ORIGINAL ARTICLE
DOI: 10.5152/EssentDent.2025.24031

ISTANBUL UNIVERSITY
CERRAHPAŞA

Evaluation of the Information Quality of Chatbot Technologies About Root Canal Treatment

Cem Sarkan[1,2], Faruk Haznedaroglu[1]

[1]Department of Endodontics, İstanbul University Faculty of Dentistry, İstanbul, Türkiye
[2]Institute of Graduate Studies in Health Sciences, İstanbul University, İstanbul, Türkiye

Cite this article as: Sarkan C, Haznedaroglu F. Evaluation of the information quality of chatbot technologies about root canal treatment. *Essent Dent.* 2025, 4, 0031, doi: 10.5152/EssentDent.2025.24031.

Abstract

Background: This study aims to evaluate the quality of information provided by artificial intelligence (AI)-based chatbot technologies regarding root canal treatment. ChatGPT-4 (OpenAI Inc.), Google Gemini (Google LLC), and Microsoft Copilot (Microsoft Corporation) were assessed for the accuracy, reliability, and scope of their responses.

Methods: Twenty frequently asked questions about root canal treatment were compiled from online sources and refined by endodontic experts. These questions were posed to the 3 chatbots, and responses were collected twice to evaluate reliability. Answers were scored using a 5-point Likert scale based on a modified Global Quality Score for accuracy and completeness. Statistical analyses, including Pearson correlation coefficients and Fisher's exact tests, were conducted to evaluate the validity and reliability of the responses.

Results: ChatGPT provided the longest responses (mean word count: 250) but did not include citations. Copilot offered concise responses (mean word count: 120) with consistent citations, while Gemini provided moderate-length answers (mean word count: 230) with limited references. Regarding dentist recommendations, ChatGPT suggested consulting a dentist in 16 of 40 responses, Copilot in 12, and Gemini in 38. At the high validity threshold, Gemini achieved a 100% validity rate, outperforming ChatGPT and Copilot. Reliability analysis showed high correlation coefficients for all chatbots: Gemini (1.0), Copilot (0.95), and ChatGPT (0.93).

Conclusion: Artificial intelligence-based chatbots show strong potential for providing patient education on root canal treatment. Gemini demonstrated superior reliability and validity, highlighting the importance of robust training datasets. However, inaccuracies and limitations in chatbot responses underline the need for continuous evaluation and improvement of these technologies to ensure they are reliable and clinically useful tools in healthcare.

Keywords: Artificial intelligence, chatbot, chatgpt, copilot, endodontics, gemini, root canal treatmentIntroduction

In dentistry, root canal treatment is often perceived by patients as a complex and anxiety-inducing procedure. Access to accurate and easily understandable information regarding treatment is critical for the successful management of the process. The integration of digital technologies into healthcare services has made it easier for patients to access such information. In particular, artificial intelligence (AI)-based chatbot technologies have emerged as effective tools for providing patients with information on topics they may find concerning, such as dental treatments.[1,2]

What is already known on this topic?

- *Artificial intelligence-based chatbots have been increasingly used as tools for delivering healthcare information. Previous studies have evaluated earlier versions of these technologies and found that while they offer potential for patient education, their responses often lacked consistency, clinical accuracy, and evidence-based detail, particularly in dental contexts such as endodontics.*

What this study adds on this topic?

- *This study evaluates the most recent versions of three major AI-based chatbots—ChatGPT-4, Google Gemini, and Microsoft Copilot—specifically in the context of endodontic patient education. It demonstrates that newer-generation models, particularly Gemini, provide significantly more reliable and valid information compared to earlier versions reported in the literature. The findings underscore the progress in AI chatbot development and highlight the need for ongoing assessment as these tools continue to evolve and impact clinical communication.*

Corresponding author: Cem Sarkan
e-mail: cemsarkan@hotmail.com

Content of this journal is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Received: November 26, 2024
Revision Requested: January 8, 2025
Last Revision Received: February 5, 2025
Accepted: February 21, 2025
Publication Date: October 8, 2025

In recent years, the use of digital solutions and AI-based healthcare technologies has gradually increased. Among these technologies, chatbots stand out in terms of providing information and consulting, particularly in the field of health.[1] Chatbots supported by AI, particularly machine learning and deep learning algorithms, contribute to decision-making processes in healthcare by providing natural language responses to users.[1,3] GPT-4 (OpenAI Inc.), Google Gemini (Google LLC), and Bing (Copilot) (Microsoft Corporation) are platforms at the forefront of this development. One of these models, GPT-4, is an advanced language model developed by OpenAI. Unlike previous versions, GPT-4 is a model that has been trained with a larger data set and has the capacity to understand and generate language with higher accuracy. GPT-4 covers a wide range of languages with human-like language processing capabilities and can deliver highly accurate results even in complex language structures.[4] This model has become an important tool for the writing and editing of scientific papers. It offers effective results in many areas such as text summarization, language translation, text production, and language analysis. GPT-4 can process and synthesize information obtained from various scientific sources by scanning a database. This saves researchers a significant amount of time in analyzing and interpreting complex data sets.[5] Gemini was developed by Google DeepMind and combines large language models (LLMs) and deep learning techniques for use in image processing, data analytics, and many other tasks besides natural language processing. In this regard, Gemini has the ability to process different types of data, both text and images, owing to its multi-modality support. Gemini's language model is positioned as a versatile AI system that can perform more complex and detailed analyses using large-scale training sets such as GPT.[6] Bing is a product of Microsoft's vast AI research and development resources. Although less is known about its specific architecture, Microsoft's extensive web-scale data benefits from significant advances in AI techniques with GPT-3.5 and machine learning. On the other hand, the Copilot used in the study uses the newest version of OpenAI GPT-4 technology as part of its basic structure.[7]

In the field of dentistry, chatbots have begun to be integrated with functions such as informing patients, providing guidance before and after treatment, and responding to routine questions.[2,8] The use of chatbot technologies in dentistry has important potential for patient information processes related to endodontic treatments. However, there are limited studies in the literature on the quality and accuracy of the information provided by chatbots.[3,9] Some studies emphasize that the effectiveness of chatbots, especially in providing medical information, may vary and that the reliability of these technologies should be evaluated comprehensively.[1] For this reason, the accuracy, timeliness, and clinical reliability of the information provided by chatbots are important.[8,10]

In this study, the accuracy, reliability, and suitability of the information provided by ChatGPT, Google Gemini, and Microsoft Copilot AI-based chatbots were examined. This evaluation aimed to reveal the potential uses and limits of chatbot technologies in the field of endodontics.

MATERIAL AND METHODS

Ethical Considerations and Informed Consent

This study did not require ethical committee approval as it did not involve human participants, animal subjects, or the collection of sensitive data.

Informed Consent: For this study, informed consent is not required since it does not involve any studies with human participants or animal subjects.

Twenty frequently asked questions (FAQs) about endodontics and root canal treatments were selected and represented a wide range of patient questions. These questions were collected from 2 sources in Türkiye:

A) First, the Google search engine was used to collect the most FAQs about endodontic treatment. The questions on the first 50 pages were analyzed individually, and 30 questions were selected.

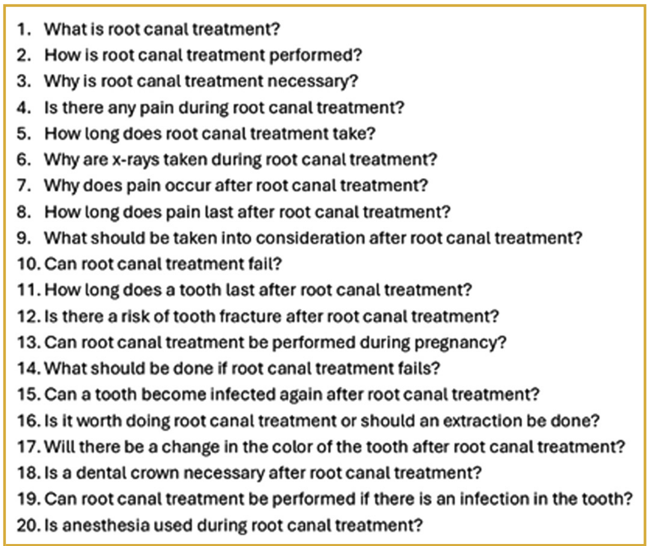
B) The 30 most commonly asked questions in the field of endodontics, provided on demand by GPT-4, were examined. The top 20 most commonly asked questions in the common set of 2 groups were selected by an experienced instructor and a doctoral student in endodontics, one of whom works full-time in the field of endodontics.

The list of these original 20 questions is presented in English translated versions in Figure 1. The questions cover a variety of areas such as terminology, diagnosis, treatment procedures/technical details, aftercare, prognosis, potential risks, possible side effects, and alternative treatment options.

The validity rates of chatbot responses at both low and high thresholds are summarized in Figure 2.

Each question was posed to 3 different AI chatbots, with each being repeated twice to evaluate reliability. The questions were repeated twice under similar conditions (posed on the same day without altering context or phrasing). This step was crucial to analyze the reliability of chatbot responses through consistency testing. To replicate real-world interactions, the process was streamlined through the primary application programming interface (API) of each chat platform. The following methods were employed:

GPT-4.0: The API was accessed at https://chat.openai.com/. A paid version of the API (i.e., GPT-4.0) was used. All questions were submitted on the same day to ensure consistency (October 05, 2024). A new chat was created for each question.



1. What is root canal treatment?
2. How is root canal treatment performed?
3. Why is root canal treatment necessary?
4. Is there any pain during root canal treatment?
5. How long does root canal treatment take?
6. Why are x-rays taken during root canal treatment?
7. Why does pain occur after root canal treatment?
8. How long does pain last after root canal treatment?
9. What should be taken into consideration after root canal treatment?
10. Can root canal treatment fail?
11. How long does a tooth last after root canal treatment?
12. Is there a risk of tooth fracture after root canal treatment?
13. Can root canal treatment be performed during pregnancy?
14. What should be done if root canal treatment fails?
15. Can a tooth become infected again after root canal treatment?
16. Is it worth doing root canal treatment or should an extraction be done?
17. Will there be a change in the color of the tooth after root canal treatment?
18. Is a dental crown necessary after root canal treatment?
19. Can root canal treatment be performed if there is an infection in the tooth?
20. Is anesthesia used during root canal treatment?

Figure 1. The list of these original 20 questions is presented in English translated versions.

Google Gemini: The API was accessed via https://gemini.google.com/app. All questions were submitted on the same day (October 05, 2024). The chat was reset before each new question was asked.

Copilot: The API was accessed over https://copilot.microsoft.com using Microsoft Edge. All questions were submitted on the same day (October 05, 2024). A separate chat session was created for each question.

Scoring

All responses were independently assessed using a 5-point Likert Scale by an experienced instructor and a PhD student in endodontics, one of whom works full-time in the field of endodontics. A modified version of the Global Quality Score developed by Bernard et al[11] (2007) was used to score responses based on "context" and "content":

Score 5 (Completely Agree): The response is accurate, with all information provided being correct and comprehensive.

Score 4 (Agree): The response is generally accurate, with most content being correct; however, it may include incomplete or slightly incorrect information.

Score 3 (Undecided): The response is partially accurate but contains significant inaccuracies, incomplete details, or irrelevant content.

Score 2 (Disagree): The response is largely incorrect but includes some correct elements.

Score 1 (Strongly Disagree): The response is entirely incorrect or irrelevant, with no accurate or meaningful content.

Using these criteria, chatbot responses were thoroughly evaluated, taking into account both the accuracy of the responses (context) and the completeness and correctness of the information provided (content). Responses were evaluated according to European Endodontic Society and American Association of Endodontists guidelines with literature-supported information. The questions were repeated twice to analyze the consistency of each chatbot and assess the reliability of the responses. After scoring was completed separately, the two reviewers shared their scoreboards (each totaling 120 points) and reviewed and discussed the different responses. Inter-rater reliability was assessed using Cohen's Kappa to evaluate the agreement between the 2 independent raters. The Kappa value was 0.83 (95% CI: 0.75-0.91), indicating almost perfect agreement beyond chance. Disputes were resolved through evidence-based discussions on context and content. Finally, a single scoring table was prepared for all 120 responses for the statistical analyses.

Statistical Analysis

For the low-threshold test, the score threshold was set to 4. If both answers to a question received a score of ≥4, the

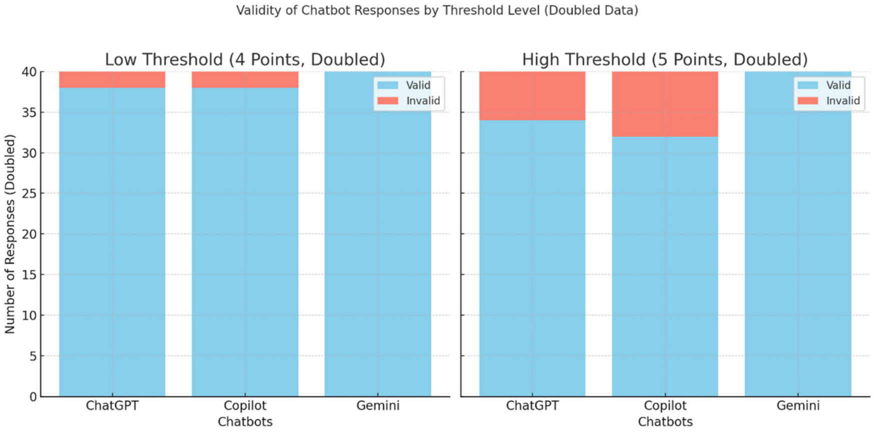


Figure 2. Validity of chatbot responses by threshold level.

chatbot's answer was considered valid. If any response scored less than 4, the chatbot's response was considered invalid. For the high-threshold test, the score threshold was set to 5. In this case, the chatbot's response was considered valid only if both received a score of 5. If any response scored less than 5, the chatbot's response was considered invalid. Fisher's exact test was used to compare the validity of the responses between the chatbots. The significance level was determined as <0.05. The validity rates of chatbot responses at both low and high thresholds are summarized in Figure 2. All analyses were performed using the R programming language version 4.1.0 (R Foundation for Statistical Computing, Vienna, Austria).

Analysis of Reliability

Reliability (the degree to which the chatbot produces similar responses when repeated with the exact same question under similar conditions) was defined by analyzing the 5-point Likert scale scores assigned to the responses given when repeated twice. To assess the consistency of the responses, the Pearson correlation coefficient was calculated for the 2 groups of responses that each chatbot gave to the 20 questions. The Pearson correlation shows the level of consistency on a standardized scale between −1 and 1: 1 denotes perfect consistency, 0 denotes no correlation, and −1 denotes perfect inverse correlation. A high correlation coefficient indicates that the chatbot consistently provides the same structure and that the scale can be considered reliable. A low correlation coefficient indicates that the chatbot does not consistently provide the same structure, and that the scale is less reliable. In medical and health research, a correlation coefficient of ≥0.70 is generally considered acceptable reliability.

All analyses were performed using the R programming language version 4.1.0.

RESULTS

Descriptive Analyses

Three different chatbots responded to all 20 repeated questions, for a total of 120 responses. A list of the responses is presented in the supplementary materials (Appendix S1). Copilot provided shorter responses (average word count: 120) and cited a bibliography in all of its responses. ChatGPT-4, on the other hand, offered longer responses (average word count: 250) but did not provide a bibliography of any of its responses. Gemini was moderate in both length and detail in its responses (average word count: 230) and provided a bibliography in only 3 responses. In terms of dentist recommendation, ChatGPT-4 directed users to consult a dentist in 16 out of 40 questions. Copilot suggested consulting with a dentist on 12 out of 40 questions. Gemini, on the other hand, encouraged users to consult a dentist in 38 responses. Furthermore, the 3 chatbots offered incorrect or irrelevant statements.

Reliability

In the study, ChatGPT: 0.93, Copilot: 0.95, and Gemini: 1 correlation was found. All chatbots showed high correlation values, and their answers showed an acceptable level of reliability.

Statistical Analysis

ChatGPT: Low Threshold (4 points): 19 valid and 1 invalid response. High Threshold (5 points): 17 valid and 3 invalid responses.

Copilot: Low Threshold (4 points): 19 valid and 1 invalid response. High Threshold (5 points): 16 valid and 4 invalid responses.

Gemini: Low Threshold (4 points): 20 valid and 0 invalid responses. High Threshold (5 points): 20 valid and 0 invalid responses.

According to these results, all chatbots have high validity rates at the low threshold level, but at the high threshold level, all of Gemini's responses are considered valid, while some of ChatGPT's and Copilot's responses are considered invalid. Low Threshold (4 points): No significant difference was found between ChatGPT, Copilot, and Gemini in terms of validity (*P*-value = .596). High Threshold (5 points): There was no significant difference between chatbots at the high threshold either, but the *P*-value was lower (*P*-value = .122).

DISCUSSION

Artificial intelligence chatbots have become a powerful and readily accessible source of information, with the potential to transform how individuals access and process information, particularly in healthcare.[12] In this study, the quality of information in patient information processes related to the root canal treatment provided by AI-based chatbots was evaluated in terms of reliability and accuracy. The results reveal that chatbots have the potential to provide information, especially in a technical field such as dentistry, but they can differ in terms of information accuracy and scope. When evaluating the validity of chatbot responses, the source and language of the questions play a crucial role. Selected questions in this article are intended to reflect patients' questions and concerns about endodontics. Twenty questions were selected by 2 endodontists who work full-time and interact with patients on a daily basis.

Performance of Chatbots

Previous studies showed that Google Bard provided non-evidence-based recommendations, such as suggesting general anesthesia for pregnant patients with allergies to local anesthetics, which could lead to serious health consequences.[13] The misinformation can have far-reaching effects, as seen during the COVID-19 pandemic when public hesitancy to visit dental clinics was fueled by misinformation.[14-16] In this study, the results the newest version of Google Bard show that Gemini performs more consistently and reliably than other chatbots, especially when all its responses are considered valid in both low- and high-threshold tests. This is due to Gemini's ability to process both text and other types of data, and the possibility that it may have benefited from a larger-scale training set. This suggests that Gemini's versatile structure and the large data sets used in the training process contribute to the accuracy and scope of the information. In similar studies, Mohammad-Rahimi et al[17] showed 85% validity in the low threshold scale and 15% validity in the high threshold achievement scale of Bard, an older version of Gemini. In the study, Gemini, which has shown more successful and reliable performance, suggests that AI-supported chatbots will provide increasingly successful results as technology advances in this field day by day. ChatGPT and Copilot, however, have not been able to provide results consistent with Gemini, as some of their responses are considered invalid at the high threshold level. However, both models performed well at the low threshold level (95%), suggesting that the overall information accuracy is acceptable. The fact that ChatGPT provides long and detailed answers is advantageous in terms of providing comprehensive information to users. However, the lack of attribution can be considered a disadvantage. On the other hand, Copilot's shorter, source-based responses may be a more reliable option for users to fact-check. Lahat et al,[18] in their study evaluating chatbots in the healthcare field, reported that GPT-4 outperformed GPT-3.5 in all evaluation dimensions. Similarly, in the study, next generation chatbots demonstrated higher performance compared to the findings of Mohammad-Rahimi et al.[17] This could be attributed to factors such as the questions being asked in different languages and being sourced from different datasets. However, it can be anticipated that advancements in AI will continue to drive further improvements in this field.

Variations in chatbot responses can be attributed to differing design philosophies of AI companies, the specific algorithms implemented, the training datasets utilized, and the objectives each AI system is designed to achieve.[17]

Reliability Analysis

The inherent probabilistic nature of LLMs can lead to variability in their responses, as highlighted by Suárez et al,[20] who evaluated ChatGPT's performance in answering endodontic questions and observed a satisfactory consistency rate of 85.4%.[19]

The findings align with those of other studies,[19,21,22] which have highlighted the promising potential of this LLM in medical education and clinical decision-making.

In this study, the fact that chatbots show a high correlation coefficient when answering the same questions twice reveals that their consistency is high. In particular, Gemini has an excellent correlation coefficient (1.0), which supports that this system is the most successful chatbot in terms of reliability. ChatGPT (0.93) and Copilot (0.95) also exhibited high levels of reliability; however, they have not performed as consistently as Gemini.

The Role of Chatbots in Healthcare

This study highlights the significance of ensuring accurate and reliable chatbot responses in the dental field to enhance patient satisfaction and reduce healthcare costs.[23] The role of chatbots in healthcare is increasing due to their potential to provide accurate and easily accessible information. However, for the full integration of these technologies into clinical practice, it is necessary to constantly evaluate the accuracy and reliability of the information they provide. In this study, it has been observed that chatbots can vary in information accuracy. In particular, providing incorrect or incomplete information may lead to misleading patients and adversely affect their treatment processes.[20] For this reason, it is important that chatbots are designed more carefully when presenting content and supported by constantly updated data sets. One thing to note is that despite their generally high validity scores, these chatbots have made critical errors in some of their responses. Some of these responses may have the potential to mislead patients on certain topics. For example, Copilot could not give a correct answer to the question "Can root canal treatment fail?".

Limitations and Future Studies

Among the limitations of this study is that chatbots were evaluated only on a specific date and with a specific data set (questions were asked and answered in Turkish and according to the Türkiye database). Similar studies conducted in different time periods or in different contexts can provide more comprehensive and generalizable results. Additionally, a limitation of the study is that the validity of the responses was evaluated by only 1 certified endodontist and 1 doctoral student. It may be desirable to use a larger pool of endodontists to examine the validity of the responses. Furthermore, the limited range of questions used in this study prevented the evaluation of the performance of chatbots in other fields of dentistry. Future studies are proposed to evaluate the ability of chatbots to provide information across a wide range of healthcare.

The results of this study show that AI-based chatbots have strong potential to provide information about root canal treatment but should be carefully evaluated for information accuracy and scope. In particular, Gemini's superior performance demonstrates that chatbots trained with versatile and up-to-date data sets can be used more effectively in healthcare. However, more research and development are required for these technologies to become truly reliable and comprehensive sources of information.

Data Availability Statement: The data that support the findings of this study are available on request from the corresponding author.

Ethics Committee Approval: The methodology of the present study is designed without any human or human-related subjects. Therefore, no ethical approval was necessarily obtained.

Informed Consent: The methodology of the present study is designed without any human or human-related subjects. Therefore, no consent was necessarily obtained.

Peer-review: Externally peer-reviewed.

Author Contributions: Concept – C.S.; Design – C.S.; Supervision – F.H.; Resources – C.SAND F.H.; Materials – C.S.; Data Collection and/or Processing – C.S.; Analysis and/or Interpretation – C.S., F.H.; Literature Search – C.S.; Writing Manuscript – C.S.; Critical Review – F.H.

Declaration of Interests: The authors have no conflicts of interest to declare.

Funding: The authors declared that this study has received no financial support.

REFERENCES

1. Schwendicke F, Samek W, Krois J. Artificial intelligence in dentistry: chances and challenges. *J Dent Res*. 2020;99(7):769-774. [CrossRef]
2. Chen YW, Stanley K, Att W. Artificial intelligence in dentistry: current applications and future perspectives. *Quintessence Int*. 2020;51(3):248-257. [CrossRef]
3. Aminoshariae A, Kulild J, Nagendrababu V. Artificial intelligence in endodontics: current applications and future directions. *J Endod*. 2021;47(9):1352-1357. [CrossRef]
4. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H, eds. Curran Associates, Inc; 2020;1877-1901. Available at: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.
5. Floridi L, Chiriatti M. GPT-3: its nature, scope, limits, and consequences. *Minds Mach (Dordr)*. 2020;30(4):681-694. [CrossRef]
6. Bommasani R, Hudson DA, Adeli E, et al. On the opportunities and risks of foundation models. *CORR* [Internet]. 2021;[abs]/2108.07258. Available at: https://arxiv.org/abs/2108.07258.
7. Rakauskas TR, Da Costa A, Moriconi C, Gill G, Kwong JW, Lee N. Evaluation of chat generative pre-trained transformer and Microsoft copilot performance on the American society of surgery of the hand self-assessment examinations. *J Hand Surg Glob Online*. 2025;7(1):23-28. [CrossRef]
8. Heo MS, Kim JE, Hwang JJ, et al. Artificial intelligence in oral and maxillofacial radiology: what is currently possible? *Dento Maxillo Facial Rad*. 2021;50(3):20200375. [CrossRef]
9. Carrillo-Perez F, Pecho OE, Morales JC, et al. Applications of artificial intelligence in dentistry: a comprehensive review. *J Esthet Restor Dent*. 2022;34(1):259-280. [CrossRef]
10. Yang S, Lee J, Sezgin E, Bridge J, Lin S. Clinical advice by voice assistants on postpartum depression: cross-sectional investigation using Apple Siri, amazon Alexa, google assistant, and Microsoft Cortana. *JMIR MHealth UHealth*. 2021;9(1):e24045. [CrossRef]
11. 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(9):2070-2077. [CrossRef]
12. Walker HL, Ghani S, Kuemmerli C, et al. Reliability of medical information provided by ChatGPT: assessment against clinical guidelines and patient information quality instrument. *J Med Internet Res*. 2023;25:e47479. [CrossRef]
13. Rampa S, Veeratrishul A, Raimondo M, Connolly C, Allareddy V, Nalliah RP. Hospital-based emergency department visits with periapical abscess: updated estimates from 7 years. *J Endod*. 2019;45(3):250-256. [CrossRef]
14. Nosrat A, Dianat O, Verma P, Yu P, Wu D, Fouad AF. Endodontics Specialists' practice during the initial outbreak of coronavirus disease 2019. *J Endod*. 2022;48(1):102-108. [CrossRef]
15. Nosrat A, Yu P, Dianat O, et al. Endodontic Specialists' practice during the coronavirus disease 2019 pandemic: 1 year after the initial outbreak. *J Endod*. 2022;48(6):699-706. [CrossRef]
16. Abbasi J. Widespread misinformation about infertility continues to create COVID-19 vaccine hesitancy. *J Am Med Assoc*. 2022;327(11):1013-1015. [CrossRef]
17. 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. [CrossRef]
18. Lahat A, Sharif K, Zoabi N, et al. Assessing generative pretrained transformers (GPT) in clinical decision-making: comparative analysis of GPT-3.5 and GPT-4. *J Med Internet Res*. 2024;26:e54571. [CrossRef]
19. Antaki F, Touma S, Milad D, El-Khoury J, Duval R. Evaluating the performance of ChatGPT in ophthalmology: an analysis of its successes and shortcomings. *Ophthalmol Sci* [Internet]. 2023;3(4):100324. [CrossRef]
20. Suárez A, Díaz-Flores García V, Algar J, Gómez Sánchez M, Llorente de Pedro M, Freire Y. Unveiling the ChatGPT phenomenon: evaluating the consistency and accuracy of endodontic question answers. *Int Endod J*. 2024;57(1):108-113. [CrossRef]
21. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States medical licensing examination (USMLE)? The implications of large language models for medical education and knowledge assessment. *JMIR Med Educ*. 2023;9:e45312. [CrossRef]
22. Huh S. Are ChatGPT's knowledge and interpretation ability comparable to those of medical students in Korea for taking a parasitology examination?: a descriptive study. *J Educ Eval Health Prof*. 2023;20:1. [CrossRef]
23. Xu L, Sanders L, Li K, Chow JCL. Chatbot for health care and oncology applications using artificial intelligence and machine learning: systematic review. *JMIR Cancer*. 2021;7(4):e27850. [CrossRef]