilot in terms of Flesch Reading Ease Score index and expert scores ($p>0.05$). However, FKGL scores were significantly higher in Microsoft Copilot ($p=0.025$), indicating more linguistically complex responses.

Conclusion: ChatGPT and Microsoft Copilot demonstrated similar scientific quality and readability based on FRES scores. However, Microsoft Copilot generated more linguistically complex responses than ChatGPT, as reflected by its significantly higher FKGL scores, indicating that a higher educational level may be required for comprehension. In conclusion, AI-based chatbots can be useful as supportive tools in patient information and education, but expert supervision remains crucial in this area.

Keywords: Artificial Intelligence, chatbots, patient education as topic, radiologic health

INTRODUCTION

Artificial Intelligence (AI) is a set of technologies that enable computer systems to perform various tasks by mimicking human cognitive abilities. These systems can learn from data, analyze information, and generate appropriate outputs. They can exhibit functions similar to human intelligence in processes such as decision-making and problem-solving.[1] Large language models (LLMs) are AI systems trained with data from articles, books, and online resources. These models utilize deep learning-based neural network architectures to model complex connections between words in textual data.[2] LLMs use deep learning techniques to process language, enabling them to perform a wide range of tasks such as translation, summarizing, and answering questions. Their ability to produce coherent and meaningful texts makes them valuable tools in fields like education and healthcare. Research has shown that LLMs also help medical professionals and patients access and interpret large amounts of data efficiently and accurately. By providing information quickly, they can improve patient care and education and contribute to the medical workflow.[3] Among LLMs, chatbots are conversational AI systems capable of simulating human-like conversation through text and voice. Thanks to their natural language processing capabilities, a subfield of AI, chatbots can understand and answer human questions. One of the key features of chatbots in the healthcare field is their ability to provide accessible information to users by managing data and answering specific, focused questions.[4] In November 2022, OpenAI (San Francisco, USA) introduced the Chat Generative Pre-trained Transformer (ChatGPT). This LLM, which combines AI and natural language processing technologies, can analyze dialogues and generate appropriate response texts.[5] Numerous studies have evaluated

the performance of chatbots in various medical specialties. In particular, ChatGPT's successful performance on the United States Medical Licensing Exam without additional training has highlighted the potential of AI-based chatbots in medicine and increased interest in their use in healthcare. In dentistry, studies evaluating chatbot performance have increased since 2023, with studies conducted across specialties such as general dentistry, periodontology, orthodontics, and oral, dental, and maxillofacial surgery.[6] Although there are studies in the literature examining the performance of ChatGPT and other chatbots in dentistry, evaluations of the scientific accuracy, readability, and user satisfaction of their responses to patient questions about head and neck imaging methods remain limited. Providing patients with incorrect or incomplete information about imaging methods and radiation exposure can lead to unnecessary anxiety, false clinical expectations, and negatively impact health decisions. Therefore, evaluating the accuracy of the information provided by AI-powered systems is important. Studies in the literature have evaluated the performance of AI chatbots in various areas of medicine and dentistry. However, there are a limited number of studies on the quality and readability of chatbot responses to patient-focused questions about head and neck imaging methods in dentistry. The aim of this study was evaluate chatbots' responses to questions patients may ask about head and neck imaging methods used in dentistry. In this context, the chatbot responses were analyzed by experts for scientific adequacy, readability, and user satisfaction to assess their potential for patient education and information.

METHODS

Ethical approval was not required for this study because no patient data or personal information was used and the study aimed solely to analyze freely accessible information sources. This study was prepared considering the basic questions that patients might have regarding head and neck imaging methods in dentistry. A set of 20 questions was deemed sufficient to represent the most frequently asked questions by patients regarding head and neck imaging and to enable a comprehensive evaluation of AI chatbot responses. The questions were selected by examining the "Frequently Asked Questions" and "People Also Ask" sections on Google; content from patient information platforms was also used. The selected questions were chosen by consensus by two expert dentists based on the current literature, 'White and Pharoah's Oral Radiology: Principles and Interpretation', and their clinical significance. The questions were posed to ChatGPT-4.5 and Microsoft Copilot in March 2026. All questions were posed to both chatbots in the same sessions, in incognito mode, using identical wording, and in English. No additional guidance were provided during the response generation process, and each question was evaluated independently. The responses were recorded in plain text format by a third evaluator. During the evaluation process, evaluators were not informed which AI model the responses belonged to. Chatbot responses were independently evaluated by two expert evaluators. The scientific accuracy and quality of the responses were assessed using 'White and Pharoah's Oral radiology: Principles and Interpretation', a reference source in the field of oral and maxillofacial radiology (**Table 1**).

Table 1. Patient questions regarding head and neck imaging methods are directed to chatbots

1. In dentistry and medicine, in what situations is tomography (CBCT/CT) preferred, and what kind of information does it provide?
2. In what situations is MRI necessary in dentistry?
3. Is MRI safe in patients with dental implants?
4. In what clinical situations is ultrasonography used in dentistry?
5. What are the main differences between ultrasonography and MRI?
6. What are the differences between CBCT and CT?
7. Are patients exposed to radiation during ultrasonography and MRI imaging?
8. What precautions are taken to protect patients from radiation during radiological imaging?
9. What level of radiation are patients exposed to during tomography (CBCT/CT) scans in dentistry and medicine?
10. Is tomography (especially CBCT) necessary before dental implant planning?
11. How many times can a patient have a CT scan in a year, and what factors determine this decision?
12. Are there any special precautions patients should take after radiographic imaging?
13. Do metal materials in the body affect image quality during a CT scan (CBCT/CT)?
14. Is MRI or CT more suitable for evaluating TMJ problems?
15. Can all diseases be definitively diagnosed with CT/CBCT in medicine and dentistry?
16. Which methods can be safely used for head and neck imaging during pregnancy?
17. What are the differences in cost and accessibility between CBCT, CT, MRI, and ultrasound methods?
18. What risks does the incorrect or unnecessary use of imaging methods pose to patients?
19. What factors do clinicians consider when choosing the appropriate imaging method for a patient?
20. What are the differences between different imaging methods (CBCT, CT, MRI, ultrasound) in terms of scanning times and patient comfort?

CBCT: Cone-beam computed tomography, CT: Computed tomography, MRI: Magnetic resonance imaging, TMJ: Temporomandibular joint

Each response was evaluated using a single overall score on a 7-point Likert scale (1=very poor, 7=excellent), assessing both scientific accuracy and comprehensibility. The readability of responses was assessed using the Flesch Readability Ease Score (FRES) and the Flesch-Kincaid Grade Level (FKGL) indices. Scientific accuracy and the quality of information in the responses were assessed using a 7-point Likert scale, aligned with assessment approaches used in similar studies. On the scale, 1 point was defined as 'very poor (incorrect and potentially harmful)', 2 points as 'poor (contains serious errors and misinformation)', 3 points as 'poor (contains significant omissions and some errors)', 4 points as 'medium (partially correct but contains significant omissions)', 5 points as 'good (contains minor errors or limited omissions)', 6 points as 'very good (almost error-free and contains very minor omissions)', and 7 points as 'excellent (completely accurate, complete, and reliable)'. The data collected in the study were statistically analyzed using IBM SPSS 29.0 software. Skewness and kurtosis values for the FRES, FKGL, and expert scoring variables ranged from -2 to +2; histograms showed a normal distribution (George & Mallery, 2010). Cohen's Kappa was calculated to assess the level of agreement among expert scores. An independent-samples t-test was used to compare index scores and expert-evaluation variables between ChatGPT and Microsoft Copilot. The significance level was set at $p<0.05$ for all statistical analyses.

RESULTS

Inter-Expert Agreement

Cohen's kappa was used to assess the agreement between the two experts' evaluations of the AI responses. As shown in **Table 2**, Cohen's kappa analysis indicated weak (McHugh, 2012) but statistically significant agreement between the scoring styles for AI responses evaluated by the experts (κ=0.399, p<0.001).

Table 2. Cohen's kappa analysis table of agreement between how two experts categorize AI responses

		Expert 2				
		Score 5	Score 6	Score 7	Kappa (κ)	p
Expert 1	Score 5	3	4	0	0.399	<0.001
	Score 6	3	14	5		
	Score 7	3	0	8		

AI: Artificial Intelligence

Chatbot Comparisons

Independent-samples t-tests were conducted to examine whether there was a significant difference between ChatGPT and Microsoft Copilot in FRES, FKGL, and expert scoring. The expert score was created by averaging the two experts' scores. According to the results, there was no significant difference between ChatGPT and Microsoft Copilot in FRES (p=0.815) or expert scoring (p=0.793). On the other hand, Microsoft Copilot's scores were found to be statistically significantly higher than ChatGPT's FKGL scores (p=0.025) (**Table 3**).

Table 3. Independent samples t-test table for comparing various index values between ChatGPT and Copilot

	ChatGPT	Copilot			
	Mean±SD	Mean±SD	t	p	Cohen's d
FRES	32.40±14.39	31.35±13.79	0.236	0.815	0.075
FKGL	10.05±2.04	11.55±2.01	-2.342	0.025	-0.741
Mean expert rating	6.13±0.60	6.08±0.59	0.265	0.793	0.084

SD: Standard deviation, FRES: Flesch Reading Ease Score, FKGL: Flesch–Kincaid Grade Level

DISCUSSION

The main findings of the present study showed that the responses of ChatGPT and Microsoft Copilot to questions regarding head and neck imaging methods in dentistry were similar in terms of scientific quality and user satisfaction. A statistically significant but weak level of agreement was found between expert evaluations. Furthermore, while no significant difference was found between the two chatbots in FRES and expert readability scores, FKGL scores were significantly higher for Copilot. This suggests that Microsoft Copilot's responses may have a more complex language structure. The low level of agreement observed among experts indicates that the evaluation of AI-generated health information is somewhat subjective. This finding may be influenced by the evaluators' clinical experience. However, the differences in scores between evaluators were mostly at the one-point level and did not indicate a significant divergence of opinion in clinical evaluations regarding the overall quality of the responses. Therefore, the low Cohen's kappa coefficient should not be interpreted as indicating a clinically significant discrepancy between evaluations. Similarly, previous studies have reported that chatbot responses can receive different scores depending on experts' experience and evaluation criteria.

Deep learning models have been increasingly used in natural language processing and computer vision. For this purpose, pre-trained LLMs such as ChatGPT and Microsoft Copilot have demonstrated capabilities in text generation and comprehension. The rate of online access to medical and health information is high, and chatbots are frequently used for this purpose.[7] With the development of chatbots, it is possible for patients and healthcare professionals to use these resources to learn about the benefits and risks of different diseases and imaging techniques.[8] However, the ability of AI systems to generate inaccurate, incomplete, or outdated information remains a significant concern for patient safety. A comprehensive evaluation of chatbots' performance in providing accurate and safe information is critical to determining the benefits of these technologies for clinical decision-making and patient education.[9] In recent years, magnetic resonance imaging (MRI), a non-invasive method that does not involve ionizing radiation, has gained prominence in dentistry for imaging soft tissues in the head and neck region, thanks to technical advancements and the development of new sequences. However, its use in dentistry is limited by constraints such as cost, accessibility, and artifact generation.[10,11] The present study also evaluated imaging methods that do not involve ionizing radiation, such as MRI. Providing accurate patient information about the advantages and limitations of MRI, which has seen increased use in the head and neck region in recent years, is crucial. Therefore, it is important to determine whether AI-based systems can provide reliable information about imaging methods that are gaining new applications, such as MRI. In their study, Zhang et al.[7] posed ultrasound-related questions in English and Chinese to the ChatGPT and ERNIE Bot chatbots and evaluated how responses varied by model, language, and question type. They reported that the ERNIE Bot generally performed better than ChatGPT, although both models experienced a lower performance on English questions. Furthermore, they emphasized that while the chatbots were successful in providing basic information, they were more limited in providing diagnostic interpretations. Zhang et al.[7] reported that chatbot performance can vary depending on language and question type. In the present study, although all questions were in English, no significant difference was found between ChatGPT and Microsoft Copilot in terms of overall quality scores. This suggests that the model's performance may converge toward patient-focused, informative questions about imaging methods. In their study, Patil et al.[8] evaluated AI chatbots' responses to scenarios involving imaging modalities such as CT and MRI, considering risks, benefits, alternative imaging options, and the potential consequences of not using contrast agents. ChatGPT has been reported to provide more accurate responses than Google Bard. However, it was emphasized that both models can provide incomplete or misleading information in critical situations and are limited in terms of clinical decision-making. However, in

the present study, ChatGPT and Microsoft Copilot showed similar performance. Furthermore, differences in results across studies conducted at different times may be due to the ongoing updates of AI models. In a study by Mohammad-Rahimi et al.[9] on chatbots, controversial topics in the literature and frequently encountered questions requiring expert knowledge in clinical practice in oral, dental, and maxillofacial radiology were used. The responses from 6 different chatbots were evaluated for quality by expert radiologists, and GPT-4 was reported to perform best. However, the study noted that some of the references provided by the chatbots were unverified. Similar to the present study, this research evaluated the performance of chatbots in patient education and information using frequently asked questions about head and neck imaging methods. The questions used in the present study focused on imaging methods and radiation safety issues that patients frequently wonder about in daily clinical practice. In this regard, the study contributes to the evaluation of chatbots from a patient education perspective.

In this study, the higher FKGL scores found in Copilot indicate that this model tends to use longer sentences and more complex word structures. A high FKGL score suggests that a higher educational level is required to understand the text. However, the lack of a significant difference between the two chatbots in FRES scores suggests that both models exhibit similar reading ease. When evaluated in terms of patient education, lower FKGL values may contribute to easier understanding of information by a wider patient population. The findings show that ChatGPT and Microsoft Copilot can serve as supportive tools for informing patients about head and neck imaging methods, but the information provided by these systems should be verified by healthcare professionals. Future healthcare-focused AI applications should prioritize not only information accuracy but also readability and user-friendly content creation.

Limitations

This study has some limitations. The evaluation was performed using only the ChatGPT and Microsoft Copilot versions available in March 2026. Since AI models are constantly updated, findings may vary across different versions. Furthermore, the study included a limited number of patient questions related to head and neck imaging techniques. Using a broader set of questions might lead to different results. Future studies should include a wider and more diverse set of patient questions, evaluate additional AI chatbots, and assess chatbot performance in different languages and clinical scenarios to contribute to a more comprehensive evaluation of the models' strengths and weaknesses in terms of patient education.

CONCLUSION

In this study, the responses of ChatGPT and Microsoft Copilot to frequently asked patient questions about head and neck imaging techniques were evaluated in terms of scientific accuracy and readability. The findings showed that, according to expert assessments, both chatbots provided similar levels of scientific accuracy and response quality. No significant difference was found between the two models in terms of FRES scores. However, Microsoft Copilot produced significantly higher FKGL scores than ChatGPT, indicating that its responses were linguistically more complex and required a higher educational level for comprehension. Both AI chatbots produced responses that could help inform patients about head and neck imaging techniques. The findings suggest that AI-based chatbots can support patient education and facilitate access to information, but should only serve as complementary tools in clinical decision-making processes under the guidance of healthcare professionals. It should be kept in mind that information from AI may contain errors or omissions and should be verified by healthcare professionals.

ETHICAL DECLARATIONS

Ethics Committee Approval

Ethical approval was not required for this study because no patient data or personal information was used and the study aimed solely to analyze freely accessible information sources.

Informed Consent

This study did not involve human participants, clinical samples, or personal health data; it analyzed only freely available information sources. Therefore, informed consent is not required.

Peer Review Process

This manuscript was subject to external peer review.

Conflict of Interest

The authors declare no conflicts of interest related to this study.

Financial Disclosure

The authors received no financial support for the conduct or publication of this research.

Author Contributions

Conceptualization: MH, AC; Methodology: MH, AC; Software: MH; Validation: MH, AC; Formal Analysis: MH; Investigation: MH; Resources: MH; Data Curation: MH, AC; Writing-Original Draft Preparation: MH; Writing-Review and Editing: MH, AC; Visualization: MH; Supervision: AC; Project Administration: MH.

REFERENCES

1. Guler R, Yalcin E. Performance of AI chatbots in preliminary diagnosis of maxillofacial pathologies. *Med Sci Monit*. 2025;31:e949076. doi:10.12659/MSM.949076
2. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med*. 2023;29(8):1930-1940. doi:10.1038/s41591-023-02448-8
3. Çetin P, Onan A. Büyük dil modellerinin tıp eğitimde soru yanıtlama sistemlerinde kullanımı: potansiyel, zorluklar ve gelecek yönelimleri. *JISMCE*. 2025;1(1):82-108.
4. Rokhshad R, Zhang P, Mohammad-Rahimi H, Pitchika V, Entezari N, Schwendicke F. Accuracy and consistency of chatbots versus clinicians for answering pediatric dentistry questions: a pilot study. *J Dent*. 2024;144:104938. doi:10.1016/j.jdent.2024.104938
5. Tan S, Xin X, Wu D. ChatGPT in medicine: prospects and challenges: a review article. *Int J Surg*. 2024;110(6):3701-3706. doi:10.1097/JS9.0000000000001312
6. Jeong H, Han SS, Yu Y, Kim S, Jeon KJ. How well do large language model-based chatbots perform in oral and maxillofacial radiology? *Dentomaxillofac Radiol*. 2024;53(6):390-395. doi:10.1093/dmfr/twae021

7. Zhang Y, Lu X, Luo Y, Zhu Y, Ling W. Performance of Artificial Intelligence chatbots on ultrasound examinations: cross-sectional comparative analysis. *JMIR Med Inform*. 2025;13:e63924. doi:10.2196/63924
8. Patil NS, Huang RS, Caterine S, et al. Artificial Intelligence chatbots' understanding of the risks and benefits of computed tomography and magnetic resonance imaging scenarios. *Can Assoc Radiol J*. 2024;75(3): 518-524. doi:10.1177/08465371231220561
9. Mohammad-Rahimi H, Khoury ZH, Alamdari MI, et al. Performance of AI chatbots on controversial topics in oral medicine, pathology, and radiology. *Oral Surg Oral Med Oral Pathol Oral Radiol*. 2024;137(5):508-514. doi:10.1016/j.oooo.2024.01.015
10. Reda R, Zanza A, Mazzoni A, Cicconetti A, Testarelli L, Di Nardo D. An update of the possible applications of magnetic resonance imaging (MRI) in dentistry: a literature review. *J Imaging*. 2021;7(5):75. doi:10.3390/jimaging7050075
11. Al-Haj Husain A, Zollinger M, Stadlinger B, et al. Magnetic resonance imaging in dental implant surgery: a systematic review. *Int J Implant Dent*. 2024;10(1):14. doi:10.1186/s40729-024-00532-3