Available online at www.medicinescience.org

ORIGINAL ARTICLE

Medicine Science International Medical Journal

Medicine Science 2026;15(3):1132-7

Performance of ChatGPT-4o and ChatGPT-5.2 in child psychiatry assessments: A comparative study

Burcu Yildirim Budak[1], Ayse Dilara Oztermeli[2]

[1]İstanbul Medeniyet University, Göztepe Prof. Dr. Süleyman Yalçın City Hospital, Department of Child and Adolescent Psychiatry, İstanbul, Türkiye
[2]Kocaeli City Hospital, Department of Emergency Medicine, Kocaeli, Türkiye

Received 15 January 2026; Accepted 01 April 2026
Available online 11 August 2026 with doi: 10.5455/medscience.2026.01.012

Abstract

Although previous studies have assessed the adequacy of ChatGPT (Generative Pre-Trained Transformer) in psychiatry, no research, to our knowledge, has focused specifically on child psychiatry. This study aimed to evaluate the performance of ChatGPT in child psychiatry by testing it with board-level questions from the question bank of a reference textbook in the field. Seven tests from the textbook were randomized using random.org, and Tests 1,2, and 5 were selected. A total of 600 questions were presented to two versions of ChatGPT: GPT-4o (July 2024) and GPT-5.2 (December 2025). Each response was recorded as correct or incorrect. The questions were further categorized into two groups based on length (short vs.long) and question type (multiple-choice questions–MCQ vs. true/false questions-TFQ). ChatGPT-4o answered 522 (87.0%) of the 600 questions correctly, while GPT-5.2 provided correct answers to 546 (91.0%). The improvement in ChatGPT-4o's and ChatGPT-5.2's accuracy was statistically significant ($p < 0.001$). No significant differences were found between short and long questions for either version (ChatGPT-4o: $p = 0.61$; ChatGPT-5.2: $p = 0.60$). Both versions achieved higher accuracy on MCQs compared to TFQs, this difference was borderline significant for ChatGPT-4o ($p < 0.047$) and clearly significant for ChatGPT-5.2 ($p < 0.009$). Both versions of ChatGPT performed successfully, with the latest version, ChatGPT-5.2, showing significantly improved accuracy. As large language models continue to enhance, improvements in accuracy may contribute to reducing certain risks in psychiatry-related tasks.

Keywords: Child psychiatry, artificial intelligence, ChatGPT, lewis, board exam questions

Introduction

ChatGPT (Generative Pre-Trained Transformer), developed by OpenAI (OpenAI's artificial intelligence, San Francisco, CA), is an artificial intelligence chatbot designed for dialogue, functioning as a large language model [1,2]. In psychiatry and other areas of medicine, these AI tools offer clinicians quick access to up-to-date guidelines, algorithms, positioning AI as a valuable complement to clinical practice. ChatGPT can assist healthcare professionals and medical students with academic tasks such as conducting literature reviews, drafting theses or manuscripts. Another potential application involves using ChatGPT as a virtual assistant to provide preliminary support to patients before they are able to reach a healthcare professional. However, the versatility of ChatGPT also introduces challenges, including limitations, medical risks, and ethical concern [3]. Some clinicians remain cautious when using AI-generated information, citing the risk of incomplete or inaccurate content [4].

While the benefits and drawbacks of ChatGPT in medicine are also relevant to psychiatry, the field faces additional challenges. Issues such as psychotherapy, emotion recognition, and empathy introduce complexities that limit its use in psychiatric settings [5].

Recent studies have assessed ChatGPT's competence in the field of medicine. ChatGPT's competence and performance across medical specialties was evaluated using questions from the US Medical Licensing Examination Step 2. ChatGPT achieved an accuracy rate of 89% when answering these board-level questions. ChatGPT demonstrated strong performance in psychiatry [6]. ChatGPT was tested with questions from

CITATION
Yildirim Budak B, Oztermeli AD. Performance of ChatGPT-4o and ChatGPT-5.2 İn child psychiatry assessments: A comparative study. Med Science. 2026;15(3):1132-7.

Corresponding Author: Burcu Yildirim Budak, İstanbul Medeniyet University, Göztepe Prof. Dr. Süleyman Yalçın City Hospital, Department of Child and Adolescent Psychiatry, İstanbul, Türkiye
Email: burcuyildirimbudak@hotmail.com

DOI: 10.5455/medscience.2026.01.012

the Türkiye Medical Specialty Examination (MSE/TUS), a mandatory exam for doctors seeking to enter specialty training after medical school. Using five years of exam data, ChatGPT performed successfully in this context as well [7]. Despite these promising results, research supporting ChatGPT's applications specifically in psychiatry remains limited. One study evaluated its competence by preparing 40 questions across various clinical psychiatry topics. The model responded with high levels of accuracy, completeness, with ChatGPT users outperforming those using other sources [8]. ChatGPT achieved a passing score on the Taiwan Psychiatric Licensing Examination [9].

Child adolescent psychiatry is a specialized medical field in our country. To become a specialist, candidates must complete six years of medical education. An additional four years of specialty training is required, which involves writing a thesis and passing a scientific examination. Although studies evaluating ChatGPT's competence in psychiatry exist in the literature, to our knowledge, no such study has been conducted specifically in the field of child adolescent psychiatry. This study aims to assess ChatGPT's performance in this area by presenting it with board-level questions from Lewis's Question Bank [10], a widely recognized reference textbook in the field.

Material and Methods

Lewis's Child and Adolescent Psychiatry [11] textbook has long been a cornerstone in the library of child and adolescent psychiatrists. Over the years, it has undergone significant revisions and updates, continuously evolving to integrate scientific principles, research methodologies, and comprehensive information. The 2010 edition of Lewis's Child and Adolescent Psychiatry Review (Question Bank), edited by Yann B. Poncin and Prakash K. Thomas [10], serves as a valuable companion to the main textbook. This review is designed to assist with preparation for child and adolescent psychiatry residency exams, specialty board exams, recertification processes, and the Child Psychiatry Resident In-Training Examination. The question bank includes seven tests, each consisting of 200 questions, featuring both multiple-choice and true/false formats.

Data collection was carried out by asking questions to different versions of ChatGPT in two separate periods, July 2024 and December 2025. Seven tests from the question bank were selected using randomization via Random.org. The 1st, 2nd, and 5th tests were chosen, resulting in a total of 600 questions that were presented to ChatGPT. These questions were administered to two versions of the GPT: the July 2024 release (ChatGPT-4o) and the December 2025 release (ChatGPT-5.2) by Open AI. The models used in the study were accessed through the ChatGPT web interface provided by OpenAI. During each evaluation, the relevant model version **(**ChatGPT-4o and ChatGPT-5.2**)** was explicitly selected by the researchers. The most up-to-date version available at the time the study was conducted and during the manuscript preparation period was used. As new versions of ChatGPT were released during the writing phase of the manuscript, the study methodology was updated accordingly, and the questions were directed to the most current version available at that time. During the evaluation process, no system prompts, custom instructions, or persistent memory features were used. All interactions were conducted using standard user prompts under the default interface settings. This approach was chosen to ensure that the model responses reflected the default behavior of the system and to enhance the reproducibility of the study.

To minimize recall bias, each question was asked in a new chat session. Before each question, the model was provided with an introductory prompt stating: "You will be asked multiple-choice or true/false questions from a child and adolescent psychiatry question bank. Please carefully evaluate the question and indicate the correct answer." Subsequently, the question and answer options were copied exactly as they appeared in the question bank. The accuracy of the model responses was evaluated using the standard answer key included in the question bank. During the evaluation process, two researchers independently reviewed the responses, and any disagreements were resolved through discussion and consensus. Some questions have more than one correct answer in the question bank, and this is also indicated in the answer key. For these questions, one of the correct answers provided by ChatGPt was counted as correct. Questions were manually entered into ChatGPT's new chat bar, and responses were recorded as either correct or incorrect. The 600 questions were categorized into two groups based on the length of the question, determined by the number of characters. The average length of all questions in the dataset was calculated. The analysis revealed that the average question length was 46 characters. Therefore, based on the dataset-specific distribution, questions were classified as "short questions" (46 characters or less) and "long questions" (over 46 characters). This approach aims to examine the potential impact of different question lengths on model performance by roughly dividing the dataset into two comparable groups. Additionally, the questions were categorized into two types: multiple-choice questions (MCQ) and true/false questions (TFQ). Some of the multiple-choice questions include more than one correct option in their original form, as presented in the question bank book, and the answer key also contains multiple possible correct responses. If the model provided any one of the answers listed in the answer key, it was accepted as correct. MCQ: A type of question where one or more correct answers are selected from several options. TFQ: A question requiring the evaluation of a clear and explicit statement as either correct or incorrect. Each category was analyzed for both ChatGPT versions (4o and 5.2), and the results were compared.

Ethical approval was not needed for the study because this is not a study about humans or animals, and only consists of questions asked to AI.

DOI: 10.5455/medscience.2026.01.012

Statistical Analysis

All statistical analyses were conducted using the JAMOVI software (Sydney, Australia; version 2.3.17). All analyses were performed using the base statistical modules available in JAMOVI (Exploration and Frequencies modules). Descriptive statistics were used to summarize the study data. Continuous variables were expressed as mean ± standard deviation (SD), while categorical variables were reported as frequency and percentage. The normality of continuous variables was assessed using the Shapiro–Wilk test. Since the distribution of the variables did not significantly deviate from normality, descriptive statistics were presented as mean ± SD. Comparisons between categorical variables were performed using the Chi-square test. A p-value of less than 0.05 was considered statistically significant.

Results

A total of three different test exams, each comprising 200 questions (600 questions in total), were administered to both the ChatGPT-4o and ChatGPT-5.2 versions of ChatGPT. The number of correct answers provided by each version, along with the corresponding accuracy percentages, are presented in Table 1.

Table 1. Correct and incorrect responses provided by ChatGPT for the test questions

Test	ChatGPT- 4o			ChatGPT- 5.2			Total
	Correct	Incorrect	Correct percentage	Correct	Incorrect	Correct percentage	
Test 1	172	28	86.0%	185	15	92.5%	**200**
Test 2	174	26	87.0%	178	22	89.0%	**200**
Test 5	176	24	88.0%	183	17	91.5%	**200**
Total	**522**	**78**	**87.0%**	**546**	**54**	**91.0%**	**600**
Chi-square test, $p < 0.001$							

When all test questions were evaluated together, ChatGPT-4o correctly answered 522 out of 600 questions (87.0%), while ChatGPT-5.2 correctly answered 546 out of 600 questions (91.0%). The higher accuracy rate of ChatGPT-5.2 was statistically significant ($p < 0.001$).

When the tests were analyzed individually:

- In Test 1, ChatGPT-4o correctly answered 172 out of 200 questions (86.0%), while ChatGPT-5.2 correctly answered 185 out of 200 questions (92.5%).
- In Test 2, ChatGPT-4o correctly answered 174 out of 200 questions (87.0%), while ChatGPT-5.2 correctly answered 178 out of 200 questions (89.0%).
- In Test 5, ChatGPT-4o correctly answered 176 out of 200 questions (88.0%), while ChatGPT-5.2 correctly answered 183 out of 200 questions (91.5%).

In all three tests, ChatGPT-5.2 demonstrated a statistically significant improvement in accuracy compared to ChatGPT-4o ($p < 0.001$).

When questions were evaluated based on their length, the 600 questions ranged from a minimum of 10 characters to a maximum of 203 characters, with an average of 46 ± 28.9 characters. The questions were categorized into two groups: short questions (fewer than 46 characters) and long questions (46 or more characters). In the short question group, ChatGPT 4o correctly answered 343 out of 392 questions (87.5%), while ChatGPT 5.2 correctly answered 355 out of 392 questions (90.6%). In the long question group, ChatGPT 4o correctly answered 179 out of 208 questions (86.1%), while ChatGPT 5.2 correctly answered 191 out of 208 questions (91.8%). There was no statistically significant difference in accuracy between the short and long question groups for either ChatGPT 4o ($p = 0.61$) or ChatGPT 5.2 ($p = 0.60$) (Table 2).

Table 2. Evaluation of responses based on question length

Question length	ChatGPT- 4o		ChatGPT- 5.2		Total
	Correct	Incorrect	Correct	Incorrect	
Short question	343 (87.5%)	49 (12.5%)	355 (90.6%)	37 (9.4%)	**392 (100%)**
Long question	179 (86.1%)	29 (13.9%)	191 (91.8%)	17 (8.2%)	**208 (100%)**
Chi-square test, ChatGPT-4o $p = 0.61$, ChatGPT-5.2 $p = 0.60$					

DOI: 10.5455/medscience.2026.01.012

When the questions answered by ChatGPT were categorized as MCQ and TFQ, ChatGPT-4o correctly answered 425 out of 481 multiple-choice questions (88.4%), while ChatGPT-5.2 correctly answered 445 out of 481 (92.5%). In the true/false question group, ChatGPT-4o correctly answered 97 out of 119 questions (81.5%), while ChatGPT-5.2 correctly answered 101 out of 119 (84.9%). When the performance of ChatGPT-4o was compared across question types, a statistically significant difference was observed between multiple-choice questions and true/false questions (Chi-square test, p = 0.047). Although this result reached statistical significance, the p value was close to the threshold, indicating that the magnitude of difference between the two question formats for ChatGPT-4o was relatively modest. In contrast, the comparison of multiple-choice versus true/false performance for ChatGPT-5.2 revealed a more robust and statistically significant difference (Chi-square test, p = 0.009), suggesting that ChatGPT-5.2 demonstrated a more consistent advantage in handling multiple-choice questions relative to true/false questions (Table 3).

Table 3. Results when questions were categorized as multiple choice or true/false questions

Question type	ChatGPT- 4o		ChatGPT- 5.2		Total
	Correct	Incorrect	Correct	Incorrect	
Multiple choice	425 (88.4%)	56 (11.6%)	445 (92.5%)	36 (7.5%)	**481 (100.0%)**
True / False	97 (81.5%)	22 (18.5%)	101 (84.9%)	18 (15.1%)	**119 (100.0%)**
Chi-square test, ChatGPT-4o p = 0.047, ChatGPT-5.2 p = 0.009					

Discussion

Recent advancements in AI technology have been remarkable. ChatGPT, a conversational AI model developed by OpenAI, is capable of generating human-like text, answering questions, and engaging in discussions on a wide range of topics. Its use in healthcare and medicine, like in many other sectors, has garnered attention for its potential benefits and drawbacks. In the specialized field of child and adolescent psychiatry, potential applications include providing clinicians and researchers with quick access to clinical knowledge, guidelines, and algorithms, as well as supporting tasks such as writing, automatic translation, and more. These capabilities can enhance work efficiency, aiding clinicians and researchers in better managing their time.

In recent years, while the number of child and adolescent psychiatry specialists has been growing, access to child and adolescent mental health services remains limited for patients and their families, particularly in developing countries and rural areas [12]. In this context, ChatGPT could serve as a tool to offer solutions and stepwise guidance to patients and families until they are able to access a mental health professional or center. However, despite these potential benefits, there are also risks, including misinformation, ethical and copyright concerns for clinicians and researchers, and the possibility of providing misleading or unsafe advice to patients and their families [3,13]. For these reasons, the findings of our study, which evaluates ChatGPT's performance and adequacy in the field of child and adolescent mental health, conclude that ChatGPT demonstrated strong performance in answering structured knowledge questions in child and adolescent psychiatry.

Child and adolescent psychiatry began to develop at the beginning of the 20th century and has continued to evolve, particularly in the realm of education [14]. Board-level exams assessing competency in child and adolescent psychiatry generally cover theoretical and conceptual knowledge from textbooks, clinical practice approaches, and current literature in the field [15]. Although the training process to become a child and adolescent psychiatrist varies across countries, it remains a demanding and lengthy endeavor. *Lewis's Child and Adolescent Psychiatry: A Comprehensive Textbook* [11] was published half a century after Leo Kanner's classic *Child Psychiatry* [16] and continues to be updated with new editions. This book is a fundamental, encyclopedic reference found in the libraries of many child and adolescent psychiatrists (CAPs), providing answers for daily clinical practice. In our study, we utilized *Lewis's Child and Adolescent Psychiatry Review (Question Bank)* [10], which contains questions designed to help clinicians in the CAP field prepare for board exams. These questions are structured to cover the core competencies required for such exams. In our study, questions from Lewis's question bank were categorized according to various characteristics. Both the ChatGPT-4o and ChatGPT-5.2 versions showed significant positive results across different types of questions, including short and long questions, as well as multiple-choice and true/false formats. The latest version, ChatGPT-5.2, demonstrated a significantly higher accuracy rate in answering questions correctly.

In our study, an analysis of ChatGPT-4o's performance across different question types revealed a statistically significant difference between multiple-choice and true/false formats; however, the p value's proximity to the significance threshold suggests that the practical magnitude of this difference may be limited. In contrast, ChatGPT-5.2 showed a clearer and more robust advantage in handling MCQs, performing substantially better than on true/false items. This finding indicates that the

DOI: 10.5455/medscience.2026.01.012

newer model provides a more stable and reliable response profile in question types requiring more complex option analysis. In another study, multiple-choice clinical case questions from the Foreign Medical Graduate Examination were presented to ChatGPT-3.5, which achieved a 90% success rate, accurately diagnosing most case [17]. However, in a different study where written examination questions from Japan's childcare worker national examination were posed to ChatGPT-3.5 and 4.0, neither version met the passing criteria, raising concerns about the accuracy of ChatGPT's responses in that context [18]. One possible explanation for the higher performance on MCQs is that the presence of several answer options may provide contextual cues that help LLMs identify the most plausible response. In contrast, TFQs require a definitive judgment about a single statement, which may increase the likelihood of error if the model misinterprets a specific concept. This difference may partly reflect the sequence-prediction mechanisms through which LLMs process textual information [5].

When evaluating the questions posed to ChatGPT based on format, both ChatGPT-4o and ChatGPT-5.2 performed well on short and long questions in our study, with no significant difference in performance observed between the two question lengths for either version.

A systematic review and meta-analysis assessing ChatGPT's effectiveness in medical education evaluated 45 studies, focusing on the performance of different ChatGPT versions in "Medical Licensing Examinations." ChatGPT-4 achieved an overall accuracy rate of 81%, while ChatGPT-3.5 achieved 58%. Additionally, ChatGPT-3.5 performed better on short-text questions compared to long-text questions, indicating that question length influenced performance. The difficulty level of the questions also affected the performance of both ChatGPT-3.5 and ChatGPT-4. ChatGPT's accuracy in MCQs ranged from 13.1% to 100%, with significantly better performance on MCQs than on open-ended questions [19].

In another study, "medical specialty exam" questions were categorized by word count, and it was found that ChatGPT-3.5 provided significantly more correct answers to short questions than to long ones [7].

There is a significant shortage of child and adolescent psychiatrists in the United States, and ongoing efforts aim to increase the number and accessibility of trained professionals in this field[15]. In situations where immediate access to specialists is not feasible, supplementary and supportive tools can offer initial recommendations, and ChatGPT appears to be a promising candidate for this role. However, the human-centered nature of psychiatry is essential, making it crucial that ChatGPT be used to enhance rather than replace treatment [20].

In our study, the performance of ChatGPT models was evaluated using board-level questions rather than real clinical decision-making scenarios. The results indicated that ChatGPT models demonstrated strong and adequate performance in answering structured, exam-style knowledge questions in the field of child and adolescent psychiatry. Therefore, our findings reflect the ability of ChatGPT models to retrieve and apply theoretical knowledge rather than their clinical reliability or safety in psychiatric practice. Additionally, LLMs may occasionally generate incorrect or fabricated information (so-called "AI hallucinations"), which may pose risks if used without professional supervision [21,22]. Therefore, our findings should be interpreted within the scope of the study design.

Limitations

The analysis was based on the model versions available at the time of data collection and may not reflect newer iterations. When model responses were examined in terms of content, clinically risky misinformation (e.g., suggesting dosages or making diagnoses) was rare. This study primarily examines the accuracy of the model's responses to knowledge-based questions. Clinical safety, the model's ability to produce hallucinations, make inappropriate suggestions, or its reliability in actual clinical decision-making processes were not evaluated within the scope of this study. Therefore, the findings should be interpreted as the model's performance in answering structured knowledge questions; they should not be directly generalized to safety or reliability in clinical psychiatric practice.

Conclusion

The use of ChatGPT in mental health may potentially serve as a first-step provider of services, offering information on psychiatric conditions, psychoeducation, self-control, self-management skills, and relaxation techniques in situations where the risk of misinformation is low, until access to a professional or center is available. Additionally, ChatGPT can be a valuable tool for medical students, residents, and doctors by providing quick access to reliable information on clinical psychiatry and research. This utility may help bridge the gap in medical education and healthcare between wealthier nations and low- and middle-income countries[9].

In our study, both versions of ChatGPT achieved positive results, with the latest version showing significantly better performance. As LLMs continue to evolve, improvements in performance may potentially reduce some risks associated with their use. However, further empirical studies are required to evaluate these developments in real clinical settings. Additionally, longitudinal studies are needed to observe these developments over time. It is crucial to conduct ethical research when relying on AI-derived medical information. While acknowledging its benefits, the associated risks should not be underestimated, and careful consideration is essential in its use and recommendations.

Future studies may include content safety evaluations to ensure the responsible use of AI tools in clinical psychiatry and education.

Ethical Approval
Ethical approval was not needed for the study because this is not a study about humans or animals, and only consists of questions asked to AI.

Patient Consent
Patient consent was not applicable because no human participants or patient data were involved in this study.

Conflict of Interest
The authors confirm the absence of any conflicts associated with this work.

Funding
The authors declare that they have received no financial support for the study.

Acknowledgements
We would like to express our sincere gratitude to Dr. Yann B. Poncin and Dr. Prakash K. Thomas from the Child Study Center, Yale University School of Medicine, for their work on Lewis's Child and Adolescent Psychiatry Review: A Board Review and Companion Guide. This comprehensive resource served as an essential reference and inspiration during the development of our study's question set and analysis framework.

References

1. OpenAI. Introducing ChatGPT. https://openai.com/index/chatgpt/ access date 10.03.2024.
2. ChatGPT. https://tr.wikipedia.org/wiki/ChatGPT access date 10.03.2024.
3. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell. 2023;6:1169595.
4. Sallam, M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel). 2023;11:887.
5. Cheng SW, Chang CW, Chang, WJ, et al. The now and future of ChatGPT and GPT in psychiatry. Psychiatry Clin Neurosci. 2023;77:592-6.
6. Mackey BP, Garabet R, Maule L, et al. Evaluating ChatGPT-4 in medical education: an assessment of subject exam performance reveals limitations in clinical curriculum support for students. Discov Artif Intell. 2024;4:38.
7. Oztermeli AD, Oztermeli A. ChatGPT performance in the medical specialty exam: an observational study. Medicine (Baltimore). 2023;102:e34673.
8. Luykx JJ, Gerritse F, Habets PC, Vinkers CH. The performance of ChatGPT in generating answers to clinical questions in psychiatry: a two-layer assessment. World Psychiatry. 2023;22:479-80.
9. Li DJ, Kao YC, Tsai SJ, et al. Comparing the performance of ChatGPT GPT-4, Bard, and Llama-2 in the Taiwan Psychiatric Licensing Examination and in differential diagnosis with multi-center psychiatrists. Psychiatry Clin Neurosci. 2024;78:347-52.
10. Poncin YB, Thomas PK. Lewis's Child and Adolescent Psychiatry Review: 1,400 Questions to Help You Pass the Boards. Philadelphia: Lippincott Williams & Wilkins. 2009.
11. Martin A, Volkmar FR, editors. Lewis's Child and Adolescent Psychiatry: A Comprehensive Textbook. 4th ed. Philadelphia: Lippincott Williams & Wilkins. 2007.
12. McBain RK, Cantor JH, Kofner A, et al. Ongoing disparities in digital and in-person access to child psychiatric services in the United States. J Am Acad Child Adolesc Psychiatry. 2022;61:926-33.
13. Borji A. A categorical archive of ChatGPT failures. https://arxiv.org/abs/2302.03494 access date 16.07.2024.
14. Sexson SB. Overview of training in the twenty-first century. Child Adolesc Psychiatr Clin N Am. 2007;16:1-16.
15. Hunt J, Reichenberg J, Lewis AL, Jacobson S. Child and adolescent psychiatry training in the USA: current pathways. Eur Child Adolesc Psychiatry. 2020;29:63-9.
16. Kanner L. Child Psychiatry. 4th ed. Springfield, IL: Charles C Thomas Publisher. 1979.
17. Sathyam PKR, Surapaneni KM. Assessing the performance of ChatGPT in psychiatry: A study using clinical cases from foreign medical graduate examination (FMGE). Indian J Psychiatry. 2024;66:408-10.
18. Kaneda Y, Namba M, Kaneda U, Tanimoto T. Artificial intelligence in childcare: assessing the performance and acceptance of ChatGPT responses. Cureus. 2023;15:e44484.
19. Liu M, Okuhara T, Chang X, et al. Performance of ChatGPT across different versions in medical licensing examinations worldwide: systematic review and meta-analysis. J Med Internet Res. 2024;26:e60807.
20. Wei Y, Guo L, Lian C, Chen J. ChatGPT: Opportunities, risks and priorities for psychiatry. Asian J Psychiatr. 2023;90:103808.
21. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: can language models be too big?. 2021 ACM Conference on Fairness, Accountability, and Transparency, 3-10 March 2021. Virtual Event, Canada, 610-23.
22. Alkaissi H, McFarlane SI. Artificial hallucinations in ChatGPT: implications in scientific writing. Cureus. 2023;15:e35179.