

RESEARCH

Open Access



Assessment of the capacity of ChatGPT as a self-learning tool in medical pharmacology: a study using MCQs

Woong Choi^{1*}

Abstract

Background ChatGPT is a large language model developed by OpenAI that exhibits a remarkable ability to simulate human speech. This investigation attempts to evaluate the potential of ChatGPT as a standalone self-learning tool, with specific attention on its efficacy in answering multiple-choice questions (MCQs) and providing credible rationale for its responses.

Methods The study used 78 test items from the Korean Comprehensive Basic Medical Sciences Examination (K-CBMSE) for years 2019 to 2021. 78 test items translated from Korean to English with four lead-in prompts per item resulted in a total of 312 MCQs. The MCQs were submitted to ChatGPT and the responses were analyzed for correctness, consistency, and relevance.

Results ChatGPT responded with an overall accuracy of 76.0%. Compared to its performance on recall and interpretation questions, the model performed poorly on problem-solving questions. ChatGPT offered correct rationales for 77.8% (182/234) of the responses, with errors primarily arising from faulty information and flawed reasoning. In terms of references, ChatGPT provided incorrect citations for 69.7% (191/274) of the responses. While the veracity of reference paragraphs could not be ascertained, 77.0% (47/61) were deemed pertinent and accurate with respect to the answer key.

Conclusion The current version of ChatGPT has limitations in accurately answering MCQs and generating correct and relevant rationales, particularly when it comes to referencing. To avoid possible threats such as spreading inaccuracies and decreasing critical thinking skills, ChatGPT should be used with supervision.

Keywords ChatGPT, Large language model, Self-directed learning, Performance, Multiple-choice questions, Rationale, Referencing

Introduction

Created by OpenAI, ChatGPT is an advanced large language model (LLM) that has been pre-trained to chat in natural language [1]. Since its launch in late 2022,

ChatGPT has drawn considerable attention from the public. Thanks to its large capacity and training text corpora [2], ChatGPT is able to produce human-like responses, going as far as to demonstrate reasoning through chain-of-thoughts mimicking human problem-solving behavior [3–6]. After ChatGPT met the passing threshold on the United States Medical Licensing Examination (USMLE) [7, 8], many authors applied ChatGPT on answering other multiple-choice questions (MCQs) in the medical domain such as physiology [9], anesthesiology [10, 11],

*Correspondence:

Woong Choi
wchoi@chungbuk.ac.kr

¹ Department of Pharmacology, College of Medicine, Chungbuk National University, Cheongju, Chungbuk 28644, Korea



© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

ophthalmology [12], and parasitology [13]. One meta-analysis reported that ChatGPT demonstrated an accuracy of 61.1% (95% CI 56.1%–66.0%) in answering MCQs in medical examinations [14].

Given that MCQs can be used as a self-learning tool [15, 16], such performance suggests that ChatGPT could act as an easy-to-access interactive learning environment, which could lead to greater retention of information and more pleasant learning experience [7].

The Korean Comprehensive Basic Medical Sciences Examination (K-CBMSE) is a minimum competency test taken by Korean medical students who have completed didactic learning and laboratory experiment for basic medical sciences (See Supplement 1 for the details) [17]. One theme of K-CBMSE focuses on pharmacology, which includes MCQs at three levels of cognitive taxonomy: recall, interpretation, and problem-solving [16]. Pharmacology is often perceived as a challenging subject by students due to (1) the introduction of numerous new terms and concepts, and (2) requirement of complex background knowledge such as pathophysiology and biochemistry. Therefore, reinforcement of key concepts by self-learning is essential to improve understanding, learning and retention [18].

ChatGPT was suggested as a self-learning tool for students facing difficulties in learning pharmacology, as it achieved a high accuracy rate when answering centric questions from a pharmacology textbook for undergraduate students [19]. However, ChatGPT's ability to answer MCQs in pharmacology have not been addressed in the past literature. In this study, the capacity of ChatGPT as a self-learning tool for pharmacology was tested on selected MCQs from the pharmacology section of K-CBMSE. ChatGPT was asked four incrementally designed prompts to provide answers, rationales (reasoning or justification) supporting its answers, references for the rationale, and relevant paragraphs or excerpts from each reference. The accuracy of answers, the soundness of rationales, and the veracity of references and relevant paragraphs were evaluated. Cases of incorrect answers and rationales were identified along with potential causes for the errors. Possible strategies to minimize the drawbacks of ChatGPT were discussed.

Methods

Aim

This study assessed ChatGPT's potential as a standalone self-learning tool for medical pharmacology by evaluating its response to 312 MCQs derived from the K-CBMSE test items. The responses were assessed based on the correctness of answer, rationale, references, and paragraph from each respective reference. As MCQs are a combination of test items and incrementally engineered lead-in

prompts, the study also tested whether the cognitive taxonomy level of the test items and the incrementally engineered prompts interacted to influence ChatGPT's performance.

Construction of test item dataset

Test items from the K-CBMSE for years 2019 to 2021 (a total of 105 test items) were used as the test item dataset.¹ Test items with figures (27 items) were excluded because ChatGPT could not interpret images. The remaining 78 test items were translated from Korean to English by the author. During the translation, long Korean sentences were split into short English sentences for better readability, and appropriate plain words or medical terms were used where required. The cognitive taxonomy level of the test items was also rated by the author as recall, interpretation, and problem-solving [16].

Prompt engineering

For the answer, references for the rationales, relevant paragraphs or excerpts in each reference), lead-in prompts were engineered incrementally for each of the four levels. This four-level prompting is an incremental prompting technique that uses four levels of prompts to guide a large language model (LLM) such as ChatGPT towards a desired response by providing multiple prompts, one after another [20]. It was hypothesized that incremental prompting might increase ChatGPT's workload and error rate.

- (1) *Prompt 1 (correct answer): Please choose the best answer for the following question,*
- (2) *Prompt 2 (rationale): Please choose the best answer for the following question and explain the rationale,*
- (3) *Prompt 3 (references): Please choose the best answer for the following question and explain the rationale. Please provide the references (Uniform Resource Locator or URL, title, and authors) that support the rationale,*
- (4) *Prompt 4 (relevant paragraph): Please choose the best answer for the following question and give a rationale for the answer. Please provide the references (URL, title, and authors) that support your rationale. Please provide the relevant paragraphs or formulas from each reference.*

¹ The test items of K-CBMSE are not publicly available, but can be provided by the Medical Education Assessment Corporation (bmeac.ac.kr) on demand for educational or research purposes.

<p>User: Please choose the best answer for the following question and give a rationale for the answer. Please provide the references (URL, title, and authors) that support your rationale. Please provide the relevant paragraphs or formulas from each reference.</p>	lead-in prompt
<p>2021_#25. A 56-year-old man presented at the hospital with heartburn. He was diagnosed with peptic ulcer after endoscopic examination and omeprazole was prescribed. What is the mechanism of action of omeprazole?</p>	test item (stem & question)
<p>1) neutralizing gastric acid 2) inhibiting proton pump 3) blocking muscarinic receptor 4) blocking histamine H2 receptor 5) promoting prostaglandin synthesis</p>	options

Fig. 1 The typical style of each multiple-choice question. Each question consisted of a lead-in prompt, a test item, and options. The lead-in prompt could ask for the answer; a rationale supporting the answer; references; and paragraphs from the references

<p>ChatGPT: The correct answer is 2) inhibiting proton pump.</p>	answer
<p>Omeprazole is a proton pump inhibitor (PPI), ... inhibiting the H+/K+-ATPase enzyme system, also known as the proton pump... omeprazole reduces acid production and thereby helps to relieve symptoms of acid-related conditions such as heartburn.</p>	rationale
<p>References: 2. Title: "Proton pump inhibitors: a review of their pharmacology and their therapeutic uses" Authors: J.J. Scarpignato and G. Zentilin URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1372933/</p>	references
<p>Relevant paragraphs: "Proton pump inhibitors (PPIs) are the most powerful acid-secretory inhibitors currently available. They irreversibly and specifically inhibit the H+/K+-ATPase enzyme system, also known as the proton pump, which is responsible for the final step in acid secretion. The proton pump is located on the parietal cell in the stomach and is a highly selective, membrane-bound enzyme that pumps protons from the cytoplasm into the stomach lumen, thus generating hydrochloric acid."</p>	relevant paragraphs

Fig. 2 The typical style of ChatGPT’s response. Depending on the level of the lead-in prompt, typical ChatGPT’s response consisted an answer; a rationale supporting the answer; references; and paragraphs from the references

Multiple-choice question dataset

A single MCQ was composed of one lead-in prompt, one blank line, and the original test item. Figure 1 shows the typical style of each MCQ. Since each test item could be paired with four different lead-in prompts, 78 test items generated a total of 312 MCQs.

ChatGPT

ChatGPT (versions between Jan/28/2023 and Jan/29/2023; OpenAI) was used. The version of ChatGPT is powered by Generative Pretrained Transformer (GPT)-3.5 and has been pretrained on a large size (from 570 gigabytes to 45 terabytes) of text data [1, 2]. There is no report whether ChatGPT received any additional pharmacology-specific pretraining.

ChatGPT inquiry and the responses thereof

ChatGPT was used to generate responses to English MCQs by feeding one question to the model at a time,

deleting previous chat history before each submission to prevent previous questions from influencing the current response. A response consisted of answers, rationales, reference lists, and relevant paragraphs from each reference, depending on the prompt. Figure 2 shows a typical example of ChatGPT’s response style.

Data summary

ChatGPT-generated responses were evaluated to check whether: (1) the answer for each MCQ was correct; (2) the answers to each test item were consistent over the prompts; (3) ChatGPT returned scientifically accurate and relevant rationales.; (4) reference lists were correct (i.e., ChatGPT provided the accurate citation information for references); and (5) a relevant paragraph could be found in the references listed, if possible. The results were binary (yes or no), and the number of yes and no results was summarized.

Table 1 ChatGPT performance for multiple-choice questions with prompt 4 (relevant paragraph)

Cognitive taxonomy level (Number of test items)	Correct answers to Prompt 4 (relevant paragraph) (%)	Overall accuracy over four lead-in prompts (%)
Recall (44)	81.8	86.4
Interpretation (20)	75.0	77.5
Problem-solving (14)	35.7	41.1
Total test items (78)	71.8	76.0

Please see Table S2-1 in Supplement 2 for the details

Statistical analysis

All results were expressed in categorical contingency tables and statistical analysis was performed online using Fisher’s exact test (<https://astatsa.com/FisherTest/>). When cell frequencies were less than 5, the Freeman-Halton extension of Fisher’s exact test was performed using the Free Statistics Calculator v4.0 (<https://www.danielsooper.com/statcalc/default.aspx>). If required, chi-square goodness-of-fit test was performed using Chi-Square Goodness of Fit Test Calculator (https://stats.libretexts.org/Learning_Objects/02%3A_Interactive_Statistics/36%3A__Chi-Square_Goodness_of_Fit_Test_Calculator). A *p*-value of less than 0.05 was considered statistically significant.

Results

Answers

The overall accuracy of ChatGPT’s answers to MCQs was 76.0% (Table 1). Responses with an incorrect answer (65/312, 20.8%), with multiple answers (4/312, 1.3%), or a response of “not determined” (6/312, 1.9%) were grouped as incorrect answers (75/312, 24.0%). Although the accuracies across the prompts varied from 71.8% to 82.1%, they did not differ significantly (See Supplement 2 for the details). ChatGPT’s accuracy was higher than its previously reported performance of 56.1%–66.0% (95% CI [14] and Korean students’ average performance of 55.3% (See Supplement 1 for the details). In terms of the cognitive taxonomy level of MCQs, the accuracy was 86.4% (152/176) for recall, 77.5% (62/80) for interpretation, and 41.1% (23/56) for problem-solving (See Supplement 2 for the details). Table 1 shows the performance for prompt 4 (relevant paragraph) in each taxonomy level, as the performance for other prompts was not significantly different.

To assess the concordance or consistency of answers to test items, the responses were aggregated by test item. The correctness for a single test item were classified as all correct, all incorrect, or partially correct across the prompts, and the all-correct and all-incorrect responses were regarded as concordant responses. Of the 78 test items, 60 items (76.9%) had concordant response across

Table 2 The concordance of answers to the test items across prompts

Answers across prompts	The number of test items	%
Concordance		
All correct	49	62.8
All incorrect	11	14.1
Discordance ^a		
Partially correct	18	23.1
Total	78	100.0

^a Discordant responses are inconsistent or conflicting set of answers to the same test item for repeated inquiries with different prompts

Table 3 Scientific accuracy of rationale for correct and incorrect responses

Answers (N)	Rationale (%)			
	Scientifically accurate	Scientifically inaccurate	Overall	<i>p</i> -value
Correct (178)	73.5	2.6	76.1	<i>p</i> < 0.05*
Incorrect (56)	4.3	19.6	23.9	
Subtotal (234)	77.8	22.2	100.0	

* The Fisher’s exact test and the follow-up chi-squared test ($\chi^2(df=1, N=234) = 152.93$) found significant interaction between the scientific accuracy of rationale and the accuracy of answers to the MCQs (*p* < 0.05)

the prompts (Table 2). For partially correct responses, the incorrect-to-correct answer ratio varied between 1-to-3 to 3-to-1 with various correct-incorrect sequences (data not shown).

Rationale

Prompts 2 (rationale) to 4 (relevant paragraph) required the rationale to be included in the response (234 MCQs). The scientific accuracy of each rationale was assessed by the author. Among 234 MCQs, 178 MCQs were correctly answered with either a correct (172/178, 96.6%) or an incorrect rationale (6/178, 3.4%). 56 MCQs were incorrectly answered with either a correct (10/56,

Table 4 The types and frequencies of reference errors

	Types of errors	N ₁ /N ₂ (%)
A	Reference URLs linked to a wrong page or a '404 page not available' (N ₁) among total references listed (N ₂)	191/274 (69.7)
B	The authors not found on PubMed, Amazon (for books), or Google (N ₁) among the authors appearing in the references (N ₂)	59/350 (16.9)
C	Incorrectly titled articles or books (N ₁) among the references (N ₂)	148/152 (97.4)

Table 5 The distribution of paragraphs between the correct and incorrect answers

Answers to the MCQs (N)	MCQs provided with Paragraphs (%)	MCQs not provided with Paragraphs (%)	Overall (%)	P-value
Correct (56)	46.2	25.6	71.8	<i>p</i> > 0.05*
Incorrect (22)	14.1	14.1	28.2	
Subtotal (78)	60.3	39.7	100.0	

* Fisher's exact test and a follow-up chi-squared test ($\chi^2(1, 78) = 1.35$) found no significant interaction between the correctness of ChatGPT's response and whether or not it was provided with relevant paragraphs from the references (*p* > 0.05)

17.9%) or an incorrect rationale (46/56, 82.1%). Overall, 22.2% (52/234) of the rationales were incorrect. Table 3 shows that correct answers were more likely supported by correct rationale and incorrect rationales were more likely to be associated with incorrect answers ($X^2(df=1, N=234) = 152.93, p < 0.05$).

The incorrect rationales could be grouped into one of the two categories: information errors (28/52, 53.8%) and reasoning errors (24/52, 46.2%). Information errors involved incorrect information or formula in the rationale (See Fig. S3-1 for correct and relevant supporting paragraph; Figs. S3-2 and S3-3 for the errors in the rationale in Supplement 3), while reasoning errors involved failed identification of the cues from the question stem, disregard of the cues in the question stem, or arithmetic errors including unit conversion (See Figs. S3-4 and S3-5 in Supplement 3).

References

Prompts 3 (references) and 4 (relevant paragraph) required references for the test items (156 MCQs). In total, 274 references were listed (Table 4). The reference lists consisted of URLs including PubMed, articles in journal citation format, and book information. Among these references, 191 (69.7%) had URLs linked to either an irrelevant or a nonexistent site, including PubMed links that did not match the relevant contents. A total of 350 authors were cited, but 59 authors (16.9%) could not be found on PubMed, Amazon, or Google. Even the combination of the existing authors did not find any relevant articles. Although 152 titles of articles or books were given, 148 titles (97.4%) were incorrect. Figure 2 shows a case of errors in referencing. The reference information

presented was "Proton pump inhibitors: a review of their pharmacology and their therapeutic uses. Scarpignato, JJ and Zentilin, G. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1372933/>". The PubMed Central Identifier PMC1372933 was directed to "Preventive Medicine in World War II. Vol. IV. Communicable Diseases Transmitted Chiefly through Respiratory and Alimentary Tracts. Am J Public Health Nations Health. 1959 Jul; 49(7): 969. PMCID: PMC1372933" and none of the listed authors could be found on PubMed, Amazon, or Google. The reference title, 'Proton pump inhibitors: a review of their pharmacology and their therapeutic uses', could not be found in a PubMed search. Even after the references were limited to textbooks and the prompts were modified to require actual books, the errors in the information on authors, book chapters, and pages were persistent (data not shown).

Relevant paragraphs

Prompt 4 (relevant paragraph) asked ChatGPT to identify the relevant paragraphs from each reference (78 MCQs). Only 47 MCQs from Table 5 were provided with paragraphs, which are presented in Table 6 (61 paragraphs). The distribution of the paragraph presentation did not differ significantly between correct and incorrect answers (Table 5; $\chi^2(1, 78) = 1.35, p > 0.05$).

Irrespective of the correctness of the reference, the contents of 61 paragraphs themselves could be grouped into (1) correct and relevant to the answer key (47/61, 77.0%), (2) correct but irrelevant to the answer key (6/61, 9.8%), and (3) incorrect information (8/61, 13.1%). As shown in Table 6, correct and relevant paragraphs are more likely to support correct answers (See the Supplement 3 for the

Table 6 The credibility of provided paragraphs between the correct and incorrect answers

Answers to the MCQs (N)	Correct and relevant Paragraph %	Correct but irrelevant Paragraph %	Incorrect Paragraph %	Overall %	P-value
Correct (49)	75.4	0.0	4.9	80.3	$p < 0.05^*$
Incorrect (12)	1.6	9.8	8.2	19.7	
Subtotal (61)	77.0	9.8	13.1	100.0	

* The Freeman-Halton extension of Fisher's exact test found significant interaction between the correctness of ChatGPT's response and the credibility of paragraphs from the references ($p < 0.05$)

details of each type of paragraph in Table 6; especially Fig. S3-1 for a correct and relevant paragraph, Fig. S3-6 for a correct but irrelevant paragraph, and Fig. S3-7 for an incorrect paragraph).

Discussion

Based on pharmacology MCQs, this study found that the current version of ChatGPT need to be improved to be used as a standalone self-learning tool. ChatGPT's overall performance (76%) in this study surpassed the ranges reported in the previous literature (61.1%, 95% CI 56.1%–66.0%) [14]. Its performance may vary depending on the number of subjects covered by each test, the numbers of options per MCQ, as well as the distribution of test items' cognitive taxonomy. However, its performance below 95% may limit its reliability as a self-learning tool [14]. ChatGPT outperformed Korean students in terms of overall accuracy in its response (76% vs 55%), but performed poorly on problem-solving MCQs (45%) despite its supposed critical thinking and problem-solving abilities. This result suggests that ChatGPT is still limited in its ability to apply critical thinking and reasoning skills to real-world problems.

Another issue with ChatGPT was the randomness of the generated responses. ChatGPT answered 23.1% of the test items inconsistently across the lead-in prompts. While randomness may be useful when generating creative content or exploring different ideas, it can be a critical problem when answering factual questions [21]. A particularly problematic form of randomness is hallucination, a phenomenon where ChatGPT generates plausible-sounding but incorrect or misleading [22–25]. The hallucinations can be caused by training data biases, lack of required information, limited real-world understanding, or algorithmic limitations [26]. The rationales for the answer and the supporting references were especially susceptible to hallucination. Among all the generated rationales, 22.2% were incorrect and involved information errors or reasoning errors. Generated URL links were often incorrect or unavailable (191/274, 69.7%), and some authors could not be found (59/350, 16.9%). Consequently, while ChatGPT did provide paragraphs to some of the paragraph-requiring prompts, it was not possible

to evaluate their veracity because most of the reference links were unavailable. This poor performance demonstrates a weakness of ChatGPT as a standalone self-learning tool. In the medical domain, it is crucial to ensure that information is accurate, as errors or inaccuracies can have detrimental consequences [27]. However, any inaccuracies and misinformation in self-study guides cannot be corrected without references, which would lead to erroneous absorption of information that can negatively impact learning outcomes. The absence of appropriate references may also deprive the students of access to additional information, which in turn could limit their comprehension and understanding related to the subject matter [28].

Despite its limitations, ChatGPT could still be useful as a self-study tool when used under supervision [29]. As a part of preparing students for the challenges in the future, they could be trained to critically evaluate and challenge factually incorrect or misleading responses from ChatGPT, such as tracing evidence to its primary sources to verify the model's assertions [28]. For instance, students can ask ChatGPT for its chain of thoughts through prompts such as "Explain your reasoning process for the answer" or "Explain your chain of thoughts for the answer" [30]. The responses to these prompts can help students understand ChatGPT's reasoning, think critically about the underlying information, and develop their own reasoning and critical thinking skills based on the experience. ChatGPT can be an engaging way of learning, but it is important to use it in moderation and not let it replace independent thinking. Students should be cautioned against overreliance on ChatGPT, as it could impair their higher-order cognitive skills, such as creativity and problem-solving [31].

This study contributed to the previous literature by providing evidence that the current version of ChatGPT is not suitable as a standalone self-learning tool and exploring the potential for supervised use of ChatGPT.

However, this study also has several limitations. Firstly, the study employed only 78 test items derived from the K-CBMSE pharmacology. While the sample size is adequate for the purposes of this study, it is still relatively small and may not fully represent all categories

of medical examination questions. As a result, future research may seek to utilize a larger and more diverse set of medical examination questions for a more comprehensive evaluation of ChatGPT's capabilities. Secondly, this study's primary focus was centered on examining ChatGPT's capacity to address medical examination MCQs, specifically those pertaining to pharmacology. The outcomes of this research may not necessarily be generalizable to other types of inquiries or domains. To enhance the transferability of the study's results, subsequent investigations may explore ChatGPT's efficacy in answering questions in fields other than medicine or in other answer formats such as essays. This approach would aid in establishing the generalizability of the findings and providing more robust support for future practical applications. Thirdly, ChatGPT is rapidly evolving. Significant advancements have occurred during the research process, which could potentially make some findings less relevant. For example, GPT-4 was released while this research was underway, and it is known to be significantly more powerful than ChatGPT [32]. Fourthly, overall performance of ChatGPT may have been overestimated in this study due to the imbalanced distribution of cognitive taxonomy levels in the test items. Only 17.9% of the test items are problem-solving, while 56.4% are recall. To ensure fair comparison across studies, the distribution of cognitive taxonomy levels should be standardized. Finally, there are several key components that can contribute to the effectiveness of learning tools, such as students' perception and interaction [33, 34]. This study did not assess ChatGPT's efficacy on these dimensions.

The introduction of new technologies such as internet, mobile devices, and ChatGPT in education presents both opportunities and threats. The introduction of new technologies such as internet, mobile devices, and ChatGPT presents both opportunities and threats in education. Artificial intelligence (AI) technology has the potential to revolutionize education [35], offering personalized virtual assistants and adaptive learning experiences for every student [31, 36, 37]. AI-powered systems can provide timely and immediate feedback, tailored recommendations, and interactive and engaging learning activities [26, 36, 38]. Although some may fear the threats of plagiarism and misinformation posed by ChatGPT, efforts to ban emerging technologies in higher education have been futile historically. ChatGPT is unlikely to be an exception [39].

Instead, we should embrace ChatGPT and other language models as self-learning tools while striving to minimize the associated risks. One possible approach is to develop strategies for appropriate supervision. For example, students can ask ChatGPT to generate a solution to a complex problem, and then evaluate the solution to

determine its feasibility or effectiveness [30]. Such setups would require the students to use their problem-solving skills and to think critically about the different factors involved in the problem. By doing so, we can stimulate students' learning and motivate them to develop higher cognitive skills such as critical thinking and problem-solving. Empirical studies also should be performed to investigate whether using ChatGPT with supervision can truly improve critical thinking and problem-solving skills.

Conclusion

The current version of ChatGPT has limitations as a useful self-study tool despite its performance in correctly answering MCQs. The answers could be inconsistent when the same inquiry is repeated; the generated rationale could be incorrect; and the generated references were nonsensical. To maximize the potential benefits of AI technology while minimizing its risks, it is imperative to develop effective supervision and moderation strategies.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12909-023-04832-x>.

Additional file 1: Supplement 1. Korean Comprehensive Basic Medical Sciences Examination (K-CBMSE). **Supplement 2.** ChatGPT's accuracy across the prompts. **Supplement 3.** The cases of incorrect responses.

Acknowledgements

WC is grateful to the Medical Education Assessment Corporation in Korea (MEAC) for providing the K-CBMSE test items and quality control data for this study. MEAC has kindly granted permission for some of the K-CBMSE test items to be included in this research manuscript. WC also thanks to Chungbuk National University for its support.

Author's contributions

WC designed and performed the current study including the translation of test items in Korean into English and statistical analysis. WC prepared the manuscript for publication.

Funding

N/A

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

N/A

The current study does not involve human subjects.

Consent for publication

N/A

The current study does not involve human subjects or private information.

Competing interests

The authors declare no competing interests.

Received: 2 March 2023 Accepted: 1 November 2023
Published online: 13 November 2023

References

1. Introducing ChatGPT. OpenAI. <https://openai.com/blog/chatgpt>. Accessed 30 Aug 2023.
2. What is the size of the training set for GPT-3?. OpenAI. <https://community.openai.com/t/what-is-the-size-of-the-training-set-for-gpt-3/360896>. Accessed 30 Aug 2023.
3. Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, Chi E, Le Q, Zhou D. Chain-of-thought prompting elicits reasoning in large language models. arXiv preprint arXiv:2201.11903v6. 2022. Accessed 30 Aug 2023.
4. Liévin V, Hother CE, Winther O. Can large language models reason about medical questions? arXiv preprint arXiv:2207.08143v3. 2022. Accessed 30 Aug 2023.
5. Wang B, Min S, Deng X, Shen J, Wu Y, Zettlemoyer L, Sun H. Towards understanding chain-of-thought prompting: an empirical study of what matters. arXiv preprint arXiv:2212.10001v2. 2023. Accessed 30 Aug 2023.
6. Patel D, Raut G, Zimlichman E, Cheetirala SN, Nadkarni G, Glicksberg BS, Freeman R, Timsina P, Klang E. The limits of prompt engineering in medical problem-solving: a comparative analysis with ChatGPT on calculation based USMLE medical questions. medRxiv. 2023. <https://doi.org/10.1101/2023.08.06.23293710>. Accessed 30 Aug 2023.
7. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, Chartash D. How does ChatGPT perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment. *JMIR Med Educ*. 2023;9:e45312. <https://doi.org/10.2196/45312>.
8. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, Madriaga M, Aggabao R, Diaz-Candido G, Maningo J, Tseng V. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLoS Digit Health*. 2023;2(2):e0000198. <https://doi.org/10.1371/journal.pdig.0000198>.
9. Subramani M, Jaleel I, Krishna MS. Evaluating the performance of ChatGPT in medical physiology university examination of phase I MBBS. *Adv Physiol Educ*. 2023;47(2):270–1.
10. Birkett L, Fowler T, Pullen S. Performance of ChatGPT on a primary FRCA multiple choice question bank. *Br J Anaesth*. 2023;131(2):e34–5. <https://doi.org/10.1016/j.bja.2023.04.025>.
11. Shay D, Kumar B, Bellamy D, Palepu A, Dershwitz M, Walz JM, Schaefer MS, Beam A. Assessment of ChatGPT success with specialty medical knowledge using anesthesiology board examination practice questions. *Br J Anaesth*. 2023;131(2):e31–4. <https://doi.org/10.1016/j.bja.2023.04.017>.
12. Mihalache A, Popovic MM, Muni RH. Performance of an artificial intelligence chatbot in ophthalmic knowledge assessment. *JAMA Ophthalmol*. 2023;141(6):589–97. <https://doi.org/10.1001/jamaophthalmol.2023.1144>.
13. 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):1.
14. Levin G, Horesh N, Brezinov Y, Meyer R. Performance of ChatGPT in medical examinations: a systematic review and a meta-analysis. *BJOG*. 2023. <https://doi.org/10.1111/1471-0528.17641>.
15. Amin Z, Eng KH. Chapter 31. Multiple choice question. In: *Basics in medical education*. Singapore: World Scientific Publishing; 2003. p. 283–298.
16. Buckwalter JA, Schumacher R, Albright JP, Cooper RR. Use of an educational taxonomy for evaluation of cognitive performance. *Acad Med*. 1981;56(2):115–21.
17. Lee YM, So YH, Ahn DS, Rhee KJ, Im H. Psychometric analysis of comprehensive basic medical sciences examination. *Korean J Med Educ*. 2002;14(2):301–6. <https://doi.org/10.3946/kjme.2002.14.2.301>. (In Korean).
18. Gaikwad N, Tankhiwale S. Crossword puzzles: self-learning tool in pharmacology. *Perspect Med Educ*. 2012;1(5–6):237–48. <https://doi.org/10.1007/s40037-012-0033-0>.
19. Nisar S, Aslam MS. Is ChatGPT a good tool for T&CM students in studying pharmacology? 2023. <https://doi.org/10.2139/ssrn.4324310>. Accessed 30 Aug 2023.
20. Eager B. Achieving better results from ChatGPT using incremental prompting. <https://broneager.com/incremental-prompting-ai-chatgpt>. Accessed 30 Aug 2023.
21. Duong D, Solomon BD. Analysis of large-language model versus human performance for genetics questions. *Eur J Hum Genet*. 2023. <https://doi.org/10.1038/s41431-023-01396-8>.
22. Limitations. OpenAI. <https://openai.com/blog/chatgpt>. Accessed 30 Aug 2023.
23. Morreel S, Mathysen D, Verhoeven V. Aye, AI! ChatGPT passes multiple-choice family medicine exam. *Med Teach*. 2023;45(6):665–6. <https://doi.org/10.1080/0142159X.2023.2187684>.
24. Alkaissi H, McFarlane SI. Artificial hallucinations in ChatGPT: implications in scientific writing. *Cureus*. 2023;15(2):e35179. <https://doi.org/10.7759/cureus.35179>.
25. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural language generation. *ACM Comput Surv*. 2022. <https://doi.org/10.1145/3571730>.
26. Abd-Alrazaq A, AlSaad R, Alhuwail D, Ahmed A, Healy PM, Latifi S, et al. Large language models in medical education: opportunities, challenges, and future directions. *JMIR Med Educ*. 2023;9:e48291. <https://doi.org/10.2196/48291>.
27. Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. *J Med Internet Res*. 2023;25:e48568. <https://doi.org/10.2196/48568>.
28. Boscardin C, Gin B, Black Golde P, Hauer KE. ChatGPT and generative artificial intelligence for medical education: potential impact and opportunity. *Acad Med*. 2023. <https://doi.org/10.1097/ACM.0000000000005439>.
29. Bair H, Norden J. Large language models and their implications on medical education. *Acad Med*. 2023;98(8):869–70. <https://doi.org/10.1097/ACM.0000000000005265>.
30. Ahn S. The impending impacts of large language models on medical education. *Korean J Med Educ*. 2023;35(1):103–7. <https://doi.org/10.3946/kjme.2023.253>.
31. Farrokhnia M, Banihashem SK, Noroozi O, Wals A. A SWOT analysis of ChatGPT: implications for educational practice and research. *Innov Educ Teach Int*. 2023;1–10. <https://doi.org/10.1080/14703297.2023.2195846>.
32. Ali R, Tang OY, Connolly ID, Zadnik Sullivan PL, Shin JH, Fridley JS, et al. Performance of ChatGPT and GPT-4 on neurosurgery written board examinations. *Neurosurgery*. 2023. <https://doi.org/10.1227/neu.00000000002632>.
33. Shoufan A. Exploring students' perceptions of ChatGPT: thematic analysis and follow-up survey. *IEEE Access*. 2023;11:38805–18. <https://doi.org/10.1109/ACCESS.2023.3268224>.
34. Haglund JH. Students acceptance and use of ChatGPT in academic settings [master's thesis]. Uppsala University; 2023. p. 47. <https://uu.diva-portal.org/smash/get/diva2:1772274/FULLTEXT01.pdf>. Accessed 30 Aug 2023.
35. Cardona MA, Rodríguez RJ, Ishmael K. Artificial intelligence and the future of teaching and learning. <https://www2.ed.gov/documents/ai-report/ai-report.pdf>. Accessed 30 Aug 2023.
36. Lee H. The rise of ChatGPT: exploring its potential in medical education. *Anat Sci Educ*. 2023. <https://doi.org/10.1002/ase.2270>.
37. Qadir J. Engineering education in the era of ChatGPT: Promise and pitfalls of generative AI for education. *TechRxiv*. 2022. <https://doi.org/10.36227/techrxiv.21789434.v1>.
38. Klutka J, Ackerly N, Magda AJ. Artificial intelligence in higher education: current uses and future applications. 2018. <https://universityservices.wiley.com/wp-content/uploads/2020/12/201811-AI-in-Higher-Education-TLH-with-new-bage.pdf>. Accessed 30 Aug 2023.
39. Saunders S. Rather than ban generative AI, universities must learn from the past. *University World News*; 2023. <https://www.universityworldnews.com/post.php?story=20230221075136410>. Accessed 30 Aug 2023.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.