Received: 3 September 2023 | Accepted: 27 September 2023

DOI: 10.1111/iej.13985

ORIGINAL ARTICLE

INTERNATIONAL ENDODONTIC JOURNAL | WILEY

Unveiling the ChatGPT phenomenon: Evaluating the consistency and accuracy of endodontic question answers

Ana Suárez[1] | **Víctor Díaz-Flores García**[1] | **Juan Algar**[2] | **Margarita Gómez Sánchez**[1] | **María Llorente de Pedro**[1] | **Yolanda Freire**[1]

[1]Department of Pre-Clinic Dentistry, School of Biomedical Sciences, Universidad Europea de Madrid, Madrid, Spain

[2]Department of Clinical Dentistry, School of Biomedical Sciences, Universidad Europea de Madrid, Madrid, Spain

Correspondence
Víctor Díaz-Flores García, Department of Pre-Clinic Dentistry, School of Biomedical Sciences, Universidad Europea de Madrid, Calle Tajo s/n, Villaviciosa de Odón, Madrid 28670, Spain.
Email: victor.diaz-flores@universidadeuropea.es

Abstract

Aim: Chatbot Generative Pre-trained Transformer (ChatGPT) is a generative artificial intelligence (AI) software based on large language models (LLMs), designed to simulate human conversations and generate novel content based on the training data it has been exposed to. The aim of this study was to evaluate the consistency and accuracy of ChatGPT-generated answers to clinical questions in endodontics, compared to answers provided by human experts.

Methodology: Ninety-one dichotomous (yes/no) questions were designed and categorized into three levels of difficulty. Twenty questions were randomly selected from each difficulty level. Sixty answers were generated by ChatGPT for each question. Two endodontic experts independently answered the 60 questions. Statistical analysis was performed using the SPSS program to calculate the consistency and accuracy of the answers generated by ChatGPT compared to the experts. Confidence intervals (95%) and standard deviations were used to estimate variability.

Results: The answers generated by ChatGPT showed high consistency (85.44%). No significant differences in consistency were found based on question difficulty. In terms of answer accuracy, ChatGPT achieved an average accuracy of 57.33%. However, significant differences in accuracy were observed based on question difficulty, with lower accuracy for easier questions.

Conclusions: Currently, ChatGPT is not capable of replacing dentists in clinical decision-making. As ChatGPT's performance improves through deep learning, it is expected to become more useful and effective in the field of endodontics. However, careful attention and ongoing evaluation are needed to ensure its accuracy, reliability and safety in endodontics.

KEYWORDS

artificial intelligence, chatbot, ChatGPT, dentistry, endodontics, large language models

INTRODUCTION

Artificial intelligence (AI) refers to the ability of a computer system to perform tasks that typically require human intelligence (Rodrigues et al., 2021). It is based on the development of algorithms that enable machines to process large amounts of data, learn from it, solve problems adapt and improve their performance over time (Deng, 2018).

At the end of 2022, the Chatbot Generative Pre-trained Transformer (ChatGPT) developed by OpenAI was launched (Cadamuro et al., 2023). This generative AI software is based on large language models (LLMs) and is designed to simulate human conversations, understanding the meaning of words and phrases much like humans do and using that understanding to generate new content based on the training data it has been exposed to (Abd-Alrazaq et al., 2023). ChatGPT works with algorithms that interpret natural language and provide predefined or real-time-generated appropriate answers (Plebani, 2023), offering a wide range of applications with endless possibilities, ranging from creative and recreational activities to idea exchange, code writing, scientific articles, text editing and many others (Cadamuro et al., 2023; Li et al., 2023).

However, the use of ChatGPT faces difficulties due to limitations in the training data, as it only has access to various Internet data until the year 2021, which hinders the generation of current texts. Furthermore, the lack of access to relevant databases further limits its ability to provide accurate and reliable information, raising questions about the credibility and trustworthiness of the results generated by ChatGPT (Arif et al., 2023). It has also been observed that ChatGPT is capable of convincingly generating false or inaccurate information, which poses a challenge for individuals who lack experience in distinguishing between accurate and deceptive information (Li et al., 2023). For this reason, authors such as Abd-Alrazaq et al. (2023) caution that this technology should be considered as an assistive tool rather than a direct provider of information.

In the field of medicine, although ChatGPT has not been specifically trained with medical data, this model has been shown to be able to pass the United States Medical Licensing Exam (USMLE) (Kung et al., 2023). Studies in specific areas have reported that ChatGPT demonstrated an accuracy and comprehension level of 86.8% when answering questions about bariatric surgery (Samaan et al., 2023), 79.1% for cirrhosis (Yeo et al., 2023), 74% for hepatocellular carcinoma (Yeo et al., 2023), 80% for microbiology (Das et al., 2023) and 60.8% for parasitology (Huh, 2023). These results highlight the potential of ChatGPT as a consultation tool in medicine, although further detailed and rigorous analysis is needed to assess its performance across different medical specialties (Kung et al., 2023).

In this regard, endodontics is considered a challenging discipline for general dentists (Dahlström et al., 2017). As a branch of dentistry, it focuses on the diagnosis, prevention and treatment of diseases affecting the dental pulp and periapical tissues (Sigurdsson, 2003). The diverse nature of the pathologies, symptoms and treatment options associated with endodontics justifies the importance of investigating whether large language models (LLMs), such as ChatGPT, could provide accurate and reliable answers to specific questions in this field. If research results show that this generative AI software is capable of providing reliable and timely answers to frequently asked questions, it could become a valuable support tool for general dentists as a virtual assistant in clinical practice, streamlining decision-making in endodontics.

Given the above, the main objective of this study was to evaluate the consistency and accuracy of the answers generated by ChatGPT compared to those provided by human experts on dichotomous questions related to endodontics.

METHOD

Ethical approval

Ethical approval was not required for this research as no human participants were involved in the study.

Question design

The Position Statements of the European Society of Endodontology (ESE) (2023) were used to design the questions. These articles were selected because they represent the consensus reached by expert committees on various topics related to the field of endodontics.

A general dentist and a specialist (A.S. and Y.F.) created a total of 91 dichotomous (Yes/No) questions with a specific prompt format (Only a yes or no answer) for ChatGPT to answer with either Yes or No. Example: 'In a tooth autotransplantation, should the tooth be placed in occlusion? Only a yes or no answer'.

The decision to ask for 'yes' or 'no' responses was based on several considerations. First, we wanted to simplify the evaluation approach to allow for a direct comparison of ChatGPT's responses with those of a group of experts. This simplification allowed for a more quantitative and objective measure of the accuracy of the answers provided by the model. In addition, by limiting the responses to a binary format, we aimed to avoid the partially correct or incorrect answers that might occur when more descriptive responses are requested.

Categorization of questions

The questions were classified by two general dentists (M.LL.dP. and M.G.S.) into three levels of difficulty (easy, medium and difficult). Since the classifying dentists were not endodontic specialists, an appropriate perspective for nonexpert dentists in this field was ensured. Prior to beginning the classification process, both raters jointly defined the criteria they would use to determine the difficulty of the question. These criteria included factors such as the complexity of the endodontic procedure involved, the level of knowledge required and the likelihood of an incorrect response by a general dentist. When discrepancies arose in certain answers, they were resolved by the intervention of a third member (A.S.). Subsequently, 20 questions from each level of difficulty were randomly selected for inclusion in the study: 20 easy questions (E), 20 medium questions (M) and 20 difficult questions (D) (Appendix S1).

Generating answers in ChatGPT

Two authors (Y.F. and A.S.), using four different accounts, performed the answer generation using the ChatGPT-4 platform. ChatGPT answers were obtained at three different times of the day (morning, afternoon and evening) by selecting the 'new chat' option over a period of 10 days, until a total of 60 answers per question were obtained. This approach allowed us to examine whether the answers were consistent or whether changes or fluctuations were observed at different times of the day, with the goal of obtaining a more accurate representation of the overall model performance. The answers were stored in an Excel spreadsheet.

Human expert answers

After completion of the ChatGPT answer generation, two endodontic experts (V.D.-F.G. and J.A.) independently answered the 60 questions. When discrepancies arose in certain answers, they were resolved by the intervention of a third member (Y.F.), who consulted reference articles to reach a consensus.

Statistical analysis

All answers were stored in an Excel spreadsheet (Microsoft) and analysed using the statistical software program STATA version BE 14 (StataCorp).

To analyse the consistency of the responses generated by ChatGPT, the proportion of matching responses was calculated for the total responses in the question set and for each individual question, along with its 95% confidence interval (binomial Wald method). To assess the accuracy of ChatGPT-generated responses compared to expert responses, the proportion of exact matches was calculated for the total set of questions and for each individual question, along with its 95% confidence interval (binomial Wald method).

Chi-square tests were used to determine the presence of significant differences in the consistency and validity of answers amongst questions of different difficulty levels ('difficult', 'medium' and 'easy'), with a significance level of 5%.

RESULTS

Analysing the results, it was found that the overall consistency of the answers generated by ChatGPT, based on the time of day when the questions were asked, was 85.44%. Regarding the analysis of the accuracy of the answers generated by ChatGPT, the results show that the model has an overall low variability, suggesting a reasonable consistency in the accuracy of the answers generated by ChatGPT (Table 1; Appendix S2).

Regarding the difficulty of the questions, it was observed that consistency did not show significant differences between questions categorized as E (easy), M (medium) or D (difficult). When analysing accuracy based on question difficulty, significant differences were found. ChatGPT showed poorer performance in generating accurate answers for questions of lower difficulty (Table 2).

DISCUSSION

In recent months, there has been a growing interest in the application of LLMs in medicine, particularly in exploring their clinical utility, as evidenced by the emergence of ChatGPT (Antaki et al., 2023; Ge & Lai, 2023; Lahat et al., 2023). Despite the promising results demonstrated by these models, it is crucial to perform a comprehensive

TABLE 1 Overall consistency and accuracy results.

Metric	Value (%)	Standard deviation	CI 95% (Wald binomial)	
Consistency	85.44	0.0295	79.54%	91.35%
Accuracy	57.33	0.0082432	55.72%	58.95%

TABLE 2 Consistency and accuracy results according to the difficulty of the question.

Difficulty	Metric	Value (%)	Standard deviation	CI 95% (Wald binomial)		*p*-Value
E	Consistency	86.75	0.0564	74.95%	98.55%	.066
	Accuracy	49.25	0.0144321	46.42%	52.08%	<.001
M	Consistency	79.41	0.0583	67.21%	91.63%	.066
	Accuracy	64.58	0.0138062	61.88%	67.29%	<.001
D	Consistency	90.17	0.0347	82.89%	97.44%	.066
	Accuracy	58.17	0.0142399	55.38%	60.96%	<.001

evaluation of their performance and potential errors before determining their viability in a clinical setting (Antaki et al., 2023). In this context, a study was conducted to evaluate the consistency and accuracy of answers provided by ChatGPT to questions related to clinical situations in endodontics.

Due to the probabilistic nature of LLMs, which can lead to variability in their answers (Antaki et al., 2023), this study analysed the agreement of the 60 answers for each question. It was observed that the model achieved a satisfactory level of consistency, reaching 85.4%. Thus, in the majority of cases, the answers obtained remained consistent at different times of the day. Furthermore, when comparing the answers of ChatGPT with those of the experts, an average accuracy of 57.33% was obtained. Umer and Habib (2022) argue that an acceptable threshold for accuracy in diagnostic tasks should be set at >90%. However, our result is similar to other studies (Antaki et al., 2023; Gilson et al., 2023; Huh, 2023; Kung et al., 2023; Thirunavukarasu et al., 2023), that have recognized the promising potential of this LLM for medical education and clinical decision-making. However, some studies have warned of the variability in response accuracy, with incomplete or incorrect answers being common (Lahat et al., 2023; Samaan et al., 2023).

These types of incorrect answers have raised concern amongst authors because of their potential implications (Juhi et al., 2023). In fact, it is already known that the model, rather than demonstrating its inability to perform a task, can fabricate information and bibliographic references in a highly convincing manner (Seth et al., 2023; Thirunavukarasu et al., 2023). This suggests that whilst ChatGPT can identify patterns and organize data, it has limitations in fully understanding the underlying meaning and context of the information (Sinha et al., 2023). To mitigate the problems of generating 'hallucinatory' or fabricated responses (Athaluri et al., 2023; Masters, 2023; Thirunavukarasu et al., 2023), this study took a cautious and specific approach to eliciting answers about endodontics. This was achieved using a prompt that clearly indicated the desired type of answer: 'Only yes or no answer'. This restriction is similar to the old practice of making a wish to a genie in a lamp, where the precise wording of the wish is as important as the wish itself. Using such a specific directive, the goal was to limit answers to a dichotomous option, thereby avoiding the generation of additional or fictitious information. However, this approach has limitations and does not fully capture the nuanced and contextual nature of actual clinical responses.

In dentistry, various applications of AI have been proposed, including aiding diagnosis, improving dental telemedicine services, supporting decision-making, optimizing clinical workflow, reducing costs, supporting research and improving teaching (Alhaidry et al., 2023; Eggmann et al., 2023; Sallam, 2023; Suárez et al., 2022). Few studies have evaluated the answers provided by AI in the field of endodontics. Suárez et al. (2022) developed an AI chatbot that simulated a virtual patient to study pulpal pathology in endodontics and achieved positive results in improving students' communication skills and confidence. However, it is important to note that there is a crucial difference between the creation and training of an AI chatbot by a group of experts and the use of general LLMs such as ChatGPT, which retrieve information from various sources, some of which may be unverified. This may result in inaccurate or unreliable content, which could be detrimental to both patients and dental professionals (Alhaidry et al., 2023). Therefore, before implementing this technology, it is critical to consider the inherent limitations and potential risks to ensure the accuracy and reliability of the information provided (Eggmann et al., 2023).

Limitations and future directions

Despite the results obtained in this study, it is important to consider that ChatGPT is a language model designed for a general audience and was not specifically trained for the field of endodontics, which may have introduced biases in the generated responses. To address this limitation, the creation of scientifically validated data sets agreed upon by endodontists is suggested. It is critical to encourage

collaboration between endodontists and medical informatics professionals to develop standardized data sets and assessments that allow for a more accurate and rigorous comparison of the performance of LLMs in the clinical context.

CONCLUSION

Currently, ChatGPT is not capable of replacing dentists in clinical decision-making. As ChatGPT's performance improves through deep learning, it is expected to become more useful and effective in the field of endodontics. It is important to conduct ongoing studies over time to assess the learning level of the AI and its ability to improve. However, careful attention and ongoing evaluation are needed to ensure its accuracy, reliability and safety in endodontics.

AUTHOR CONTRIBUTIONS

Conceptualization, investigation, formal analysis, software and writing—original draft preparation (A.S., Y.F. and V.D.-F.G.); Investigation and writing—review and editing (J.A., M.G.S. and M.LL.dP.).

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no conflicts of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

ORCID

Ana Suárez https://orcid.org/0000-0003-2448-6669
Víctor Díaz-Flores García https://orcid.org/0000-0001-6141-0546
Margarita Gómez Sánchez https://orcid.org/0000-0002-8305-6354
Yolanda Freire https://orcid.org/0000-0002-0727-777X

REFERENCES

Abd-Alrazaq, A., AlSaad, R., Alhuwail, D., Ahmed, A., Healy, P.M., Latifi, S. et al. (2023) Large language models in medical education: opportunities, challenges, and future directions. *JMIR Medical Education*, 9, e48291. Available from: https://doi.org/10.2196/48291

Alhaidry, H.M., Fatani, B., Alrayes, J.O., Almana, A.M. & Alfhaed, N.K. (2023) ChatGPT in dentistry: a comprehensive review. *Cureus.*, 15, e38317. Available from: https://doi.org/10.7759/cureus.38317

Antaki, F., Touma, S., Milad, D., El-Khoury, J. & Duval, R. (2023) Evaluating the performance of ChatGPT in ophthalmology. *Ophthalmology Science*, 3, 100324. Available from: https://doi.org/10.1016/j.xops.2023.100324

Arif, T.B., Munaf, U. & Ul-Haque, I. (2023) The future of medical education and research: is ChatGPT a blessing or blight in disguise? *Medical Education Online*, 28, 2181052. Available from: https://doi.org/10.1080/10872981.2023.2181052

Athaluri, S.A., Manthena, S.V., Kesapragada, V.S.R.K.M., Yarlagadda, V., Dave, T. & Duddumpudi, R.T.S. (2023) Exploring the boundaries of reality: investigating the phenomenon of artificial intelligence hallucination in scientific writing through ChatGPT references. *Cureus*, 15, e37432. Available from: https://doi.org/10.7759/cureus.37432

Cadamuro, J., Cabitza, F., Debeljak, Z., De Bruyne, S., Frans, G., Perez, S.M. et al. (2023) Potentials and pitfalls of ChatGPT and natural-language artificial intelligence models for the understanding of laboratory medicine test results. An assessment by the European Federation of Clinical Chemistry and Laboratory Medicine (EFLM) Working Group on Artificial Intelligence (WG-AI). *Clinical Chemistry and Laboratory Medicine*, 61, 1158–1166. Available from: https://doi.org/10.1515/cclm-2023-0355

Dahlström, L., Lindwall, O., Rystedt, H. & Reit, C. (2017) 'Working in the dark': Swedish general dental practitioners on the complexity of root canal treatment. *International Endodontic Journal*, 50, 636–645. Available from: https://doi.org/10.1111/iej.12675

Das, D., Kumar, N., Longjam, L.A., Sinha, R., Deb Roy, A., Mondal, H. et al. (2023) Assessing the capability of ChatGPT in answering first- and second-order knowledge questions on microbiology as per competency-based medical education curriculum. *Cureus*, 15, e36034. Available from: https://doi.org/10.7759/cureus.36034

Deng, L. (2018) Artificial intelligence in the rising wave of deep learning: the historical path and future outlook [Perspectives]. *IEEE Signal Processing Magazine*, 35(1), 180–187. Available from: https://doi.org/10.1109/MSP.2017.2762725

Eggmann, F., Weiger, R., Zitzmann, N.U. & Blatz, M.B. (2023) Implications of large language models such as ChatGPT for dental medicine. *Journal of Esthetic and Restorative Dentistry*, 5. Available from: https://doi.org/10.1111/jerd.13046

European Society of Endodontology. (2023) *Resources for clinicians*. [WWW Document]. Available from: https://www.e-s-e.eu/for-professionals/resources-for-clinicians/ [Accessed 6th August 2023].

Ge, J. & Lai, J.C. (2023) Artificial intelligence-based text generators in hepatology: ChatGPT is just the beginning. *Hepatology Communications*, 7, e0097. Available from: https://doi.org/10.1097/HC9.0000000000000097

Gilson, A., Safranek, C.W., Huang, T., Socrates, V., Chi, L., Taylor, R.A. et al. (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, e45312. Available from: https://doi.org/10.2196/45312

Huh, S. (2023) Are ChatGPT's knowledge and interpretation ability comparable to those of medical students in Korea for taking a parasitology examination?: A descriptive study. *Journal of*

Educational Evaluation for Health Professions, 20, 1. Available from: https://doi.org/10.3352/jeehp.2023.20.1

Juhi, A., Pipil, N., Santra, S., Mondal, S., Behera, J.K. & Mondal, H. (2023) The capability of ChatGPT in predicting and explaining common drug-drug interactions. *Cureus*, 15, e36272. Available from: https://doi.org/10.7759/cureus.36272

Kung, T.H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C. et al. (2023) Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLoS Digital Health*, 2, e0000198. Available from: https://doi.org/10.1371/journal.pdig.0000198

Lahat, A., Shachar, E., Avidan, B., Glicksberg, B. & Klang, E. (2023) Evaluating the utility of a large language model in answering common patients' gastrointestinal health-related questions: are we there yet? *Diagnostics*, 13, 1950. Available from: https://doi.org/10.3390/diagnostics13111950

Li, H., Moon, J.T., Purkayastha, S., Celi, L.A., Trivedi, H. & Gichoya, J.W. (2023) Ethics of large language models in medicine and medical research. *The Lancet Digital Health*, 5, e333–e335. Available from: https://doi.org/10.1016/S2589-7500(23)00083-3

Masters, K. (2023) Medical teacher's first ChatGPT's referencing hallucinations: lessons for editors, reviewers, and teachers. *Medical Teacher*, 45, 673–675. Available from: https://doi.org/10.1080/0142159X.2023.2208731

Plebani, M. (2023) ChatGPT: angel or demond? Critical thinking is still needed. *Clinical Chemistry and Laboratory Medicine*, 61, 1131–1132. Available from: https://doi.org/10.1515/cclm-2023-0387

Rodrigues, J.A., Krois, J. & Schwendicke, F. (2021) Demystifying artificial intelligence and deep learning in dentistry. *Brazilian Oral Research*, 35, e094. Available from: https://doi.org/10.1590/1807-3107bor-2021.vol35.0094

Sallam, M. (2023) ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare*, 11, 887. Available from: https://doi.org/10.3390/healthcare11060887

Samaan, J.S., Yeo, Y.H., Rajeev, N., Hawley, L., Abel, S., Ng, W.H. et al. (2023) Assessing the accuracy of responses by the language model ChatGPT to questions regarding bariatric surgery. *Obesity Surgery*, 33, 1790–1796. Available from: https://doi.org/10.1007/s11695-023-06603-5

Seth, I., Sinkjær Kenney, P., Bulloch, G., Hunter-Smith, D.J., Bo Thomsen, J. & Rozen, W.M. (2023) Artificial or augmented authorship? A conversation with a chatbot on base of thumb arthritis. *Plastic and Reconstructive Surgery. Global Open*, 11, e4999. Available from: https://doi.org/10.1097/GOX.0000000000004999

Sigurdsson, A. (2003) Pulpal diagnosis. *Endodontic Topics*, 5, 12–25. Available from: https://doi.org/10.1111/j.1601-1546.2003.00024.x

Sinha, R.K., Deb Roy, A., Kumar, N. & Mondal, H. (2023) Applicability of ChatGPT in assisting to solve higher order problems in pathology. *Cureus.*, 15, e35237. Available from: https://doi.org/10.7759/cureus.35237

Suárez, A., Adanero, A., Díaz-Flores García, V., Freire, Y. & Algar, J. (2022) Using a virtual patient via an artificial intelligence chatbot to develop dental students' diagnostic skills. *International Journal of Environmental Research and Public Health*, 19, 8735. Available from: https://doi.org/10.3390/ijerph19148735

Thirunavukarasu, A.J., Hassan, R., Mahmood, S., Sanghera, R., Barzangi, K., El Mukashfi, M. et al. (2023) Trialling a large language model (ChatGPT) in general practice with the applied knowledge test: observational study demonstrating opportunities and limitations in primary care. *JMIR Medical Education*, 9, e46599. Available from: https://doi.org/10.2196/46599

Umer, F. & Habib, S. (2022) Critical analysis of artificial intelligence in endodontics: a scoping review. *Journal of Endodontia*, 48, 152–160. Available from: https://doi.org/10.1016/j.joen.2021.11.007

Yeo, Y.H., Samaan, J.S., Ng, W.H., Ting, P.-S., Trivedi, H., Vipani, A. et al. (2023) Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. *Clinical and Molecular Hepatology*, 29, 721–732. Available from: https://doi.org/10.3350/cmh.2023.0089

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Suárez, A., Díaz-Flores García, V., Algar, J., Gómez Sánchez, M., Llorente de Pedro, M. & Freire, Y. (2023) Unveiling the ChatGPT phenomenon: Evaluating the consistency and accuracy of endodontic question answers. *International Endodontic Journal*, 00, 1–6. Available from: https://doi.org/10.1111/iej.13985