Journal of Oral Medicine and Dental Research

Potential Use of ChatGPT and GEMINI for Patient Information in Endodontics

Büşra Zengin[1]*, Elifnur Güzelce Sultanoğlu[2] and Erkan Sultanoğlu[3]

[1]Endodontist, Usküdar Oral and Dental Health Center, Istanbul, Turkey

[2]Assistant Professor, Department of Proshodontics, University of Health Sciences, Istanbul, Turkey

[3]Orthodontist, Department of Orthodontics, University of Biruni, Istanbul, Turkey

*Corresponding author: Büşra Zengin, Endodontist, Usküdar Oral and Dental Health Center, Istanbul, Turkey

Abstract

The aim of this study is to evaluate the accuracy, completeness and readability of the answers given by the artificial intelligence robots ChatGPT-4 and GEMINI to frequently asked questions about root canal treatment. 20 frequently asked questions by patients about root canal treatment were created. The questions were answered by artificial intelligence robots. The answers were scored by experts for accuracy and completeness using a Likert scale. The FRES score was calculated to assess readability. Statistically, no significant difference was found between ChatGPT and GEMINI in terms of readability. When evaluating the overall average of accuracy and completeness, no meaningful difference was observed ($p < 0.05$). Artificial intelligence robots are continuously advancing over time. However, consideration should also be given to the possibility of breaches of patient privacy, risks of misuse and misinformation.

Keywords

Artificial intelligence; Large language model; ChatGPT; GEMINI; Root canal treatment.

Introduction

Artificial Intelligence (AI) is technology designed to enable machines to think, learn and make decisions like humans [1]. Great strides are being made in artificial intelligence, particularly large language models (LLMs) such as ChatGPT (San Francisco, California, United States) and Google's Gemini AI (Google Ireland Limited, Dublin, Ireland). Chat Generative Pre-Trained Transformer (ChatGPT) is an artificial intelligence-based chatbot developed by Open AI. It will be released in November 2022 [2]. This generative AI software is based on large language models and provides human-like answers to questions posed to it. AI technology essentially consists of a neural network architecture that resembles the human brain and mimics how human beings think [3]. The neural architecture model consists of highly interconnected nerve cells that basically work as a data processing system to solve a specific task [4].

Patients want to be able to consult a doctor, to decide whether they need to see a doctor, and to get an explanation and a diagnosis for the complaints they have [5]. As technology advances, the use of artificial intelligence (AI) chatbots among healthcare professionals and patients is growing every day as an accurate source of access to medical and dental information [6]. Artificial intelligence chatbots allow patients to apply whenever they want and get instant answers, allowing people to ask whatever they want without hesitating.

ChatGPT can provide many services to healthcare professionals in dentistry and medicine, including better diagnosis, decision support, digital data capture, image analysis, disease prevention, disease treatment, reducing treatment delays, and enabling discovery and research [7,8].

It is important to objectively evaluate health-related content generated by artificial intelligence. In addition to these features, the ability of the system to give incorrect answers, produce irrelevant content, and present false information and disinformation as if it were real also raises serious concerns in critical areas such as health [9].

The reliability and quality of the information source in chatbots is crucial, as it can affect patient cooperation and compliance with treatment, doctor-patient communication and trust. In order to achieve the best results, accuracy in diagnosis and clinical decision making is of paramount importance. Advances in science and technology have led to the development of various diagnostic tools and treatment modalities that have opened up new horizons in the diagnosis, clinical decision making and planning of the best treatment for endodontic disease [10].

The aim of this study is to evaluate the accuracy, completeness and readability of the answers given by the artificial intelligence robots ChatGPT-4 and GEMINI to frequently asked questions about root canal treatment.

Materials and Methods

Ethical approval was not required for this research as no human participants were involved in the study. 20 questions were created using the Quora platform and Google search tool that patients often ask about root canal treatment (11,12) (Table 1). Questions included the need for root canal treatment, the process and content of treatment, the relationship between root canal treatment and orthodontic treatment, and the planning of prosthetic superstructures after root canal treatment. ChatGPT-4 and Gemini AI chatbots were used to answer these questions (13).

Questions were answered on the same day. All responses were evaluated by an endodontist and an orthodontist, taking into account current literature and clinical practice.

A Likert scale was used to evaluate the responses, questioning accuracy and completeness. A five-point Likert scale was used for accuracy. According to this scale, 1: I strongly disagree with the accuracy of the answers. 2: I disagree with the accuracy of the answers, 3: I am undecided about the accuracy of the answers, 4: I agree with the accuracy of the answers. 5: I strongly agree with the accuracy of the answers. A 3-point Likert scale was used for the integrity of the answers. According to this scale, 1: Incompleteness of the answers, 2: Completeness of the answers is sufficient and 3: Completeness of the answers is comprehensive.

The readability of the responses was then assessed utilizing the Flesch–Kincaid readability test to generate a reading ease score. The online resource https://goodcalculators.com/flesch-kincaid-calculator/ was utilized to calculate Flesch Reading Ease Score (FRES) for each question.

The Flesch Reading Ease Score is a measure of how easy a text is to read. It is calculated using the average sentence length and the average number of syllables per word. The score typically ranges from 0 to 100, with higher scores indicating easier readability. In this scoring system, a score between 0 and 100 is obtained according to a calculation and a reading level is determined between easy and difficult.

	QUESTIONS
1	How painful is a root canal?
2	What can go wrong during root canal treatment?
3	What are the common reasons for needing a root canal treatment?
4	What is a successful root canal treatment?
5	What are some signs that indicate you need a root canal?
6	How long does a root canal take?
7	How long will the treated tooth last?
8	What happens during a root canal?
9	What care is needed after a root canal?
10	Can I have orthodontic treatment if I have a root canal in my tooth?
11	Can you use Invisalign with root canals?
12	How long after Root Canal can i get Braces?

13	How long after a root canal can I get Invisalign?
14	Can you get a root canal with braces?
15	Can you get a root canal during Invisalign treatment?
16	Will the tooth be weaker after a root canal?
17	What is the best restoration after a root canal?
18	Is a root canal treatment better than a dental implant?
19	Can a root canal be performed on a tooth with a crown or bridge?
20	Why are dental crowns recommended after a root canal?

Table 1: Questions

Statistical Analysis

Cohen's kappa statistic has been used by researchers to determine how well the subjective ratings of two observers on the same categorical variable are in agreement with each other.

The intraclass correlation coefficient (ICC) was used for the assessment of consistency and reliability between observers. The ICC is used to determine how consistent measurements are, especially when more than one measurement is made within the same group.

T test was used to examine the ChatGPT and GEMINI results of the mean values of accuracy, completeness and readability. Independent sample t test was used to test whether there was a statistically significant difference between two independent groups by looking at the means.

Results

For readablitiy, the Chat GPT average is 52.58, the Gemini average is 55.14, and there is no significant difference between the scores (p>0.05) (Table 2).

	N	Mean	Std. Deviation	p
ChatGPT	20	52,58	8,16	0,348
Gemini	20	55,14	14,16	

Table 2: Readability score.

Kappa reliability coefficient and ICC intraclass consistency coefficient were obtained to measure the agreement. These coefficients were obtained as 0.816 and 0.886 for CHAT GPT, respectively; and 0.745 and 0.819 for Gemini.

In Expert 1 assessments, accuracy and completeness levels do not differ significantly according to the type of artificial intelligence (p>0.05). In Expert 2 assessments, accuracy levels do not differ significantly according to the type of artificial intelligence (p>0.05), while completeness levels differ significantly according to the type of artificial intelligence (p<0.05). In Expert 2 completeness assessment, ChatGPT average (2.80) is significantly higher than gemini assessment (2.35) (Table 3).

			N	Mean	Std. Deviation	p
Expert 1	Accuracy	Chat GPT	20	4,30	0,92	0,860
		Gemini	20	4,25	0,85	
	Completeness	Chat GPT	20	2,70	0,47	0,059
		Gemini	20	2,40	0,50	
Expert 2	Accuracy	Chat GPT	20	4,35	0,81	0,841
		Gemini	20	4,40	0,75	
	Completeness	Chat GPT	20	2,80	0,41	0,003*
		Gemini	20	2,35	0,49	

Table 3: Accuracy and completeness mean value.

Discussion

In this study, the accuracy, completeness and readability of the answers given by ChatGPT-4 and GEMINI to questions regarding root canal treatment were evaluated. The investigation revealed no substantial disparities between ChatGPT-4 and GEMINI with respect to readability. Similarly, no statistically significant disparities were observed between the two AI assessments in terms of accuracy. There was no significant difference between the two AI assessments for Completeness. In the present study, the ChatGPT-4 English version was utilised. Research has demonstrated that ChatGPT-4 is more efficacious than ChatGPT-3.5 in the medical field (18). The ChatGPT 4 model was evaluated as exhibiting superior creativity and reliability in comparison to the ChatGPT-3.5 model. Additionally, the AI was reported to demonstrate proficiency in addressing medical scenarios and to perform adequately under the stipulated conditions (14).

The utilisation of language models, such as ChatGPT and GEMINI, holds considerable promise for enhancing clinical practice and research in the field of dentistry. The success of these chatbots in medical language applications is particularly noteworthy. For instance, recent studies have underscored ChatGPT's capacity to attain performance levels that meet or approximate the passing criteria on all three stages of the United States Medical Licensing Exams (USMLE). This finding underscores the potential of

these chatbots not only as a tool for medical education, but also as a resource to support healthcare professionals and patients in navigating medical information [15].

ChatGPT can provide many services to healthcare professionals in dentistry and medicine, including better diagnosis, decision support, digital data capture, image analysis, disease prevention, disease treatment, reducing treatment delays, and enabling discovery and research [7,8]. It can be a valuable medical tool as it can be useful for diagnosis, treatment planning and recording of medical notes or patient information [16]. Despite the advances and successful results of these models, it is important to evaluate their performance, limitations and potential errors before applying them to patients in a clinical setting [17].

The utilisation of artificial intelligence (AI) in the field of endodontics has been demonstrated to facilitate accurate diagnoses and prognostication. Its integration into treatment planning has been demonstrated to enhance outcomes, leading to an increase in treatment success. AI has been incorporated into a variety of clinical applications in endodontics, including the detection of root fractures and periapical pathologies, the determination of working lengths, the monitoring of the apical foramen, the analysis of root morphology, and the prediction of disease [18].

In this study, Rahimi et al. evaluated and compared the validity and reliability of responses to frequently asked questions in the field of endodontics via GPT-3.5, Google Bard, and Bing. Responses were independently evaluated by two endodontists using a 5-point Likert scale and a modified Global Quality Score (GQS). The study concluded that GPT-3.5 responses demonstrated significantly higher validity and all three chatbots exhibited an acceptable level of reliability [19].

Künzle et al. posed 151 questions from the RDE question pool to ChatGPT-3.5, -4.0, -4.0o and Google (Gemini 1.0), subsequently ascertaining that ChatGPT-4.0o demonstrated the highest level of overall answer accuracy. The researchers concluded that ChatGPT-4 models were most suitable for utilisation in dentistry education, owing to their higher success rates in comparison to the other models [20].

Dursun and Bilici geçer [12] evaluated the accuracy and quality of responses provided by ChatGPT-3.5, ChatGPT-4, Gemini and Copilot to frequently asked questions concerning orthodontic clear aligners. The accuracy of responses was assessed using a five-point Likert scale, the reliability was assessed using the modified DISCERN scale, the quality was assessed using the Global Quality Scale (GQS), and the readability was assessed using the Flesch Reading Ease Score (FRES). It was found that responses from ChatGPT-4 received the highest mean Likert score. Gemini's Readability FRES average was found to be higher than the others, and this value shows that Gemini's responses are more readable when compared to other chatbots [12]. Güzelce Sultanoğlu et al. selected fifteen frequently asked questions by patients about missing teeth treatment from the Quora platform. These questions were posed to ChatGPT-4 and Copilot artificial intelligence robots. The responses were evaluated by two experts using a Likert scale to ascertain the accuracy of the responses, and the Global Quality Scale (GQS) to assess the quality of the responses. The study found that the responses produced by ChatGPT-4 and Copilot generally produced reasonably accurate responses, and were evaluated by experts and found to be of similar accuracy [11]. In their study, Acar et al. posed a total of twenty questions to ChatGPT, Microsoft Bing and Google Bard artificial intelligence robots on the subject of oral surgery. The accuracy and completeness of the answers were measured using a Likert Scale (LS), while the clarity of the answers was evaluated using a Global

Quality Scale (GQS). The study's findings indicated that ChatGPT outperformed the other AI robots in both LS and GQS scores [21].

Antaki et al. (2023) evaluated the performance of ChatGPT by posing questions to it and comparing the responses with those of two major ophthalmology reference books. The study found that ChatGPT demonstrated an average accuracy of 57.33%, suggesting that it is capable of answering questions in a manner comparable to that of experts [17]. Gurnoor S. et al. evaluated the accuracy and relevance of ChatGPT-4.0 and Google's Gemini Advanced in answering complex ophthalmic subspecialty medical questions. The findings of this study indicated that ChatGPT-4.0 demonstrated superior performance in comparison to Gemini Advanced. The study further noted that Gemini Advanced exhibited diminished effectiveness in addressing medical-specific inquiries in comparison to ChatGPT-4.0 [22]. It is hypothesised that Google's Gemini Advanced adopts a more cautious and ethical stance to avoid potential harm, which may result in a more risk-averse approach and reduced precision in answering medical-specific questions. This potential difference in performance may have implications for the efficacy of both systems in addressing specific medical inquiries [22, 23]. The findings of our study demonstrated that the mean completeness value was higher and more significant in ChatGPT-4 than in GEMINI.

Nevertheless, certain concerns, including ethical issues and confidentiality, must be given due consideration. In the event of inadvertent entry of private health data into the system by patients or healthcare professionals, the information becomes part of the LLM database, which may result in a violation of patient privacy. The LLM's inability to replicate the human capacity for complex judgment, which is frequently demanded in clinical examinations and treatments, underscores its limited capacity to substitute for human expertise [24].

There are some limitations in our study. Firstly, the results of the study were evaluated by two experts. The use of a limited number of experts may not be representative of the broader range of opinions and criteria found in the wider clinical professional community. Secondly, intelligence robots instantly update every piece of information uploaded to the internet in their memory. Consequently, the response to a query is subject to modification over time. In this study, participants were asked to respond to a question on a single occasion, and the initial answers received were then evaluated.

Conclusion

The present study contributes to the extant knowledge that chatbots can provide logical, accurate and high-quality responses to inquiries regarding root canal treatment. While artificial intelligence is undergoing rapid advancements, it is imperative that it remains within the expert's control.

Acknowledgements

Funding

Competing and Conflicting Interests

The authors declare no competing and conflicting interests.

References

1. Thurzo A., Urbanova W., Novak B., Czako L., Siebert T., Stano P., et al. (2022). Where is the Artificial Intelligence Applied in Dentistry? Systematic Review and Literature Analysis. Healthc (Basel), 10(7).

2. Brown T., Mann B., Ryder N., et al. (2020). Language models are few-shot learners. Adv Neural Inf Process Syst, 33: 1877–1901.

3. Brickley M. R., Shepherd J. P., Armstrong R. A. (1998). Neural networks: a new technique for development of decision support systems in dentistry. J Dent, 26(4): 305–309.

4. Tripathy M., Maheshwari R. P., Verma H. K. (2020). Power transformer differential protection based on optimal probabilistic neural network. IEEE Trans Pow.

5. Mokkink D. H., Dijkstra R., Verbakel D. (2008). Surfende Patiënten. Huisarts & Wetenschap. Retrieved from https://www.henw.org/artikelen/surfende-patienten.

6. Shen Y., Heacock L., Elias J., et al. (2023). ChatGPT and other large language models are double-edged swords. Radiological Society of North America, 23016.

7. Eggmann F., Weiger R., Zitzmann N. U., Blatz M. B. (2023). Implications of large language models such as ChatGPT for dental medicine. J Esthet Restor Dent, 1-5.

8. Strunga M., Urban R., Surovková J., Thurzo A. (2023). Artificial intelligence systems assisting in the assessment of the course and retention of orthodontic treatment. Healthcare (Basel), 11: 683.

9. Shi W., Zhuang Y., Zhu Y., Iwinski H., Wattenbarger M., Wang M. D., editors. (2023). Retrieval-augmented large language models for adolescent idiopathic scoliosis patients in shared decision-making. In: Proceedings of the 14th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics.

10. BOREAK N. (2020). Effectiveness of artificial intelligence applications designed for endodontic diagnosis, decision-making, and prediction of prognosis: A systematic review. The Journal of Contemporary Dental Practice.

11. Güzelce Sultanoğlu, E. (2024). Can natural language processing (NLP) provide consultancy to patients about edentulism teeth treatment? Cureus, 16(10), e70945.

12. Dursun D., Bilici Geçer R. (2024). Can artificial intelligence models serve as patient information consultants in orthodontics? BMC Med Inform Decis Mak, 24: 211.

13. Meyer M. K. R., Kandathil C. K., Davis S. J., Durairaj K. K., Patel P. N., Pepper J. P., Spataro E. A., & Most S. P. (2024). Evaluation of rhinoplasty information from ChatGPT, Gemini, and Claude for readability and accuracy. Aesthetic Plastic Surgery. Advance online publication.

14. Snigdha N. T., Batul R., Karobari M. I., Adil A. H., Dawasaz A. A., Hameed M. S., Mehta V., Noorani T. Y. (2024). Assessing the performance of ChatGPT 3.5 and ChatGPT 4 in operative dentistry and endodontics: An exploratory study. Hum. Behav. Emerg. Tech, 2024, 8.

15. Gilson A., Safranek C. W., Huang T., Socrates V., Chi L., Taylor R. A., & Chartash D. (2023). How does ChatGPT perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment. JMIR Medical Education, 9(1), Article e45312.

16. Alser M., & Waisberg E. (2023). Concerns with the usage of ChatGPT in academia and medicine: A viewpoint. American Journal of Medicine Open, 9, 100036.

17. Antaki F., Touma S., Milad D., El-Khoury J., & Duval R. (2023). Evaluating the performance of ChatGPT in ophthalmology. Ophthalmology Science, 3, 100324.

18. Karobari M. I., Adil A. H., Basheer S. N., et al. (2023). Evaluation of the diagnostic and prognostic accuracy of artificial intelligence in endodontic dentistry: A comprehensive review of literature. Computational and Mathematical Methods in Medicine, 2023, 7049360.

19. Mohammad-Rahimi H., Ourang S. A., Pourhoseingholi M. A., Dianat O., Dummer P. M. H., & Nosrat A. (2024). Validity and reliability of artificial intelligence chatbots as public sources of information on endodontics. International Endodontic Journal, 57(3), 305-314.

20. Künzle P., & Paris S. (2024). Performance of large language artificial intelligence models on solving restorative dentistry and endodontics student assessments. Clinical Oral Investigations, 28(11), 575.

21. Acar A. H. (2024). Can natural language processing serve as a consultant in oral surgery? J Stomatol Oral Maxillofac Surg, 125: 101724.

22. Gill G. S., Tsai J., Moxam J., et al. (2024). Comparison of Gemini Advanced and ChatGPT 4.0's performances on the Ophthalmology Resident Ophthalmic Knowledge Assessment Program (OKAP) Examination Review Question Banks. Cureus, 16(9): e69612.

23. Patil N. S., Huang R. S., van der Pol C. B., Larocque N. (2024). Comparative performance of ChatGPT and Bard in a text-based radiology knowledge assessment. Can Assoc Radiol J, 75: 344-350.

24. Barker F. G., & Rutka J. T. (2023). Editorial: Generative artificial intelligence, chatbots, and the Journal of Neurosurgery Publishing Group. Journal of Neurosurgery, 139(3): 901-903.