Assessing the Accuracy of AI Models in Orthodontic Knowledge: A Comparative Study Between ChatGPT-4 and Google Bard

Sadia Naureen and Huma Ghazanfar Kiani

Department of Orthodontics, Rawal Institute of Health Sciences, Islamabad, Pakistan

ABSTRACT

Objective: To compare the knowledge accuracy of ChatGPT-4 and Google Bard in response to knowledge-based questions related to orthodontic diagnosis and treatment modalities.

Study Design: Cross-sectional comparative study.

Place and Duration of the Study: Department of Orthodontics, Rawal Institute of Health Sciences, Islamabad, Pakistan, from June 23rd to August 30th 2023.

Methodology: A comprehensive content analysis was designed based on a mini implant-assisted rapid palatal expansion (MARPE), clear aligners (CA), and cone beam computed tomography (CBCT), involving 30 questions for each category (total = 90) derived from recent review articles. Questions were prepared and presented to two large language models (LLMs): Google Bard and Chat-GPT-4. Two independent raters evaluated the accuracy of the responses using a scoring system ranging from one to five, by comparing the answers to a standard key. Statistical analyses, including the paired sample t-test, were used to assess the performance of the two language models.

Results: GPT-4 demonstrated superior performance, outperforming Google Bard significantly in the MARPE, CBCT, and CA categories, and achieved a higher mean score. A p-value was found to be ($p = 0.001$) for MARPE and CBCT, while it was ($p = 0.013$) for CA. Overall, GPT-4 achieved a total score of 92.6%, surpassing Google Bard's which was 72%.

Conclusion: GPT-4 is more efficient than Google Bard in providing accurate and up-to-date information regarding recent trends in orthodontic treatment modalities.

Key Words: *Aligners, Cone beam computed tomography, ChatGPT-4, Google Bard, Mini implant-assisted rapid palatal expansion.*

How to cite this article: Naureen S, Kiani HG. Assessing the Accuracy of AI Models in Orthodontic Knowledge: A Comparative Study Between ChatGPT-4 and Google Bard. *J Coll Physicians Surg Pak* 2024; **34(07)**:761-766.

INTRODUCTION

The advancing field of artificial intelligence (AI) has seen a surge in popularity, notably with the advent of widely used chatbots such as Google Bard and ChatGPT.[1] These sophisticated conversational agents, also labeled as large language models (LLMs) aim to furnish users with precise and current information across diverse domains, including complex medical problems, the interpretation of radiology reports, and the composition of scientific articles.[2] Their efficacy in addressing various tasks, from diagnosing diseases to generating medical examination queries, has been subjected to varying degrees of scrutiny.[3]

Correspondence to: Dr. Sadia Naureen, Department of Orthodontics, Rawal Institute of Health Sciences, Islamabad, Pakistan
E-mail: drsaadis12@gmail.com

Received: January 08, 2024; Revised: April 12, 2024; Accepted: July 01, 2024
DOI: https://doi.org/10.29271/jcpsp.2024.07.761

A recent study revealed that ChatGPT-3.5 demonstrated superior accuracy compared to its counterparts. However, neither GPT3.5 nor Google Bard exhibited flawless responses to all queries with absolute consistency.[4,5] They cannot give a reason creatively, comprehend emotions, or exercise moral judegment.[6] The advent of ChatGPT-4 in March 2023 marked a pivotal stride in Open AI's pursuit of scaling-up deep learning. Functioning as a large multi-modal model, accepting both image and text inputs while producing text outputs, GPT-4, falling short of human capabilities in certain real-world scenarios, showcases human-level performance across professional and academic benchmarks.[7] This latest iteration of ChatGPT is purported to possess enhanced problem-solving capabilities and a more extensive knowledge base.[8]

Within the realm of dentistry, revolutionary strides in orthodontic treatment strategies have reshaped how oral healthcare professionals address issues such as tooth malalignment and bite problems.[9] AI stands as a pivotal asset throughout the orthodontic workflow, serving as a decision-making aid and facilitating the development of more streamlined treatment approaches. AI applications are being used in dental diagnostics, cephalometric evaluation, skeletal age determination,

temporomandibular joint (TMJ) evaluation, decision-making, and patient telemonitoring.[10] The integration of AI in orthodontics not only curtails costs but also expedites the diagnosis and treatment planning processes, potentially diminishing the reliance on manpower.[11]

According to a recent study, both ChatGPT 3.5 and Google Bard generated responses were rated with a high level of accuracy and completeness to the posed general orthodontic questions by the patients.[12] Similarly, Tanaka *et al.* assessed the reliability of ChatGPT 3.5 in answering orthodontic questions related to mini implants, clear aligners (CA), and digital imaging.[13] Despite these advancements, ChatGPT-4 is recently introduced and the precision of information about recent developments in orthodontic treatment strategies, as provided by GPT-4, has yet to undergo comprehensive evaluation. Consequently, the objective of this study was to conduct an exhaustive content analysis, scrutinising the responses of GPT-4 and Google Bard to the questions linked with mini implant-assisted rapid palatal expansion (MARPE), CA, and cone beam computed tomography (CBCT) in orthodontics.

As internet access and smartphone usage continue to rise, the percentage of patients relying on AI-based platforms to obtain health-related information is increasing. Researchers claim that health information provided by AI tools can be false and should be used with caution.[14] This study has the potential to guide orthodontic professionals, students, and teachers in understanding and relying on the capabilities of these AI tools in providing accurate and up-to-date information, specifically in the field of orthodontics. The objective of this study was to compare the knowledge accuracy of ChatGPT-4 and Google Bard in response to knowledge-based questions related to orthodontic diagnosis and treatment modalities.

METHODOLOGY

This cross-sectional study was initiated within the Department of Orthodontics at Rawal Institute of Health Sciences, Islamabad, Pakistan, from June 23rd to August 30th 2023. The Ethics Committee granted approval for the study. Initially, a pilot study was executed, during which, two reviewers independently scored ten questions from each category to calculate Cronbach's alpha for both Bard and GPT-4. The Cronbach's alpha for questionnaire reliability was 0.864 for Bard and 0.896 for GPT-4, both of which are deemed acceptable. The inter-rater score agreement with a 95% confidence interval was also determined for Bard and GPT-4, separately. The intraclass correlation coefficients were 0.760 for Bard and 0.811 for GPT-4.

A thorough content analysis was conducted using a total of 90 questions related to the most recent orthodontic diagnosis and treatment trends, focusing on MARPE, CA, and CBCT. A comprehensive online survey of recent meta-analyses and systematic review articles on these subjects was conducted using Google Scholar, Web of Science, and Pub Med search engines. A total of 20 recent review articles were collected for each category, and a pair of orthodontists, who were also authors, developed a questionnaire cantered on MARPE, CA, and CBCT. The orthodontists followed the procedures of prompt engineering while formulating the questions. The questionnaire initially comprised 50 questions on appliance design, activation, mechanics, treatment protocol, and recent advancements pertaining to a specific diagnosis and treatment approach. The researchers excluded irrelevant questions from the study through electronic randomisation, 30 pertinent questions were selected for each section of MARPE, CA, and CBCT. The answers to the questions were derived from the review articles. The textbook "Orthodontics: Current Principles and Techniques" by Lee W Graber (7th edition) was consulted to address any inadequacies or uncertainties, and a reference key was created to score the responses to each question. The questionnaire served as a guide, and each question was posed as a prompt to two AI tools, Google Bard Experiment (https://bard.google.com) and ChatGPT-4 Research version (https://chat.openi.com), with the initial response from both LLMs being considered as final. To evaluate the accuracy of the LLMs' responses, two independent raters with postgraduate qualifications in orthodontics were recruited. These raters assessed the accuracy of the LLM-generated answers based on the predesigned key. The scoring scheme was based on a scale, ranging from one to five. The average score of the two raters was considered the final. The detailed scoring method was as follows; 5 - highly accurate: The answer matched the key thoroughly; 4 - moderately accurate: The answer was mostly accurate with only minor differences from the key; 3 - somewhat accurate: The answer contained moderate inaccuracies that did not match the key; 2 - slightly accurate: The answer had considerable inaccuracies and did not match the key, and 1 - inaccurate: The answer was incorrect and did not match the key at all.

The scores obtained for each question in all three sections (MARPE + CA + CBCT) were summed up for both LLMs, and the percentage was calculated (Figure 1). The passing score was set as ≥80%.

IBM SPSS Statistics for Windows version 23 was used for the data analysis. The data were expressed as numbers, means, standard deviations, frequencies, and percentages. The results of the evaluators' scores were tabulated and compared using a paired sample t-test. All statistical analyses were performed at a significance level of $p < 0.05$.

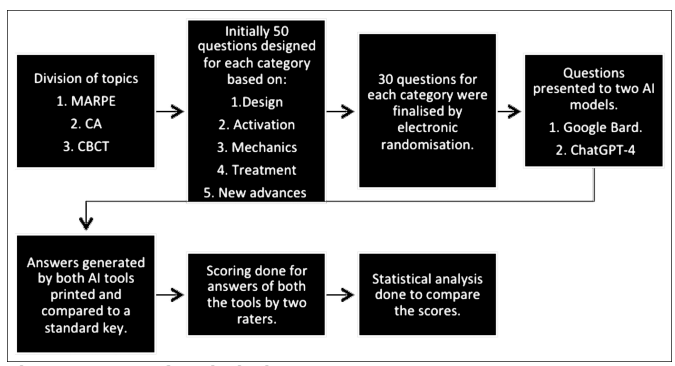

Figure 1: Steps of study design.

Table I: Frequency of performance between Google Bard and ChatGPT-4.

Question category	Number of questions	Google Bard performance			GPT-4 performance		
		Frequency of questions	Score	Percentage	Frequency of questions	Score	Percentage
MARPE	30	9 = Somewhat accurate	3	30%	15 = Moderately accurate	4	50%
		21 = Moderately accurate	4	70%	15 = Highly accurate	5	50%
Clear aligners	30	3 = Slightly accurate	2	10%	3 = Somewhat accurate	2	10%
		12 = Somewhat accurate	3	40%	6 = Moderately accurate	4	20%
		15 = Highly accurate	5	50%	21 = Highly accurate	5	70%
CBCT	30	6 = Slightly accurate	2	20%	3 = Moderately accurate	4	10%
		15 = Somewhat accurate	3	50%	27 = Highly accurate	5	90%
		6 = Moderately accurate	4	20%			
		3 = Highly accurate	5	10%			

Table II: Paired sample t-test for questionnaire categories.

Questionnaire	Paired differences					t	df	Sig. (2 tailed) p-value
	Mean	Standard deviation	Standard error mean	95% confidence interval of the difference				
				Lower	Upper			
Pair 1 = MARPE Bard-GPT-4	-0.800	0.761	0.139	-1.084	-0.516	-5.757	29	0.001
Pair 2 = Aligner Bard-GPT-4	-0.700	1.442	0.263	-1.238	-0.162	-2.659	29	0.013
Pair 3 = CBCT Bard-GPT-4	-1.022	1.022	0.187	-2.082	-1.318	-9.109	29	0.001

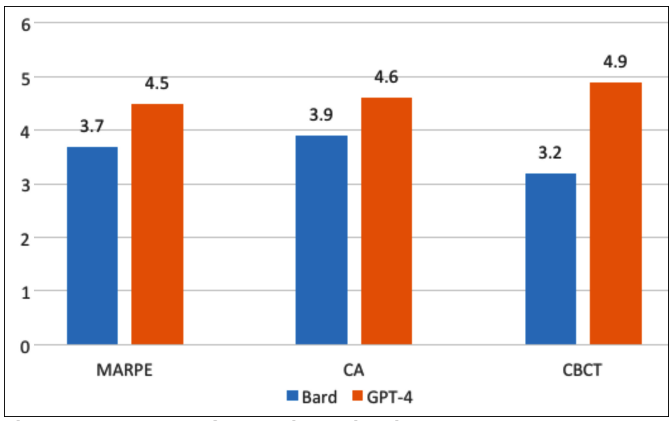

Figure 2: Mean score for Google Bard and GPT-4.

RESULTS

A total of 90 questions (30 in each category) were analysed by two raters. Figure 2 shows the mean scores obtained by Bard and GPT-4 for MARPE, CA, and CBCT. The answers given by GPT-4 were more accurate than those given by Bard. Table I shows the frequency, percentage, and score for the questions in each category. ChatGPT-4 showed a better response in all three categories: MARPE, CA, and CBCT. A paired sample t-test (Table II) was used to compare Google Bard and GPT-4 scores. The p-values for all categories were statistically significant.

In summary, GPT-4 generally showed better performance than Bard in all three categories: MARPE, CBCT, and CA. The total score of the two LLMs was significantly different, with a higher score attained by GPT-4 (92.6%) than Google Bard (72%), indicating that GPT-4 passed the questionnaire with outstanding performance.

DISCUSSION

The results of this study underscore the varying proficiency levels exhibited by the two language models when tasked with simultaneously answering identical questions. Notably, GPT-4 outperformed Bard in assessing the knowledge of orthodontic diagnosis and treatment strategies. Figure 2 illustrates a higher mean score for GPT-4 than for Bard, emphasizing the importance of choosing the most suitable language model for specific tasks, such as staying updated on orthodontic knowledge.

LLMs are continually evolving. GPT-4 launched in March 2023, showing the ability to provide information on the latest literature in a given field. This advancement opens avenues for specialsing in personalised tutoring, homework help, concept learning, standardised test preparation, discussion and collaboration, and mental health support.[1] As described in Table I, both the LLMs did not answer any question inaccurately, however, ChatGPT-4 performed better. These results are comparable to those of a recent study, but that study used ChatGPT-3.5 to generate quality answers related to clear aligners, temporary anchorage devices, and digital imaging within the context of interest of orthodontics.[13]

In this study, the answers given by GPT-4 were higher in accuracy than those given by Bard. The authors adopted a standardised method for presenting a single prompt for both LLMs. This was done to control the bias between the answers of the two models. The first answer to each prompt was considered the final answer in this study. It has been reported that Chat-GPT may potentially provide different and faster responses when asked the same question again or at different time points, whereas Google Bard generates three versions or drafts of each response.[12] Consequently, all questions were posed only once, and the initial response was selected for further evaluation. There is also a crucial need for users to verify every single response from Chat-GPT with a qualified healthcare professional, as the model's answers are generated on the basis of patterns on the data it was trained on and may not be accurate or safe.[14] However, research shows that the prompt should be specific and should explain the context of the question. The results are even better if the user is proficient in prompt engineering and follows its steps accurately.[14] Furthermore when a prompt takes the form of a question or request with a definitive answer, potentially derived from a documented source on the internet or through straightforward logical or mathematical computation, the responses generated by GPT-4 exhibit a high degree of accuracy.[15] However, false responses (hallucinations) by LLMs occur when the user enters prompts that have no single correct answer.[14]

Notably, neither Bard nor ChatGPT-4 have undergone specialised training for healthcare or medical applications, as their training objectives primarily revolved around achieving broad cognitive capability. The main difference between Google Bard and ChatGPT is that the former is trained on a dataset that includes text from the internet, while the latter is trained on a dataset that includes text from books and articles. This means that Google Bard is better equipped to provide current event-related information, while ChatGPT is more likely to offer accurate responses to factual queries.[15] This study revealed that both Google Bard and ChatGPT provided incorrect and fake references in all categories, and their effectiveness is limited due to their inability to critically analyse research findings and lack of scientific precision and reliability.

Upon further analysis, the authors found that GPT-4 outperformed Google Bard in overall performance. This result is consistent with previous studies by Ali *et al.*,[16] who observed that ChatGPT-3.5 achieved a higher percentage of accurate responses in neurosurgery oral board examinations than Google Bard (62.4% *vs*. 44.2%). GPT-4 demonstrated superior performance in all sleep medicine exam categories and achieved a higher overall score of 68.1% when compared against both GPT-3.5 (46.8%) and Google Bard (45.5%).[17] However, very few studies have explored the efficiency of GPT-4 owing to its recent launch. Gilson *et al.* observed that ChatGPT marks a significant improvement in natural language processing models on the tasks of medical question answering.[18] Toyama *et al.* found that ChatGPT plus based on GPT-4 scored 65% when answering the Japanese questions, outperforming ChatGPT-3.5 and Google Bard.[19] This highlights the potential of using LLMs to address advanced clinical questions in the field of radiology in Japan.[19] Naureen *et al.* claims that in a country like Pakistan, there is a strong need to improve knowledge and introduce a positive attitude towards the use of AI tools such as ChatGPT in dental students.[20] The research related to assessing the efficiency of AI tools is useless if students are not motivated to use them for the study purposes. In another local study, Husain *et al.* evaluated the potential of ChatGPT to help students in their assessments *via* MCQs at different levels of cognition using different subjects of internal medicine.[21] It solved C2 MCQ's by 80% but scored 69% and 54% in C1 and C3 categories, respectively.[21] Conversely, Huh's investigation revealed that ChatGPT's proficiency in parasitology falls short when juxtaposed with that of a Korean student.[22] Likewise, the research conducted by Juhi *et al*. indicates that ChatGPT exhibits only partial reliability in forecasting and elucidating drug-drug interactions within the realm of pharmacology.[23] The present study results affirm the notion that in clinical settings employing language models, it is imperative for the users to assess the originality and relevance of the model's response about a certain topic, disregarding the apparent level of confidence expressed.[24]

Even though GPT-4 gave good and strong answers about the three studied topics, still orthodontics requires more precise answers because ChatGPT includes both science and false information found in advertisements, social media, and websites.[13]

The strength of this investigation lies in the significant consensus observed among the evaluators (ICC 0.865 for Bard and 0.869 for GPT-4), which is potentially attributed to the provision of a precisely defined assessment key. This ensured a uniform scoring method for each question. Evaluators' familiarity and proficiency in evaluating orthodontic queries might have played a role in fostering more harmonised judgements. The study acknowledges limitations, including a subjective bias in the scoring approach, despite preparing the answer key beforehand. Additionally, Google Bard, which is based on PaLM2, lacks fine-tuning for medical purposes, unlike Med-Palm 2. The authors anticipate that GPT-4's continued evolution and potential improvements will contribute to the field of dentistry, specifically orthodontics, and adapt to the rapid evolution of AI in dental healthcare.

CONCLUSION

The comprehensive evaluation demonstrated that GPT-4 surpassed Google Bard in all three domains: MARPE, CA, and

CBCT. This suggested that the majority of the scores achieved by GPT-4 were categorised as moderate to highly accurate. With its improved problem-solving abilities and broadened knowledge base, GPT-4 has shown to be a dependable source of information in orthodontics. As AI advances, it holds potential as a valuable asset in orthodontics. However, it is crucial to recognise the intricate nature of AI capabilities and continually assess their performance in specific fields.

ETHICAL APPROVAL:
The Ethics Committee of the Rawal Institute of Health Sciences approved this project. Applicable ethical guidelines were followed throughout the study.

PATIENTS' CONSENT:
Patients' consent was not needed as this was a computer-based study.

COMPETING INTEREST:
The authors declared no conflict of interest.

AUTHORS' CONTRIBUTION:
SN: Conception of the study and data collection.
HGK: Data analysis and discussion.
Both authors approved the final version of the manuscript to be published.

REFERENCES

1. Labadze L, Grigolia M, Machaidze L. Role of AI chatbots in education: Systematic literature review. *Int J Educ Technol High Educ* 2023; **20**:56. doi:10.1186/s41239-023-00426-1.
2. Elkassem AA, Smith AD. Potential use cases for ChatGPT in radiology reporting. *Am J Roentgenol* 2023; **221(3)**:373-6. doi: 10.2214/AJR.23.29198.
3. Agarwal M, Sharma P, Goswami A. Analysing the applicability of ChatGPT, Bard, and bing to generate reasoning-based multiple-choice questions in medical physiology. *Cureus* 2023; **15(6)**:e40977. doi: 10.7759/cureus.40977.
4. Kumari A, Kumari A, Singh A, Singh SK, Juhi A, Dhanvijay AKD, *et al*. Large language models in hematology case solving: A comparative study of ChatGPT-3.5, Google Bard, and Microsoft Bing. *Cureus* 2023; **15(8)**:e43861. doi: 10.7759/cureus.43861.
5. 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* 2023; **307(5)**: e230922. doi: 10.1148/radiol.230922.
6. Jiang L, Wu Z, Xu X, Zhan Y, Jin X, Wang L, *et al*. Opportunities and challenges of artificial intelligence in the medical field: Current application, emerging problems, and problem-solving strategies. *J Int Med Res* 2021; **49(3)**: 3000 605211000157. doi: 10.1177/03000605211000157.
7. Kataoka Y, So R. Benefits, limits, and risks of GPT-4 as an AI Chatbot for medicine. *N Engl J Med* 2023; **388(25)**:2399. doi: 10.1056/NEJMc2305286.
8. Elkhatat AM. Evaluating the authenticity of ChatGPT responses: A study on text-matching capabilities. *Int J Educ Integr* 2023; **19**:15. doi:10.1007/s40979-023-00137-0.
9. Khanagar SB, Al-Ehaideb A, Vishwanathaiah S, Maganur PC, Patil S, Naik S, *et al*. Scope and performance of artificial intelligence technology in orthodontic diagnosis, treatment planning, and clinical decision-making - A systematic review. *J Dent Sci* 2021; **16(1)**:482-92. doi: 10.1016/j.jds.2020.05.022.
10. Kazimierczak N, Kazimierczak W, Serafin Z, Nowicki P, Nozewski J, Janiszewska-Olszowska J. AI in orthodontics: Revolutionizing diagnostics and treatment planning-A comprehensive review. *J Clin Med* 2024; **13(2)**:344. doi: 10.3390/jcm13020344.
11. Akdeniz BS, Tosun ME. A review of the use of artificial intelligence in orthodontics. *J Exp Clin Med* 2021; **38(S2)**:157-62 doi: 10.52142/omujecm.38.si.dent.
12. Daraqel B, Wafaie K, Mohammed H, Cao L, Mheissen S, Liu Y, *et al*. The performance of artificial intelligence models in generating responses to general orthodontic questions: ChatGPT *vs.* Google Bard. *Am J Orthod Dentofacial Orthop* 2024; **165(6)**:652-62. doi: 10.1016/j.ajodo.2024.01.012.
13. Tanaka OM, Gasparello GG, Hartmann GC, Casagrande FA, Pithon MM. Assessing the reliability of ChatGPT: A content analysis of self-generated and self-answered questions on clear aligners, TADs and digital imaging. *Dental Press J Orthod* 2023; **28(5)**:e2323183. doi: 10.1590/2177-6709.28.5.e2323183.
14. Mesko B. Prompt engineering as an important emerging skill for medical professionals: Tutorial. *J Med Internet Res* 2023; **25**:e50638. doi: 10.2196/50638.
15. AlZu'bi S, Mughaid A, Quiam F, Hendawi S. Exploring the capabilities and limitations of ChatGPT and alternative big language models. *Artif Intell Appl* 2023; **2(1)**:28-37. doi:10.47852/bonviewAIA3202820.
16. Ali R, Tang OY, Connolly ID, Fridley JS, Shin JH, Zadnik Sullivan PL, *et al*. Performance of ChatGPT, GPT-4, and Google Bard on a neurosurgery oral boards preparation question bank. *Neurosurgery* 2023; **93(5)**:1090-8. doi: 10.1227/neu.0000000000002551.
17. Cheong RCT, Pang KP, Unadkat S, Mcneillis V, Williamson A, Joseph J, *et al*. Performance of artificial intelligence chatbots in sleep medicine certification board exams: ChatGPT *versus* Google Bard. *Eur Arch Otorhinolaryngol* 2024; **281(4)**: 2137-43. doi: 10.1007/s00405-023-08381-3.
18. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, *et al*. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. *JMIR Med Educ* 2023; **9**:e45312. doi: 10.2196/45312.
19. Toyama Y, Harigai A, Abe M, Nagano M, Kawabata M, Seki Y, *et al*. Performance evaluation of ChatGPT, GPT-4, and Bard on the official board examination of the Japan Radiology Society. *Jpn J Radiol* 2024; **42(2)**:201-7. doi: 10.1007/s11604-023-01491-2.
20. Naureen S, Kiani HG, NaureenN0, Shafique M. Comparison of knowledge and attitude towards Chat GPT in first and final-year dental students. *JRMC* 2024; **28(1)**:440-5. doi:10.37939/jrmc.v28i1.2425.

21. Husain S, Ansari Z, Hussain A, Abbasi S, Ayoob T, Mujahid R. To evaluate the efficiency of ChatGPT in medical education: An analysis of MCQ-based learning and assessment. *Ann Abbasi Shaheed Hosp Karachi Med Dent Coll* 2023; **28(4)**:194-200.

22. Huh S. Are ChatGPT's knowledge and interpretation ability comparable to those of medical students in Korea for taking a parasitology examination?: A descriptive study. *J Educ Eval Health Prof* 2023; **20**:1. doi: 10.3352/jeehp.2023.20.1.

23. Juhi A, Pipil N, Santra S, Mondal S, Behera JK, Mondal H. The capability of ChatGPT in predicting and explaining common drug-drug interactions. *Cureus* 2023; **15(3)**:e36272. doi: 10.7759/cureus.36272.

24. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, *et al*. Large language models encode clinical knowledge. *Nature* 2023; **620(7972)**:172-80. doi: 10.1038/s41586-023-06291-2.