Scientific Research Report

Comprehensiveness of Large Language Models in Patient Queries on Gingival and Endodontic Health

Qian Zhang [a,b,c], Zhengyu Wu [a,b,c], Jinlin Song [a,b,c], Shuicai Luo [d], Zhaowu Chai [a,b,c*]

[a] College of Stomatology, Chongqing Medical University, Chongqing, China
[b] Chongqing Key Laboratory for Oral Diseases and Biomedical Sciences, Chongqing, China
[c] Chongqing Municipal Key Laboratory of Oral Biomedical Engineering of Higher Education, Chongqing, China
[d] Quanzhou Institute of Equipment Manufacturing, Haixi Institute, Chinese Academy of Sciences, Quanzhou, China

ARTICLE INFO

Article history:
Received 24 January 2024
Received in revised form
12 June 2024
Accepted 19 June 2024
Available online 14 August 2024

Key words:
Artificial intelligence
Large language models
Oral healthcare
Gingival and endodontic health

ABSTRACT

Aim: Given the increasing interest in using large language models (LLMs) for self-diagnosis, this study aimed to evaluate the comprehensiveness of two prominent LLMs, ChatGPT-3.5 and ChatGPT-4, in addressing common queries related to gingival and endodontic health across different language contexts and query types.

Methods: We assembled a set of 33 common real-life questions related to gingival and endodontic healthcare, including 17 common-sense questions and 16 expert questions. Each question was presented to the LLMs in both English and Chinese. Three specialists were invited to evaluate the comprehensiveness of the responses on a five-point Likert scale, where a higher score indicated greater quality responses.

Results: LLMs performed significantly better in English, with an average score of 4.53, compared to 3.95 in Chinese (Mann–Whitney *U* test, $P < .05$). Responses to common sense questions received higher scores than those to expert questions, with averages of 4.46 and 4.02 (Mann–Whitney *U* test, $P < .05$). Among the LLMs, ChatGPT-4 consistently outperformed ChatGPT-3.5, achieving average scores of 4.45 and 4.03 (Mann–Whitney *U* test, $P < .05$).

Conclusions: ChatGPT-4 provides more comprehensive responses than ChatGPT-3.5 for queries related to gingival and endodontic health. Both LLMs perform better in English and on common sense questions. However, the performance discrepancies across different language contexts and the presence of inaccurate responses suggest that further evaluation and understanding of their limitations are crucial to avoid potential misunderstandings.

Clinical Relevance: This study revealed the performance differences of ChatGPT-3.5 and ChatGPT-4 in handling gingival and endodontic health issues across different language contexts, providing insights into the comprehensiveness and limitations of LLMs in addressing common oral healthcare queries.

Introduction

The trend towards virtual healthcare, accelerated by the COVID-19 pandemic, continues. Patients have become accustomed to using online portals for accessing health information and communicating with healthcare providers.[1] However, existing healthcare systems face challenges in managing the increasing electronic workload. Many clinicians spend significant time handling a large volume of online messages, often beyond their regular working hours.[2,3] Addressing online messages without protected time or specific reimbursement often leads to patients receiving inaccurate and flawed responses about health-related information.[4]

Approximately two-thirds of adults in the United States seek health information on the internet, with more than one-third using online search engines for self-diagnosis.[5] This trend is particularly notable in the field of oral health, specifically regarding gingival and endodontic issues, where

* *Corresponding author*. Stomatological Hospital of Chongqing Medical University Chongqing, 426 Songshibei Road, Chongqing 401147, China.
E-mail address: 500732@hospital.cqmu.edu.cn (Z. Chai).

https://doi.org/10.1016/j.identj.2024.06.022

patients often turn to online resources.[6,7] Recent research has indicated a preference for responses from large language models (LLMs) over those from physicians on social media platforms, suggesting that artificial intelligence (AI) can significantly improve the quality of medical advice available online.[8] The rise of LLMs reflects a growing reliance among patients on these models for information on managing oral health.

LLMs are AI systems that mimic human language processing via deep learning and neural networks trained on extensive text data from various sources, including books, articles, and websites.[9] ChatGPT, a leading example, leverages self-supervised learning to produce responses that closely resemble human conversation, thanks to its training on vast amounts of text data.[10,11] In dental medicine, LLMs have recently attracted significant attention. A study revealed that recent versions of the ChatGPT have shown high proficiency in answering multiple-choice questions derived from dental licensing exams. In the context of professional licensure, the accuracy of oral health information provided by the latest ChatGPT models can rival that of human dental graduates.[12] This indicates that LLMs could offer substantial benefits for dental medicine education and clinical decision-making. Furthermore, several studies have explored the potential of LLMs to help patients access valuable information.[13,14]

Suárez et al. conducted a study to assess the accuracy and consistency of responses generated by ChatGPT compared to those provided by human experts for binary questions in endodontics. The findings indicated that ChatGPT exhibited a high level of consistency and provided valuable insights. However, ChatGPT is not currently capable of replacing dentists in clinical decision-making.[15] Despite the advancements in LLMs, there remains a lack of conclusive evidence on their performance in specific domains of dental medicine, particularly regarding the effectiveness of LLMs in addressing oral healthcare queries.[16] Furthermore, the exploration of the utility of LLMs in non-English-speaking regions, such as China, is an area for further investigation.

By systematically organizing questions that patients frequently pose about gingival and endodontic healthcare, we simulated a series of real clinical scenarios from the patient's perspective. In these scenarios, patients could leverage LLMs to seek medical information and advice in either Chinese or English. We also considered the medical professionalism of the questions, exploring the potential of LLMs across diverse question categories. We aimed to provide a comprehensive understanding of the capabilities and limitations of LLMs in addressing patient queries related to oral healthcare in different language settings. This work may contribute to facilitating the development and application of LLMs in the field of dental medicine, thereby improving the quality of care and information available to patients.

Methods

Question design

This study was collaborated on by two experts specializing in endodontic diseases and one specializing in periodontal diseases, each with over a decade of extensive clinical and academic experience. Their collaboration involved synthesizing insights from authoritative position statements issued by the European Society of Endodontology[17] and the American Association of Endodontists.[18] Questions were developed based on the guidelines of WebMD Health Corporation (WebMD) (*www.webmd.com/oral-health/dental-health-faq*) and the International Dental Health Foundation (*www.dentalhealth.org/Pages/FAQs/Category/general-faqs*) 'Frequently Asked Oral Health Questions' webpage. Subsequently, a series of real questions frequently encountered by patients in their clinical settings were compiled to further refine the question design. Ultimately, a set of 33 common questions related to gingival and endodontic healthcare was crafted.

Categorization of questions

To gain a deeper understanding of the strengths and weaknesses of LLMs across different domains, the questions were categorized into two groups: 17 Common Sense Questions and 16 Expertise Questions (Supplemental Table S1).

Study design

The study design, illustrated in Figure 1, involved a carefully planned set of 33 questions in both Chinese and English. The English version was a translation of the original Chinese questions. ChatGPT-3.5 and ChatGPT-4 were presented with the same questions from 6 December 2023 to 14 January 2024. The conversation was initiated with the prompt 'I have some questions about oral healthcare' to set the context. Subsequently, the full texts of the 33 questions were given to ChatGPT-3.5 and ChatGPT-4, and their responses were recorded. To avoid memory bias, the conversation was reset after each query, and each query was performed only once to prevent model optimization or answer alteration. All responses were presented in plain text to prevent distinguishing between the LLMs. The responses were then shuffled and evaluated by three experts in two rounds, with a 72-hour washout period between assessments to reduce recency bias. The process was repeated with the language setting changed to Chinese.

Scoring system and mechanism

The scoring panel comprised the three experts mentioned previously. To preserve objectivity, the identities of the LLMs were concealed. Our evaluation focused on 'comprehensiveness' using a five-point Likert scale[14,19] based on the NCC MERP[20] classification system. The criteria for assessment were as follows:

1) Very poor: responses with inaccuracies that could mislead and harm patients.
2) Poor: responses with factual errors but unlikely to mislead.
3) Acceptable: moderately comprehensive with a fair amount of detail.
4) Good: comprehensive, covering most necessary aspects.
5) Very good: very comprehensive, providing exhaustive details.

The evaluators' main task was to assess the 'comprehensiveness' of each response generated by LLMs

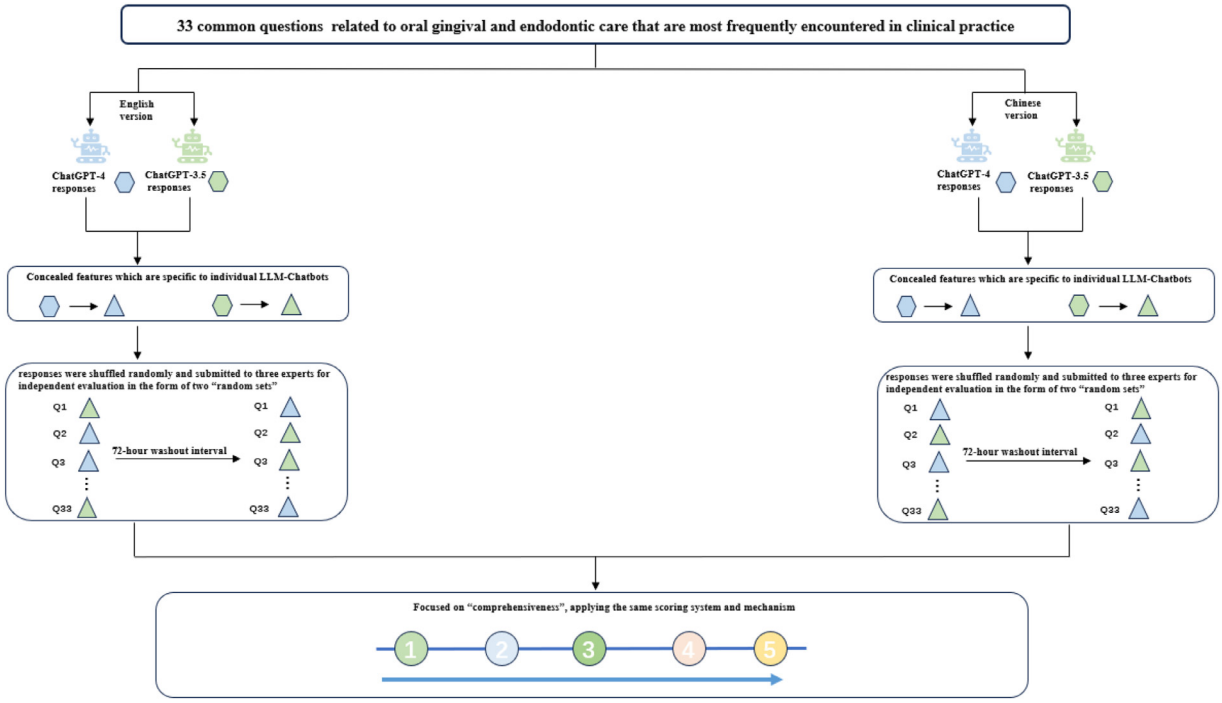

Fig. 1 – Flowchart of overall study design.

independently. The same scoring system was applied to each response, where higher quality responses were closer to a score of 5 and lower quality responses closer to a score of 1. Additionally, a majority consensus method was employed to determine the final score of each chatbot response based on the most common scores among the three evaluators. In cases where there was no consensus among the evaluators (ie, each evaluator provided a different score), the approach taken was to assign the lowest score to the LLMs' responses.[13]

Detailed qualitative analysis of LLMs' responses containing incorrect information

To gain a deeper understanding of the potential drawbacks and risks of relying solely on LLMs for oral healthcare information, we conducted a thorough analysis of responses scoring below 3. This analysis included a detailed review of responses from LLMs that contained incorrect information.

Statistical analysis

Statistical calculations were performed using IBM SPSS Statistics version 26.0 to compare the average comprehensiveness scores across two different LLMs, two question categories, and two language environments. Due to the non-normal distribution of the data, nonparametric tests were utilized. The Mann–Whitney *U* test was applied, with p values less than 0.05 considered to indicate statistical significance. Furthermore, Fleiss's kappa was used to assess the interrater reliability among the three experts.

Results

Comprehensive evaluation

All responses from the LLMs are summarized in the Supplemental Materials. The answers in both Chinese and English are shown in Supplemental Table S3-S6. The three experts

Table – The Mann–Whitney *U* test of average comprehensive scores among two different LLMs, two question categories, and two language environments.

	Mean	Median	Minimum	Maximum	Std. D	N	P value
LLMs type							
ChatGPT-3.5	4.03	4	1	5	.803	66	.000
ChatGPT-4.0	4.45	5	1	5	.807	66	
Category							
Common sense	4.46	5	3	5	.656	68	.004
Expertise	4.02	4	1	5	.934	64	
Language type							
Chinese	3.95	4	1	5	.867	66	.000
English	4.53	5	2	5	.684	66	

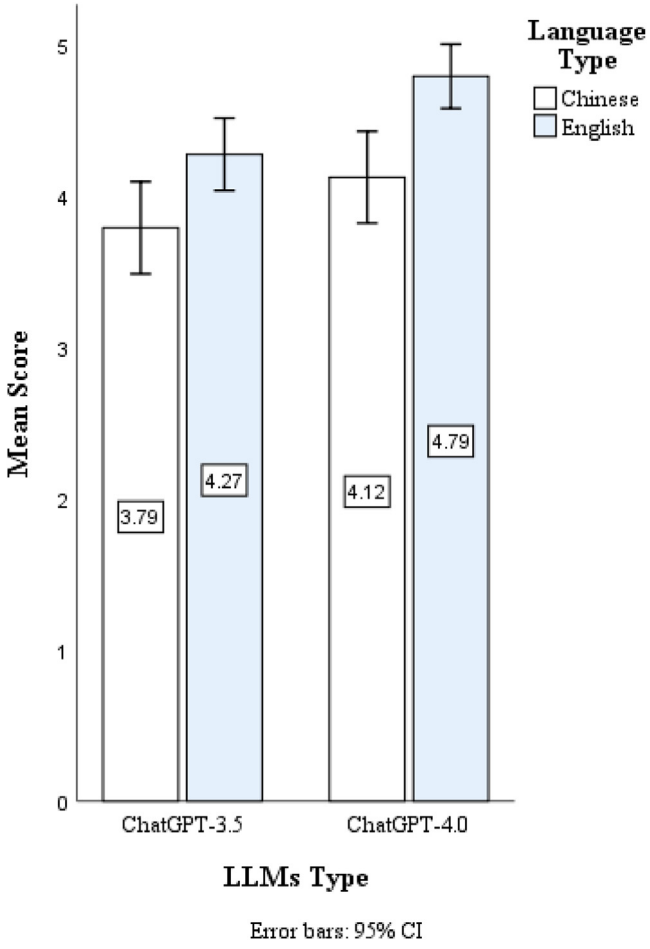

Fig. 2 – Detailed presentation of the average scores of two different LLM-Chatbots in different language environments.

demonstrated substantial agreement in evaluating the responses (Fleiss's $\kappa = 0.870$). The distribution of scores among the three experts is provided in Supplemental Table S8.

Table shows that LLMs achieved a significantly higher average score of 4.53 in English compared to 3.95 in Chinese. Statistically significant differences in the scores between the different language versions were observed ($P < .05$).

Figure 2 provides a detailed presentation of the average scores of two different LLMs in different language environments, Chinese and English.

Table also provides descriptive statistics for LLMs across different question categories. The average score for expertise questions was significantly lower than that for common sense questions ($P < .05$).

Figure 3 visually represents the average scores of LLMs in different question categories.

Furthermore, regardless of the language context or question type, ChatGPT-4 achieved a higher average comprehensive score than ChatGPT-3.5 (ChatGPT-3.5: 4.03; ChatGPT-4: 4.45). A statistically significant difference in scores between the two LLMs was observed ($P < .05$, Table).

Supplemental Table S2 offers a detailed explanation of the scores assigned by each LLM to each question.

Analysis of misinformation in LLM responses

Supplemental Table S7 displays examples of misinformation conveyed by LLMs, with sections containing errors highlighted in bold, based on insights from three experienced expert evaluators. The experts provided further explanations of these errors.

Discussion

In this study, we evaluated the ability of the ChatGPT-3.5 and ChatGPT-4 to address frequently asked questions related to oral healthcare in different language environments (English and Chinese) and across types of queries (common sense and expertise). Distinguishing from previous research that relied on standardized tests, we examined the practical application of LLMs in responding to patients' questions, specifically focusing on oral healthcare. We simulated realistic scenarios in which patients might consult LLMs for medical information and advice, emphasizing the importance of evaluating the comprehensiveness and impact of LLMs' responses in these practical scenarios.

Overall, the English version of the ChatGPT-4 demonstrated exceptional performance. These findings reinforce previous research showing that ChatGPT-4 is more effective than ChatGPT-3.5 in the medical field due to advancements in its training processes and model architecture.[8,21,22]

In different question categories, all LLMs exhibited strong performance in answering common sense questions in both Chinese and English. However, their performance declined when addressing questions requiring expertise, especially in Chinese. This underscores the accuracy and effectiveness of LLMs in responding to general medical inquiries while highlighting potential inaccuracies in more specialized fields, such as disease treatment and control. One plausible explanation is that the models are general-purpose LLMs, not specifically designed for medical contexts. Fine-tuning the models with additional training in the medical domain could prove beneficial.[23] Previous research has shown that fine-tuning LLMs can improve their accuracy in answering medical queries.[24] Future enhancements of LLMs in medical questions may involve expanding medical databases and including keyword prompts in queries.[25]

Considering the vital importance of accurate medical information, potential errors in LLMs warrant concern, especially in high-risk areas related to treatment. While ChatGPT-4 and similar products signify substantial advancements in AI-powered knowledge bases and understanding capabilities, there is a risk that LLMs might unintentionally incorporate and disseminate biases or errors from their training data. Therefore, when using LLMs such as ChatGPT-4 for diagnostic or therapeutic assistance, it is crucial that patients are aware of their current limitations and capabilities. The capacity of LLMs to comprehend and analyse text is restricted to a specific context window, lacking the ability to incorporate knowledge from previous interactions. Patients should verify advice across different sources, particularly in fields with high variability, as AI-generated outputs may contain errors or fictional information.

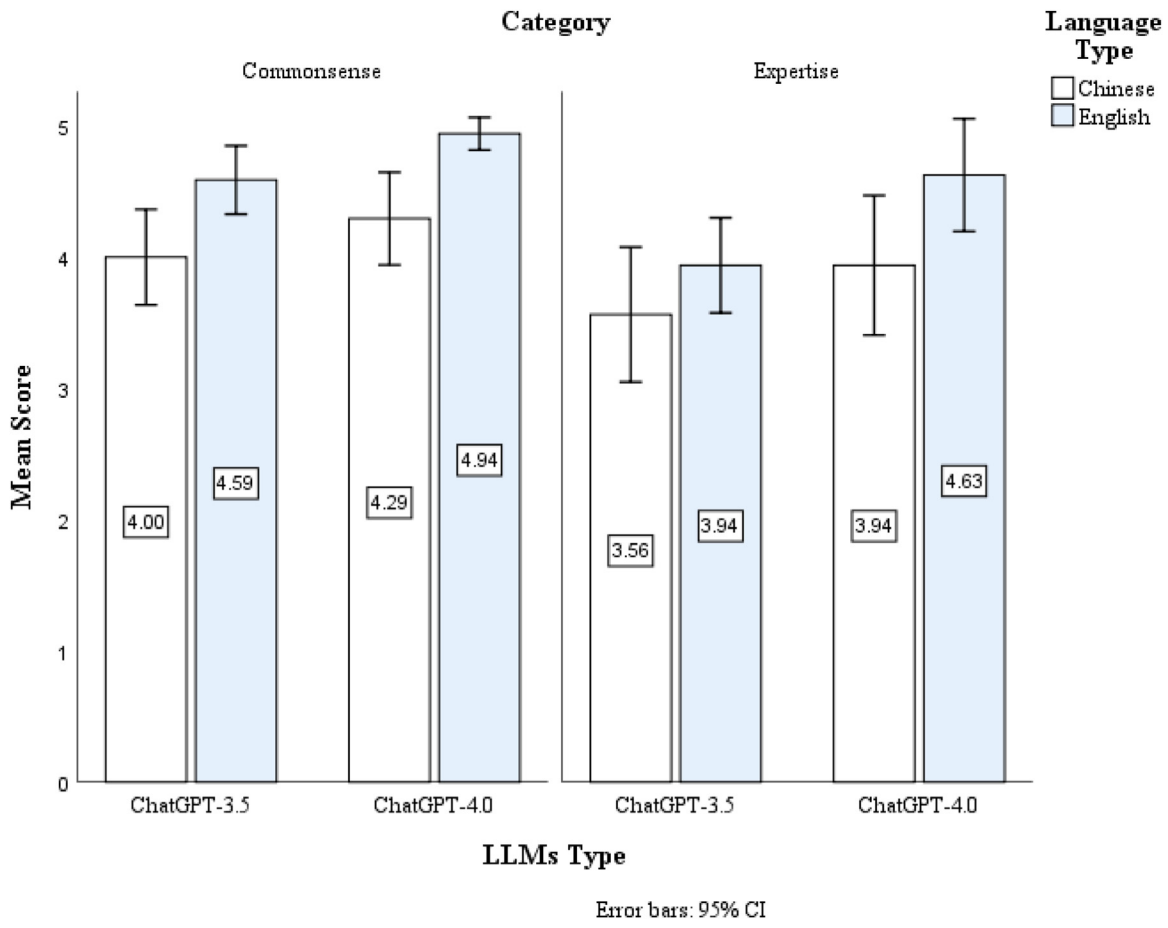

Fig. 3 – Average scores of LLM-Chatbots in different question categories.

Additionally, in the Chinese version, even ChatGPT-4 did not demonstrate satisfactory performance. This outcome may be attributed to variations in standardized medical terms and concepts across different language environments.[26] Furthermore, since LLM training data predominantly focus on English-speaking regions, language database training may lack sufficient Chinese information,[27] leading to decreased efficacy when patients are presented with questions in Chinese. Therefore, as the use of LLMs expands to non-English-speaking regions, future research should focus on reducing performance disparities across different linguistic contexts to enhance the reliability of the models.

Limitations of the study

Our study has several limitations. First, the limited number of questions for each category may not fully capture the complexity and multidimensionality inherent in real clinical settings.

Second, the subjectivity of the scoring panel cannot be overlooked. To address this issue, we selected two experts in endodontic diseases and one in periodontal diseases, each with more than a decade of clinical and academic experience, and employed a consensus-based rating approach. However, the small number of experts might not represent the wide range of opinions and criteria present in the broader clinical professional community.

Finally, the time involved in drafting manuscripts, peer review, and submission to journals must be considered. Since LLMs continuously evolve based on user feedback and updates to their training datasets, the interpretation of these findings should be contextualized within specific time periods. Future research may yield different outcomes.

Implications of the study and future

With continuous advancements in the medical field, it is believed that LLMs such as ChatGPT-4 could play a vital role as tools for diagnostic or therapeutic support in future clinical scenarios. Our study highlights the potential benefits of LLMs, specifically ChatGPT-4, in addressing patient inquiries related to oral healthcare, demonstrating the ability to provide thorough and coherent answers to a wide range of questions. However, it is important to recognize that LLMs are not specifically designed for oral healthcare and may lack the necessary medical sensitivity in certain contexts.[28] While LLMs are making progress in addressing real-world patient queries, the reliance of patients on internet-based information for health decisions suggests that inaccuracies from LLMs could lead to serious consequences.[6,7] Thus, integrating LLMs into medicine presents challenges, including high computational demands, laborious data annotation and a lack of interpretability.[29] Furthermore, data privacy remains a paramount

concern in today's digital landscape. As efforts to enhance LLMs continue, healthcare professionals and regulatory bodies must ensure that these tools are reliable, ethical, and transparent in their decision-making processes.[30] Harnessing the potential of LLMs while prioritizing data protection and privacy issues is crucial.

Conclusions

In the comparison between ChatGPT-3.5 and ChatGPT-4, significant advancements in AI-powered knowledge bases and comprehension were highlighted, demonstrating a promising future for the further improvement of LLMs in clinical applications. However, future research should focus on exploring approaches to reduce performance disparities across different linguistic contexts, thereby enhancing adaptability and effectiveness. Moreover, the potential applications of LLMs in the dental field require further exploration to refine and confirm their efficacy and reliability, which will be crucial for advancing their use in dentistry.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT to edit and proofread the manuscript for improved readability. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Date availability statement

The data that supports the findings of this study are available in the Supplemental Material of this article.

Funding

This research was supported by Chongqing Science and Health Joint Project (2021MSXM188), by Chongqing Graduate Tutor Team Construction Project (dstd201903), by Project for Youth Innovation in Future Medicine, Chongqing Medical University (W0129), by Graduate Research Innovation Project of Stomatological Hospital of Chongqing Medical University (KQY202308), and by Technological Innovation Project of Chongqing Municipal Education Commission (KJCX2020017).

Ethical approval statement

This research was conducted in adherence to the Declaration of Helsinki and ethical approval was not required for this research as no patients were involved in our study.

Conflict of interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Zhaowu Chai reports financial support was provided by Stomatological Hospital of Chongqing Medical University. Zhaowu Chai reports a relationship with Stomatological Hospital of Chongqing Medical University that includes: employment. Zhaowu Chai has patent pending to Zhaowu Chai. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this article.

CRediT authorship contribution statement

Qian Zhang: Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft. **Zhengyu Wu:** Data curation, Formal analysis, Visualization, Validation. **Jinlin Song:** Visualization, Validation, Funding acquisition. **Shuicai Luo:** Data curation, Software. **Zhaowu Chai:** Conceptualization, Supervision, Methodology, Validation, Funding acquisition, Writing – review & editing.

Acknowledgements

Zhaowu Chai was supported by Chongqing Science and Health Joint Project (2021MSXM188), by Chongqing Graduate Tutor Team Construction Project (dstd201903), by Project for Youth Innovation in Future Medicine, Chongqing Medical University (W0129); Jinlin Song was supported by Technological Innovation Project of Chongqing Municipal Education Commission (KJCX2020017). Qian Zhang was suported by Graduate Research Innovation Project of Stomatological Hospital of Chongqing Medical University (KQY202308).

Supplementary materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.identj.2024.06.022.

REFERENCES

1. Holmgren AJ, Downing NL, Tang M, Sharp C, Longhurst C, Huckman RS. Assessing the impact of the COVID-19 pandemic on clinician ambulatory electronic health record use. J Am Med Inform Assoc 2022;29(3):453–60.
2. Zulman DM, Verghese A. Virtual care, telemedicine visits, and real connection in the era of COVID-19: unforeseen opportunity in the face of adversity. JAMA 2021;325(5):437–8.
3. Praditpapha A, Mattheos N, Pisarnturakit PP, Pimkhaokham A, Subbalekha K. Dentists' stress during the COVID-19 pandemic: a repeated cross-sectional study. Int Dent J 2023;74(2):294–302.
4. Schulz PJ, Nakamoto K. The perils of misinformation: when health literacy goes awry. Nat Rev Nephrol 2022;18(3):135–6.
5. Kuehn BM. More than one-third of US individuals use the Internet to self-diagnose. JAMA 2013;309(8):756–7.
6. Wong DK, Cheung MK. Online health information seeking and eHealth literacy among patients attending a primary care clinic in Hong Kong: a cross-sectional survey. J Med Internet Res 2019;21(3):e10831.

7. Zhang D, Zhan W, Zheng C, et al. Online health information-seeking behaviors and skills of Chinese college students. BMC Public Health 2021;21(1):736.
8. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 2023;388(13):1233–9.
9. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. J Med Syst 2023;47(1):33.
10. De Angelis L, Baglivo F, Arzilli G, et al. ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. Front Public Health 2023;11:1166120.
11. Egli A. ChatGPT, GPT-4, and other large language models: the next revolution for clinical microbiology? Clin Infect Dis 2023;77(9):1322–8.
12. Chau RCW, Thu KM, Yu OY, Hsung RT, Lo ECM, Lam WYH. Performance of generative artificial intelligence in dental licensing examinations. Int Dent J. 2024;74(3):616–21.
13. Balel Y. Can ChatGPT be used in oral and maxillofacial surgery? J Stomatol Oral Maxillofac Surg 2023;124(5):101471.
14. Lim ZW, Pushpanathan K, Yew SME, et al. Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. EBioMedicine 2023;95:104770.
15. 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):108–13.
16. Suárez A, Jiménez J, Llorente de Pedro M, Andreu-Vázquez C, et al. Beyond the Scalpel: assessing ChatGPT's potential as an auxiliary intelligent virtual assistant in oral surgery. Comput Struct Biotechnol J 2024;24:46–52.
17. Duncan HF, Kirkevang LL, Peters OA, et al. Treatment of pulpal and apical disease: the European Society of Endodontology (ESE) S3-level clinical practice guideline. Int Endod J 2023;56(Suppl 3):238–95.
18. Chugal N, Assad H, Markovic D, Mallya SM. Applying the American Association of Endodontists and American Academy of Oral and Maxillofacial Radiology guidelines for cone-beam computed tomography prescription: impact on endodontic clinical decisions. J Am Dent Assoc (1939) 2024;155(1):48–58.
19. Biswas S, Logan NS, Davies LN, Sheppard AL, Wolffsohn JS. Assessing the utility of ChatGPT as an artificial intelligence-based large language model for information to answer questions on myopia. Ophthalmic Physiol Opt 2023;43(6):1562–70.
20. Stone A, Jiang ST, Stahl MC, Yang CJ, Smith RV, Mehta V. Development and interrater agreement of a novel classification system combining medical and surgical adverse event reporting. JAMA Otolaryngol Head Neck Surg 2023;149(5):424–9.
21. Raimondi R, Tzoumas N, Salisbury T, Di Simplicio S, Romano MR. Comparative analysis of large language models in the Royal College of Ophthalmologists fellowship exams. Eye (London, England) 2023;37(17):3530–3.
22. Ali R, Tang OY, Connolly ID, et al. Performance of ChatGPT, GPT-4, and google bard on a neurosurgery oral boards preparation question bank. Neurosurgery. 2023;93(5):1090–8.
23. Varshney D, Zafar A, Behera NK, Ekbal A. Knowledge graph assisted end-to-end medical dialog generation. Artif Intell Med 2023;139:102535.
24. Varshney D, Zafar A, Behera NK, Ekbal A. Knowledge grounded medical dialogue generation using augmented graphs. Sci Rep 2023;13(1):3310.
25. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature 2023;620(7972):172–80.
26. Soto-Chávez MJ, Bustos MM, Fernández-Ávila DG, Muñoz OM. Evaluation of information provided to patients by ChatGPT about chronic diseases in Spanish language. Digital Health 2024;10:20552076231224603.
27. Kao YS, Chuang WK, Yang J. Use of ChatGPT on Taiwan's examination for medical doctors. Ann Biomed Eng 2024;52(3):455–7.
28. Sezgin E. Artificial intelligence in healthcare: complementing, not replacing, doctors and healthcare providers. Digital Health 2023;9:20552076231186520.
29. Alsadhan A, Al-Anezi F, Almohanna A, et al. The opportunities and challenges of adopting ChatGPT in medical research. Front Med 2023;10:1259640.
30. Shamszare H, Choudhury A. Clinicians' Perceptions of artificial intelligence: focus on workload, risk, trust, clinical decision making, and clinical integration. Healthcare (Basel, Switzerland) 2023;11(16):2308.