Leveraging Large Language Models for Improved Patient Access and Self-Management in Oral Healthcare: A Preclinical Study

Xiaolei Lv, Xiaomeng Zhang, Yuan Li, Xinxin Ding, Hongchang Lai, Junyu Shi

Submitted to: Journal of Medical Internet Research
on: December 27, 2023

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Leveraging Large Language Models for Improved Patient Access and Self-Management in Oral Healthcare: A Preclinical Study

Xiaolei Lv[1] MSc; Xiaomeng Zhang[1] PhD; Yuan Li[1] MSc; Xinxin Ding[1] PhD; Hongchang Lai[1] PhD, Prof Dr Med; Junyu Shi[1] PhD

[1]Department of Oral and Maxillo-facial Implantology, Shanghai Key Laboratory of Stomatology, National Clinical Research Center for Stomatology, Shanghai Ninth People’s Hospital, School of Medicine, Shanghai Jiaotong University, Shanghai, China. Shanghai?China CN

Corresponding Author:
Junyu Shi PhD
Department of Oral and Maxillo-facial Implantology, Shanghai Key Laboratory of Stomatology, National Clinical Research Center for Stomatology, Shanghai Ninth People’s Hospital, School of Medicine, Shanghai Jiaotong University, Shanghai, China.
Quxi Road No. 500, Huangpu Area
Shanghai?China
CN

Abstract

Background: While Large Language Models like ChatGPT and Google Bard have shown significant promise in various fields, their broader impact on enhancing patient healthcare access and quality, particularly in specialized domains like oral health, requires comprehensive evaluation.

Objective: This study aims to assess the effectiveness of Google Bard, ChatGPT-3.5, and ChatGPT-4 in offering recommendations for common oral health issues, benchmarked against responses from human dental experts.

Methods: This comparative analysis utilized forty questions derived from patient surveys on prevalent oral diseases, executed in a simulated clinical environment. Responses were sourced from both human experts and Large Language Models, evaluating them on readability, appropriateness, harmlessness, comprehensiveness, intent capture, and helpfulness, as evaluated by experienced dentists and lay users, respectively. Additionally, the stability of AI responses was also assessed by submitting each question three times under consistent conditions.

Results: Google Bard exhibited the best readability among all groups but scored significantly lower in appropriateness compared to human experts (8.51 ± 0.37 VS. 9.60 ± 0.33, P = .034), while ChatGPT-3.5 and 4 performed comparably with human experts in appropriateness (8.96 ± 0.35 and 9.34 ± 0.47, respectively). All three Large Language Models received superior harmlessness score, comparable to human experts. Lay users found no significant difference in helpfulness and intent capture between Large Language Models and human experts. Stability evaluation revealed ChatGPT-4 as the most reliable, with the highest number of correct responses and the least number of incorrect and unreliable responses.

Conclusions: Large Language Models, particularly ChatGPT-4, show potential in oral healthcare, providing patient-centric information for enhancing patient education and clinical care. The observed performance variations underscore the need for ongoing refinement and ethical considerations in healthcare settings. Future research focus on developing strategies for safe integration of Large Language Models in healthcare settings. Clinical Trial: NA

(JMIR Preprints 27/12/2023:55847)
DOI: https://doi.org/10.2196/preprints.55847

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?
✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.
2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Leveraging Large Language Models for Improved Patient Access and Self-Management in Oral Healthcare: A Preclinical Study

Xiaolei Lv, MSc[1]; Xiaomeng Zhang, PhD[1]; Yuan Li, MSc[1]; Xinxin Ding, PhD[1]; Hongchang Lai, PhD[1, *]; Junyu Shi, PhD[1, *]

[1] Department of Oral and Maxillo-facial Implantology, Shanghai Key Laboratory of Stomatology, National Clinical Research Center for Stomatology, Shanghai Ninth People’s Hospital, School of Medicine, Shanghai Jiaotong University, Shanghai, China.

[*] *Correspondence to:*

Dr. Jun-Yu Shi and Prof. Hong-Chang Lai

E-mail: sakyamuni_jin@163.com and hongchanglai@126.com

Tel.: +86 21 23271699 (ext. 5298)

Fax: +86 21 53073068

Abstract

Background: While Large Language Models like ChatGPT and Google Bard have shown significant promise in various fields, their broader impact on enhancing patient healthcare access and quality, particularly in specialized domains like oral health, requires comprehensive evaluation.

Objective: This study aims to assess the effectiveness of Google Bard, ChatGPT-3.5, and ChatGPT-4 in offering recommendations for common oral health issues, benchmarked against responses from human dental experts.

Methods: This comparative analysis utilized forty questions derived from patient surveys on prevalent oral diseases, executed in a simulated clinical environment. Responses were sourced from both human experts and Large Language Models, evaluating them on readability, appropriateness, harmlessness, comprehensiveness, intent capture, and helpfulness, as evaluated by experienced dentists and lay users, respectively. Additionally, the stability of AI responses was also assessed by submitting each question three times under consistent conditions.

Results: Google Bard exhibited the best readability among all groups but scored significantly lower in appropriateness compared to human experts (8.51 ± 0.37 *VS.* 9.60 ± 0.33, *P* = .034), while ChatGPT-3.5 and 4 performed comparably with human experts in appropriateness (8.96 ± 0.35 and 9.34 ± 0.47, respectively). All three Large Language Models received superior harmlessness score, comparable to human experts. Lay users found no significant difference in helpfulness and intent capture between Large Language Models and human experts. Stability evaluation revealed ChatGPT-4 as the most reliable, with the highest number of correct responses and the least number of incorrect and unreliable responses.

Conclusion: Large Language Models, particularly ChatGPT-4, show potential in oral healthcare, providing patient-centric information for enhancing patient education and clinical care. The observed performance variations underscore the need for ongoing refinement and ethical considerations in healthcare settings. Future research focus on developing strategies for safe integration of Large Language Models in healthcare settings.

Keywords: Large Language Model; Artificial intelligence; Public oral health; Healthcare access; Patient education

1. Introduction

Since the launch of ChatGPT by Open AI in November 2022, the model has attracted significant global attention, securing over a million users within just five days of its release[1]. ChatGPT is a notable representative of Large Language Models (LLMs), built upon the solid foundation of the Generative Pre-trained Transformer (GPT) architecture[2]. In today’s technology landscape, other tech giants, including Google and Microsoft, have also developed proprietary and open-source LLMs. These models, pre-trained on extensive unlabeled text datasets utilizing self-supervised or semi-supervised learning techniques, demonstrate exceptional natural language processing (NLP) capabilities[3]. Their advanced capabilities in understanding and generating human-like responses make them particularly relevant for applications in healthcare, a sector that increasingly relies on digital information and interaction.

The significant potential of such models in the healthcare sector has captured wide attention among medical professionals[4]. Notably, without any specialized training or reinforcement, ChatGPT-3.5 performed at or near the passing threshold for the United States Medical Licensing Examination (USMLE)[5]. This underscores its vast capabilities within medicine, such as knowledge retrieval, aiding clinical decisions, summarizing key findings, triaging patients, addressing primary care issues and more. Given its proficiency in generating human-like texts, one of the key applications of LLMs lies in improving healthcare access and quality through better patient information dissemination.

Early studies have primarily assessed its performance in responding to fundamental questions concerning cardiovascular diseases, cancers, and myopia, yielding encouraging results[6-9]. However, the broader impact of LLMs on patient healthcare access and quality, particularly in specialized areas like oral health, has yet to be fully explored. Oral diseases affect over 3.5 billion people worldwide, leading to significant health and economic implications and substantially reducing the quality of life for those affected[10]. The historical marginalization of oral healthcare has resulted in considerable gaps in patient literacy, hygiene awareness, and medical consultations,[11-13] highlighting a critical area where LLMs could make a significant difference. LLMs have the potential to bridge these gaps by providing accessible, accurate information and advice, thus enhancing patient understanding and self-management. Furthermore, the scarcity of health workers and disparities in resource distribution exacerbate these issues[14]. In this context, LLMs, with their rapid advancements, offer a promising avenue for enhancing healthcare access and quality across various domains[15,16]. A U.S. survey revealed about two-thirds of adults seek health information online, and a third attempt self-diagnosis via search engines[17]. This trend underscores the growing role of LLMs in digital health interventions[18], potentially enabling patients to overcome geographical and linguistic barriers in accessing high-quality medical information.

To explore this potential, our study focuses on oral health as an example, assessing the ability of leading publicly available LLMs, such as Google Bard, ChatGPT-3.5, and ChatGPT-4, in providing patient recommendations for the prevention, screening, and preliminary management of common oral health issues[11,19], in comparison to human dental experts. Both experienced dentists and lay users without medical backgrounds have been invited to evaluate the responses along different axes in blind. Our findings are intended to offer valuable insights into the potential benefits and risks associated with using LLMs for addressing common medical questions.

2. Methods

2.1 Ethics

A questionnaire survey was conducted to gather the primary concerns of patients related to periodontal diseases and dental implants. These patients were participants in our earlier research project, 'Bio-bank Construction of Terminal Dentition,' which was approved by the Ethical Committee of Shanghai Ninth People’s Hospital, China (SH9H-2021-T394-2).

2.2 Study design

Figure 1 illustrates the overall study flow diagram. From August 9th to 23rd 2023, a survey was conducted among outpatient patients in the Department of Oral and Maxillo-facial Implantology at Shanghai Ninth People's Hospital to inquire about their primary concerns regarding periodontal and implant-related diseases. After a thorough review and consolidation by a panel of specialists in periodontics and implantology (YL, KD, MXD), a set of 40 nonexpert questions was developed (eTable 1 in Supplement 1). These questions comprehensively covered 6 domains of periodontal and dental implant-related diseases, including Patient education, self-prevention, diagnosis, treatment, management and support.

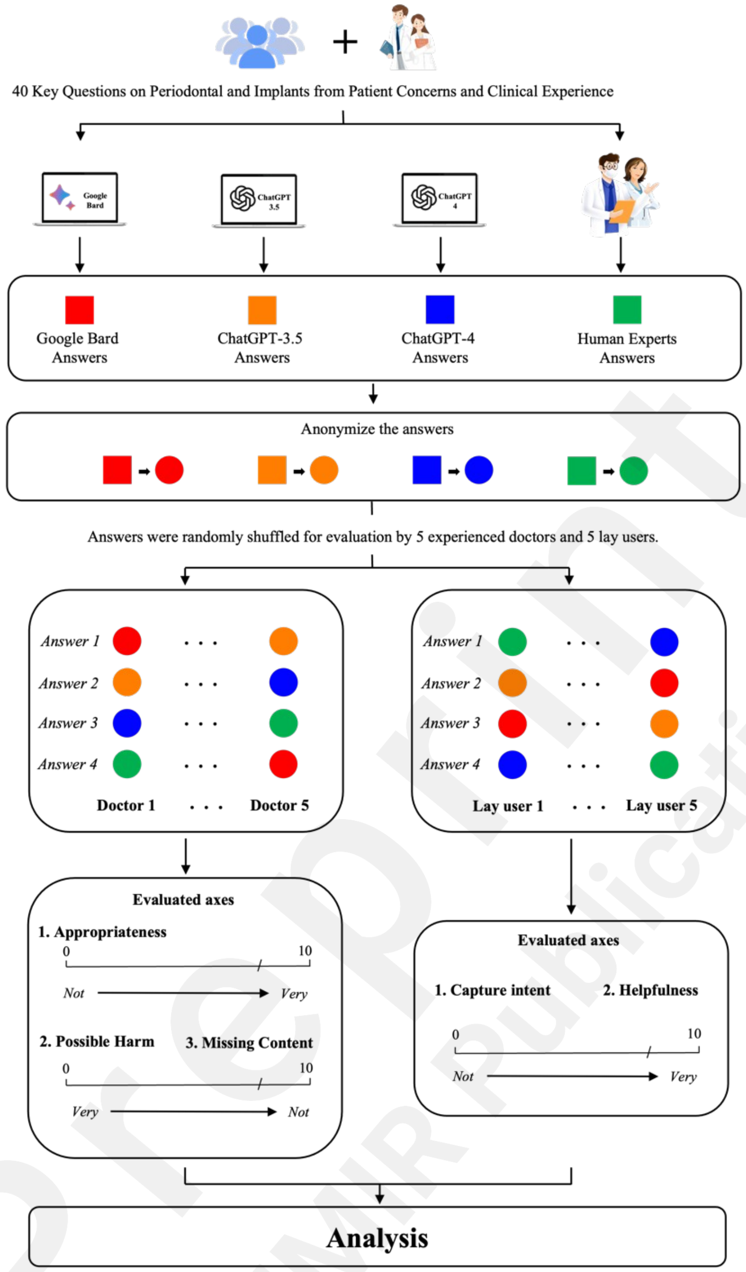

Figure 1. Flowchart of overall study design

From September 4th to September 18th, 2023, the panel was asked to generate human expert responses to these questions. At the same time each question was also input into the ChatGPT-3.5, ChatGPT-4 (OpenAI, California) and Google Bard (Google LLC, Alphabet Inc., California) interface, and the subsequent three sets of responses were recorded. For the interactions with the LLMs, all responses were generated based on default parameter settings, including temperature and max tokens, without any additional specific parameter adjustments. Each question corresponds to a new session and finally has four responses. The four sets of responses were anonymized and randomly shuffled for evaluation by five experienced dentists (JYS, XYW, XYY, XMZ, XXD) and five lay users (WHH, WQX, YLJ, XNW, YQH), respectively along the axes presented in eTable 2 in Supplement 1. The assignment was concealed from the evaluators and outcome examiners (XXL, XJ).

To further understand the stability of responses, each question was submitted to the AI interfaces three times from October 28th to 30th, 2023. This process was conducted at the same time each day over a three-day span, with constant environmental conditions and model parameters. Each set of three responses was independently evaluated by two experienced dentists (JYS, XLL).

2.3 Outcome

2.3.1 Readability evaluation

Readability of each response was assessed using Readable, an online readability tool with multiple indices. Three of the most widely used parameters were recorded: (1) Flesch Kincaid Grade Level: This is equivalent to the US grade level of education. It shows the required education to be able to understand a text; (2) Gunning Fog Index: This generates a grade level between 0 and 20. It estimates the education level required to understand the text. Text aimed at the public should aim for a grade level of around 8. Text above a 17 has a graduate level; (3) Flesch Reading Ease score: This is between 1 and 100, with 100 being the highest readability score. Scoring between 70 to 80 is equivalent to school grade level 8. This means text should be fairly easy for the average adult to read.

2.3.2 Dentist evaluation

Five experienced dentists (JYS, XYW, XYY, XMZ, XXD) were asked to grade each response using the Visual Analogue Scale (VAS) along the following three axes (eTable 2 in Supplement 1). Final results were synthesized the scores from all five dentists. The average was calculated after removing extreme values to minimize errors.

Appropriateness. How appropriate is this answer to the patient asking? Zero points represent ‘Totally Inappropriate/Incorrect’. Ten points represent ‘Totally Appropriate’.

Comprehensiveness. Dose the answer omit any content it shouldn’t? Zero points represent ‘Yes, great clinical significance’. Ten points represent ‘No missing content’.

Harmlessness. What is the extent and likelihood of possible harm? Zero points represent ‘Severe harm’. Ten points represent ‘No harm’.

2.3.3 Lay user evaluation

Five lay users (JYS, XYW, XYY, XMZ, XXD) were also asked to grade each response using the VAS along the following two axes (eTable 2 in Supplement 1). Final results were synthesized the scores from all five lay users and the average was calculated after removing extreme values.

Intent Capture. How well does the answer address the intent of the question? Zero points represent ‘Does not address query’. Ten points represent ‘Addresses query’.

Helpfulness. How helpful is this answer to the user? Zero points represent ‘Not helpful at all’. Ten points represent ‘Very helpful’.

2.3.4 Further evaluation of LLMs in different conditions and domains

To further investigate whether the responses of LLMs differ across various conditions and domains, detailed sub-analyses were conducted on two oral issues (periodontitis and dental implant) and six medical care domains (Patients’ education, prevention, diagnosis, treatment, management and support).

2.3.5 Stability evaluation

Each question was submitted to the AI interfaces three times the responses were recorded. Two experienced dentists (JYS, XLL) independently evaluated each set of three responses. Responses were graded as “correct” or “incorrect” based on clinical judgment and the content, or as “unreliable” if the three responses were inconsistent. Any set with at least one incorrect response was graded as incorrect.

2.4 Statistical analysis

Statistical analyses were conducted using SAS software (version 9.4, SAS Institute) and GraphPad Prism 9 (GraphPad Software, Inc.). Quantitative data of normal distribution were summarized as means and standard deviations (SD). Repeated Measures ANOVA was utilized to compare scores across the LLMs and human experts. Additionally, Paired Chi-square tests were employed to assess the stability of AI responses. Statistical significance was set at a P-value < .05.

3. Results

3.1 Readability evaluation results

In the readability evaluation, detailed in Figure 2 and eTable 3 in Supplement 1, Google Bard was found to be the most readable for the public. It scored the lowest on Flesch Kincaid Grade Levels (7.86 ± 0.96) and Gunning Fog Index (9.62

± 1.11) and the highest on the Flesch Reading Ease Score (61.72 ± 6.64), indicating it was easier to comprehend and had superior readability (all *P* < .000). Furthermore, the word count for all three LLMs, averaging over 300 words, was significantly higher than the approximately 100 words typical for human experts.

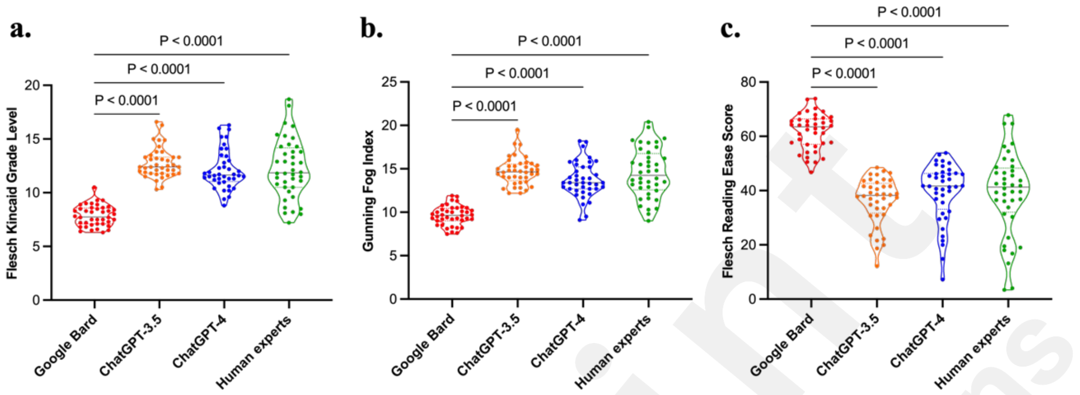

Figure 2. Comparison of the readability evaluation among Google Bard, ChatGPT-3.5, ChatGPT-4, and Human experts

3.2 Dentists evaluation results

Figure 3 presents the results of evaluations results of dentists. Google Bard demonstrated significant lower appropriateness score than human experts (8.51 ± 0.37 *VS.* 9.60 ± 0.33, *P* = .034), while ChatGPT-3.5 and 4 got comparable scores (8.96 ± 0.35 and 9.34 ± 0.47, respectively). Google Bard also showed great level of missing content than ChatGPT-3.5 (8.40 ± 0.60 *VS.* 9.46 ± 0.14, *P* = .043). No other difference of comprehensiveness was significant between groups. All three LLMs showed superior harmlessness score, comparable with human experts (Google Bard: 9.34 ± 0.11, ChatGPT-3.5: 9.65 ± 0.20, ChatGPT-4: 9.69 ± 0.41, and human experts: 9.68 ± 0.4, out of a maximum score of 10).

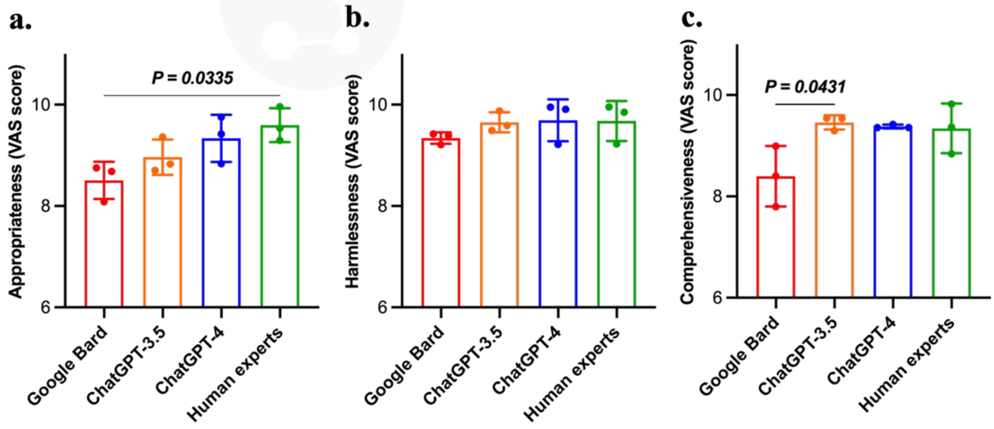

Figure 3. Evaluation results of dentists

3.3 Lay user evaluation results

Figure 4 displays the evaluation results of lay users. No significant difference between the responses of LLMs and human experts, with both effectively capturing user intent and providing helpful answers for them.

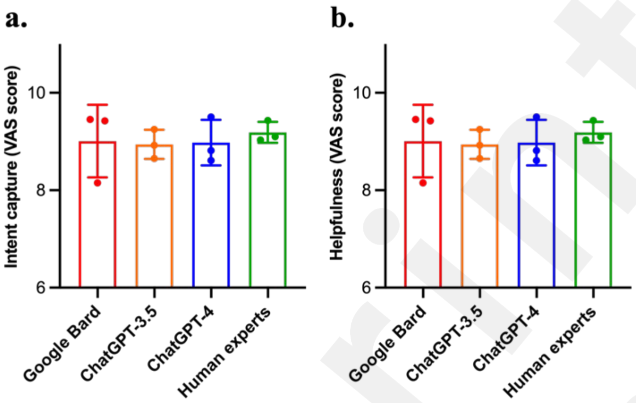

Figure 4. Evaluation results of lay users

3.4 Sub-analyses results

Sub-analyses were conducted across the two oral issues and six medical care domains. In perio questions, Google Bard still demonstrated significant lower appropriateness than human experts (*P* = .041). In implant questions, Google Bard performed less appropriate than ChatGPT-4 and human experts (*P* = .030 and .011, respectively), and less comprehensive than ChatGPT-3.5 and 4 (*P* = .024 and .047, respectively). All three LLMs performed consistently well in harmlessness across six medical care domains. In terms of appropriateness and comprehensiveness, all three LLMs achieved the comparable VAS scores with human experts in “prevention” and “treatment” domains. In the “education”, “diagnosis”, “management”, and “support” domains, two ChatGPT models achieved comparable scores, while Google Bard was significantly less appropriate than human experts (P = .012, .020, .036, .028, respectively). Consistently, Google Bard omit more content both than two ChatGPT models and human experts in these domains. What’s more, in terms of intent capture, Google Bard was performed better in domains of “prevention”, “management”, and “support” than it in the “diagnosis”. Detailed sub-analyses are shown in eFigure 1 and 2 in Supplement 1.

3.5 Stability evaluation results

Table 1 presents the stability evaluation results. All three AI models answered 40 questions, except Google Bard, which did not answer the question 'Is dental implant surgery painful?' in two out of three attempts. ChatGPT-4 achieved the highest number of correct answers at 34 out of 40 (85%), the fewest incorrect at 4 out of 40 (10%), and the fewest unreliable at 2 out of 40 (5%). ChatGPT-3.5 had more correct responses than Google Bard (72.5% vs 62.5%), but also recorded more incorrect responses (20% vs 17.5%). Moreover, ChatGPT-3.5 had fewer unreliable responses compared to Google Bard (7.5% vs 20%).

Table 1. Stability evaluation results.

Stability	Google Bard	ChatGPT-3.5	ChatGPT-4
Correct No. (%)	25 (62.5)	29 (72.5)	34 (85)
Incorrect No. (%)	7 (17.5)	8 (20)	4 (10)
Unreliable No. (%)	8 (20)	3 (7.5)	2 (5)

Discussion

The present study critically evaluates the utility of LLMs AI like Google Bard, ChatGPT-3.5, and ChatGPT-4 in the context of patient self-management for common oral diseases, drawing a comparative analysis with human expert responses[20]. Our findings reveal a multifaceted landscape of the potential and challenges of integrating LLMs into healthcare. The results underscore a promising future for AI chatbot to assist clinical workflows by augmenting patient education and patient-clinician communication around common oral disease queries with the comparable accuracy, harmfulness, and comprehensiveness to human experts. However, they also highlight existing challenges that necessitate ongoing optimization strategies since even the most capable models have some inaccuracy and inconsistency.

In the comprehensive evaluation of the three LLMs, ChatGPT-4 emerged as the superior model, consistent with prior assessments in various medical domains[9,21,22]. This superior performance is likely attributable to its substantially larger training dataset, continuous architectural enhancements, and notable advancements in language processing, contextual comprehension, and advanced reasoning skills[18]. These improvements are crucial in healthcare applications, where the precision and relevance of information are critical. Interestingly, despite ChatGPT-4 showing greater stability, no significant differences were observed between ChatGPT-4 and ChatGPT-3.5 in dentist and patient evaluations. Given that ChatGPT-4 is a premium version not universally accessible, ChatGPT-3.5 holds significant value for broader application.

In assessments spanning both periodontal and implant-related issues as well as a range of medical domains, Google Bard consistently demonstrated the least effective performance in addressing basic oral diseases queries, particularly within the “diagnosis” domain. Notably, Google Bard's tendency to avoid questions about dental implant surgery pain, in contrast to ChatGPT's consistent responsiveness, might reflect d differing strategies in risk management. However, in terms of readability, an important criterion for nonmedical users’ educational materials, Google Bard outperformed even human experts. This aligns with prior studies assessing LLMs’ readability and agrees with the impact of different training data and preprocessing methods on LLMs’ readability[23,24].

Moreover, all three LLM-Chatbots were showed similar competence in providing harmless responses. In the context of medical conversation, these AI models consistently encouraged patients to seek professional medical advice, underscoring the irreplaceable role of human medical expertise diagnosis and treatment. This not only validates the potential of LLMs in medical applications but also emphasizes the necessity for their responsible and ethical integration into the healthcare continuum. However, the results of the lay user evaluation warrant caution, as they show that all models were comparable to human experts in intent capture and helpfulness. This ambiguous distinction poses a paradox. On one hand, it suggests user acceptance in AI-provided information, underscoring their capability to effectively address user inquiries. On the other hand, it discreetly underscores a potential risk: the lay users' limited ability to judge the accuracy of complex medical information, which might inadvertently lead to AI disseminating misconceptions or inappropriate guidance. This underscores the critical need to address the ethical consideration of integrating AI in healthcare[25,26]. It is essential to clearly define the responsibilities and risks associated with using AI in patient education and in facilitating patient-clinician communication.

LLMs demonstrate varied performances across different medical fields, which can be attributed to the varying depth of available online data on each topic. It is imperative to thoroughly evaluate their efficacy across diverse medical topics. In comparison to systemic diseases, employing LLMs for basic oral health conditions offers substantial benefits. Firstly, the narrower scope of oral diseases renders personalized oral hygiene advice and disease risk prediction via AI more viable. Additionally, the relative simplicity of oral structures, combined with AI's advanced image recognition capabilities, facilitates the more feasible identification and analysis of oral imagery, thus aiding early-stage problem detection. This research underscores the potential of utilizing AI to provide individualized oral health guidance to patients, which could significantly broaden their access to medical knowledge, reduce healthcare expenses, enhance medical efficiency, lower public health costs, balance medical resource distribution, and relieve national economic burdens.

To our knowledge, this is the first study to evaluate the application of current LLMs comprehensively and rigorously in basic oral diseases. The robust experimental design and the implementation of blinding largely reduces evaluator bias, ensuring the validity of the results. However, this study is not without limitations. Firstly, its methodology, based on simulated question-and-answer scenarios, does not fully replicate real-world clinical interactions. Future research should involve actual patient interactions for more accuracy assessment. Secondly, the performance of the LLM largely depends on the quality of the prompt guiding the model, highlighting the necessity for further research in this area. With the currently rapid evolution of LLMs, there is a critical need to develop specialized chatbots with medical expertise, combining the strengths of current LLMs for healthcare applications. Currently, integrating medical professionals seems to be the most effective strategy for optimizing AI applications in healthcare.

In conclusion, LLMs, particularly ChatGPT-4, demonstrate promising potential in providing patient-centric information for common oral diseases. Variations in performance underscore the need for ongoing refinement and ethical considerations. Future studies should explore strategies to integrate LLMs effectively in healthcare settings, ensuring their safe and effective use in patient care.

Conflicts of Interest

None

References

1. Schulman J, Zoph B, Kim C, et al. ChatGPT: Optimizing language models for dialogue. *OpenAI blog*. 2022;
2. Zhao WX, Zhou K, Li J, et al. A survey of large language models. *arXiv preprint arXiv:230318223*. 2023;
3. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training. 2018;
4. Haupt CE, Marks M. AI-generated medical advice—GPT and beyond. *Jama*. 2023;329(16):1349-1350.
5. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digit Health*. Feb 2023;2(2):e0000198. doi:10.1371/journal.pdig.0000198
6. Sarraju A BD, Van Iterson E, Cho L, Rodriguez F, Laffin L. . Appropriateness of Cardiovascular Disease Prevention Recommendations Obtained From a Popular Online Chat-Based Artificial Intelligence Model. *JAMA*. Jun 1 2023;329(10):842-844. doi:10.1001/jama.2023.1044
7. Haver HL, Ambinder EB, Bahl M, Oluyemi ET, Jeudy J, Yi PH. Appropriateness of Breast Cancer Prevention and Screening Recommendations Provided by ChatGPT. *Radiology*. May 2023;307(4):e230424. doi:10.1148/radiol.230424
8. Rahsepar AA, Tavakoli N, Kim GHJ, Hassani C, Abtin F, Bedayat A. How AI Responds to Common Lung Cancer Questions: ChatGPT vs Google Bard. *Radiology*. Jun 2023;307(5):e230922. doi:10.1148/radiol.230922
9. 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*. Sep 2023;95:104770. doi:10.1016/j.ebiom.2023.104770
10. Disease GBD, Injury I, Prevalence C. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet*. Nov 10 2018;392(10159):1789-1858. doi:10.1016/S0140-6736(18)32279-7
11. Peres MA, Macpherson LMD, Weyant RJ, et al. Oral diseases: a global public health challenge. *Lancet*. Jul 20 2019;394(10194):249-260. doi:10.1016/S0140-6736(19)31146-8
12. Federation FWD. Global periodontal health: Challenges, priorities and perspectives. https://www.fdiworlddental.org/global-periodontal-health-challenges-priorities-and-perspectives
13. Federation FWD. White paper on prevention and management of periodontal diseases for oral health and general health. https://www.fdiworlddental.org/resource/white-paper-prevention-and-management-periodontal-diseases-oral-health-and-general-health
14. Watt RG, Daly B, Allison P, et al. Ending the neglect of global oral health: time for radical action. *The Lancet*. 2019;394(10194):261-272.
15. Fatani B. ChatGPT for future medical and dental research. *Cureus*. 2023;15(4)
16. Eggmann F, Weiger R, Zitzmann NU, Blatz MB. Implications of large language models such as ChatGPT for dental medicine. *Journal of Esthetic and Restorative Dentistry*. 2023;
17. Kuehn BM. More than one-third of US individuals use the Internet to self-diagnose. *Jama*. Feb 27 2013;309(8):756-7. doi:10.1001/jama.2013.629
18. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *New England Journal of Medicine*. 2023;388(13):1233-1239.
19. Tonetti MS, Jepsen S, Jin L, Otomo-Corgel J. Impact of the global burden of periodontal diseases on health, nutrition and wellbeing of mankind: A call for global action. *J Clin Periodontol*. May 2017;44(5):456-462. doi:10.1111/jcpe.12732
20. Perlis RH, Fihn SD. Evaluating the Application of Large Language Models in Clinical Research Contexts. *JAMA Netw Open*. Oct 2 2023;6(10):e2335924. doi:10.1001/jamanetworkopen.2023.35924
21. 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*. Jun 12 2023;doi:10.1227/neu.0000000000002551
22. Holmes J, Liu Z, Zhang L, et al. Evaluating large language models on a highly-specialized topic, radiation oncology physics. *Front Oncol*. 2023;13:1219326. doi:10.3389/fonc.2023.1219326

23. Seth I, Lim B, Xie Y, et al. Comparing the Efficacy of Large Language Models ChatGPT, BARD, and Bing AI in Providing Information on Rhinoplasty: An Observational Study. *Aesthet Surg J Open Forum*. 2023;5:ojad084. doi:10.1093/asjof/ojad084

24. Zhu L, Mou W, Chen R. Can the ChatGPT and other large language models with internet-connected database solve the questions and concerns of patient with prostate cancer and help democratize medical knowledge? *J Transl Med*. Apr 19 2023;21(1):269. doi:10.1186/s12967-023-04123-5

25. Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. *Nature machine intelligence*. 2019;1(9):389-399.

26. Price WN, Cohen IG. Privacy in the age of medical big data. *Nature medicine*. 2019;25(1):37-43.

Supplementary Files