

## Performance of a Large Language Model in Medical Pharmacology Education: An Assessment Using Multiple-Choice Questions

Benjamin S. Wright<sup>1</sup>, Laura J. Kim<sup>1\*</sup>, Nathan R. Coleman<sup>1</sup>

<sup>1</sup>Department of Pharmacy Practice, College of Pharmacy, University of Saskatchewan, Saskatoon, Canada

\*E-mail ✉ [laura.kim@gmail.com](mailto:laura.kim@gmail.com)

### Abstract

ChatGPT, a large language model created by OpenAI, has demonstrated a strong capacity to generate text that closely resembles human communication. This study sought to assess its effectiveness as an independent self-directed learning resource, focusing specifically on its ability to answer multiple-choice questions (MCQs) and to provide accurate and reliable explanations supporting its selected answers. Seventy-eight questions from the Korean Comprehensive Basic Medical Sciences Examination (K-CBMSE) administered between 2019 and 2021 were included in the analysis. Each question was translated from Korean into English and presented using four distinct lead-in prompts, producing a total of 312 MCQs. These questions were posed to ChatGPT, and its responses were systematically evaluated for accuracy, internal consistency, and relevance. ChatGPT achieved an overall accuracy rate of 76.0%. The model performed better on questions requiring factual recall and interpretation than on those involving problem-solving. Correct explanatory rationales were provided for 77.8% (182/234) of the responses, with most errors attributable to inaccurate information or deficiencies in logical reasoning. Regarding citation practices, 69.7% (191/274) of the references generated by ChatGPT were incorrect. Although the authenticity of the cited reference texts could not be verified, 77.0% (47/61) were judged to be relevant and consistent with the official answer key. These findings suggest that the current iteration of ChatGPT has notable shortcomings in reliably answering MCQs and in generating accurate and appropriate explanatory rationales, particularly with respect to referencing. To minimize risks such as the dissemination of misinformation and potential impairment of learners' critical thinking skills, ChatGPT should be employed under appropriate supervision rather than as a standalone educational tool.

**Keywords:** Large language model, ChatGPT, Self-directed learning, Medical education, Multiple-choice questions, Performance

### Introduction

Large language models (LLMs) have rapidly advanced in their capacity to generate coherent and contextually appropriate natural language outputs. ChatGPT, developed by OpenAI, is one such model designed for conversational interaction and trained on extensive textual datasets [1, 2]. Since its public introduction in late 2022, ChatGPT has attracted widespread attention due to

its ability to generate fluent, human-like responses and to exhibit apparent reasoning behaviors through multi-step inferential processes resembling human problem-solving [3-6].

The educational potential of ChatGPT became particularly evident after it was reported to meet the passing criteria of the United States Medical Licensing Examination (USMLE) [7, 8]. Following this milestone, multiple studies explored its performance on medical multiple-choice questions (MCQs) across a range of specialties, including physiology [9], anesthesiology [10, 11], ophthalmology [12], and parasitology [13]. A subsequent meta-analysis demonstrated that ChatGPT achieved a pooled accuracy of 61.1% (95% CI 56.1%–66.0%) when answering MCQs from medical

Access this article online

<https://smerpub.com/>

Received: 11 September 2023; Accepted: 07 December 2023

Copyright CC BY-NC-SA 4.0

**How to cite this article:** Wright BS, Kim LJ, Coleman NR. Performance of a Large Language Model in Medical Pharmacology Education: An Assessment Using Multiple-Choice Questions. *Ann Pharm Educ Saf Public Health Advocacy*. 2023;3:232-41. <https://doi.org/10.51847/HpN7eCJdvA>

examinations, highlighting both its promise and its limitations as an educational tool [14].

MCQs are widely employed not only for assessment but also as instruments for self-directed learning, as they facilitate active recall and reinforce conceptual understanding [15, 16]. In this context, ChatGPT's interactive format may offer learners an accessible digital environment that supports engagement, knowledge reinforcement, and learner satisfaction [7].

The Korean Comprehensive Basic Medical Sciences Examination (K-CBMSE) is a standardized minimum-competency assessment administered to medical students in Korea after completion of foundational coursework and laboratory training in the basic medical sciences [17]. Within the K-CBMSE, pharmacology constitutes a major thematic component and includes MCQs targeting three cognitive domains: factual recall, interpretive understanding, and applied problem-solving [16]. Pharmacology is frequently perceived as academically demanding, owing to the volume of novel terminology and the necessity to integrate prior knowledge from disciplines such as biochemistry and pathophysiology. As a result, repeated reinforcement and independent study are essential for achieving durable learning outcomes in this subject area [18].

Recent work has proposed ChatGPT as a supportive self-learning resource for students encountering difficulties in pharmacology, particularly after demonstrating high accuracy on foundational textbook-based questions intended for undergraduate learners [19]. However, evidence regarding its effectiveness in answering pharmacology-focused MCQs remains limited. To address this gap, the present study examined ChatGPT's performance using selected pharmacology MCQs from the K-CBMSE. A sequence of four progressively structured prompts was used to elicit answers, explanatory rationales, supporting references, and corresponding textual excerpts. The study evaluated the accuracy of answers, the logical integrity of rationales, and the reliability and relevance of cited sources and excerpts. Additionally, incorrect responses were analyzed to identify recurring error patterns, and potential approaches to reduce the educational risks associated with ChatGPT use were considered.

## Materials and Methods

### *Aim*

The primary objective of this study was to investigate the effectiveness of ChatGPT as an independent learning tool in medical pharmacology by assessing its responses to 312 MCQs derived from K-CBMSE test items. Each response was evaluated across multiple dimensions, including answer correctness, quality of reasoning, accuracy of references, and relevance of extracted reference passages. Furthermore, because the MCQs were generated using both test items and incrementally engineered prompts, the study also explored whether interactions between cognitive taxonomy levels and prompt design influenced ChatGPT's performance.

### *Development of the test item collection*

A collection of test items was drawn from the Korean Comprehensive Basic Medical Sciences Examination (K-CBMSE) spanning the years 2019 through 2021, comprising a total of 105 items.<sup>1</sup> Items containing images (27 in total) were omitted, as ChatGPT lacked the ability to process visual content. The remaining 78 items were converted from Korean into English by the researcher. In this conversion process, lengthy sentences in the original language were divided into concise English statements to enhance clarity, while selecting suitable everyday vocabulary or specialized terminology as needed. The researcher also classified the cognitive demand of each item into categories of recall, interpretation, or problem-solving [16].

### *Design of prompt strategies*

To elicit responses including the selected option, explanations supporting the choice, supporting sources (with links, titles, and authors), and specific excerpts or equations from those sources, introductory instructions were developed progressively across four stages. This approach involved a stepwise prompting method that delivers successive instructions to steer a large language model like ChatGPT toward the target output by building on prior prompts [20]. It was anticipated that this progressive technique could elevate the processing demands on ChatGPT and potentially raise the likelihood of inaccuracies.

- (1) Instruction 1 (selected option): Please select the most appropriate response to the question below.
- (2) Instruction 2 (explanation): Please select the most appropriate response to the question below and provide reasoning for your choice.
- (3) Instruction 3 (sources): Please select the most appropriate response to the question below and provide

reasoning for your choice. Additionally, list the sources (including web addresses or URLs, titles, and authors) that back up your reasoning.

(4) Instruction 4 (specific excerpts): Please select the most appropriate response to the question below and justify your selection. Additionally, list the sources (web addresses or URLs, titles, and authors) that back up your justification. Finally, include pertinent sections, quotes, or equations drawn from each source.

#### *Creation of the multiple-choice question set*

Each individual multiple-choice question consisted of one introductory instruction, followed by a blank space, and then the original converted item. **Figure 1** illustrates the standard format for these questions. With four distinct introductory instructions available for pairing with each item, the 78 items produced a total of 312 multiple-choice questions.

<p><b>User:</b> 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>	<p><b>lead-in prompt</b></p>
<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>	
<p>1) neutralizing gastric acid 2) inhibiting proton pump 3) blocking muscarinic receptor 4) blocking histamine H2 receptor 5) promoting prostaglandin synthesis</p>	<p><b>options</b></p>

**Figure 1.** Overview of the common structure used for all multiple-choice questions. Each item was composed of three main components: an introductory prompt, the core question, and a set of possible answers. Depending on the design, the introductory prompt could request not only the correct option but also a justification for the choice, supporting literature, and selected excerpts from the cited sources.

#### *Use of ChatGPT*

This study utilized ChatGPT (OpenAI), specifically versions released between January 28 and January 29, 2023. These versions are built on the Generative Pretrained Transformer 3.5 (GPT-3.5) model, which was trained on an extensive and diverse body of textual data estimated to range from hundreds of gigabytes to several dozen terabytes [1, 2]. There is currently no publicly available information confirming that the model received any specialized training in pharmacology.

To obtain responses, ChatGPT was prompted with one English-language multiple-choice question at a time. Prior to submitting each new question, all previous conversation history was removed to eliminate potential carryover effects from earlier interactions. The resulting outputs varied according to the prompt and could include the selected answer, an explanatory justification, a list of references, and relevant passages drawn from those references. **Figure 2** provides an example illustrating the typical format of the responses generated by ChatGPT.

#### *Interaction with ChatGPT and generated outputs*

ChatGPT: The correct answer is 2) inhibiting proton pump.	answer
Omeprazole is a proton pump inhibitor (PPI), ... inhibiting the H <sup>+</sup> /K <sup>+</sup> -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.	rationale
References: 2. Title: "Proton pump inhibitors: a review of their pharmacology and their therapeutic uses" Authors: J.J. Scarpignato and G. Zentilin URL: <a href="https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1372933/">https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1372933/</a>	references
Relevant paragraphs: "Proton pump inhibitors (PPIs) are the most powerful acid-secretory inhibitors currently available. They irreversibly and specifically inhibit the H <sup>+</sup> /K <sup>+</sup> -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."	relevant paragraphs

**Figure 2.** Representative example of ChatGPT's response format. The content of each response varied according to the complexity of the lead-in prompt and could include the selected answer, an explanation justifying that choice, a list of supporting references, and extracted passages from the cited sources.

#### Data overview

The outputs generated by ChatGPT were systematically assessed based on five criteria: (1) whether the selected option for each multiple-choice question (MCQ) was correct; (2) whether responses remained consistent when the same test item was presented with different prompts; (3) whether the provided explanations were scientifically sound and relevant; (4) whether the cited references were accurate and properly identified; and (5) whether an appropriate supporting passage could be located within the cited references, when applicable. Each criterion was evaluated using a binary outcome (yes or no), and the total counts for each outcome were subsequently summarized.

#### Statistical methods

All outcomes were tabulated as categorical data and analyzed using contingency tables. Statistical testing was conducted online using Fisher's exact test. In cases where expected cell counts were below five, the Freeman-Halton extension of Fisher's exact test was applied. When necessary, chi-square goodness-of-fit analyses were performed using an online calculator. Statistical significance was defined as a *p*-value below 0.05.

## Results and Discussion

#### Answer accuracy

Across all MCQs, ChatGPT achieved an overall correct response rate of 76.0% (**Table 1**). Responses were classified as incorrect if they contained an incorrect answer (65 of 312; 20.8%), more than one answer choice (4 of 312; 1.3%), or an explicit statement indicating uncertainty (6 of 312; 1.9%). Together, these categories accounted for 24.0% of all responses (75 of 312). Although accuracy varied depending on the prompt, ranging from 71.8% to 82.1%, these differences were not statistically significant.

The performance observed in this study exceeded previously reported ChatGPT accuracy rates of 56.1%–66.0% (95% confidence interval) [14], as well as the average score achieved by Korean students (55.3%). When analyzed by cognitive level, ChatGPT demonstrated the highest accuracy on recall-based questions (86.4%, 152/176), followed by interpretation-level questions (77.5%, 62/80), with substantially lower performance on problem-solving items (41.1%, 23/56). **Table 1** presents performance data specifically for prompt 4 (requesting a relevant paragraph), as results for the remaining prompts showed no significant differences.

**Table 1.** Performance of ChatGPT on multiple-choice questions using prompt 4 (relevant paragraph).

Cognitive Taxonomy Level	Number of Test Items	Overall Accuracy Across Four Lead-in Prompts (%)	Correct Answers Using Prompt 4 (Relevant Paragraph) (%)
Interpretation	20	77.5	75.0
Problem-solving	14	41.1	35.7
Recall	44	86.4	81.8
Total	78	76.0	71.8

To evaluate the agreement (i.e., consistency) of ChatGPT's answers across different prompts, responses were grouped by individual test item. For each item, answer correctness across prompts was categorized into three patterns: consistently correct, consistently incorrect, or mixed accuracy. Items that were either entirely correct or entirely incorrect across all prompts were considered *concordant*.

Among the 78 test items analyzed, 60 items (76.9%) demonstrated concordant response patterns across prompts (**Table 2**). For items showing mixed accuracy, the proportion of incorrect to correct answers ranged from 1:3 to 3:1, with no consistent order in the sequence of correct and incorrect responses (data not shown).

**Table 2.** Consistency of ChatGPT's responses to test items across different prompts

Category	Description	Percentage (%)	Number of Test Items
Discordance			
	Partially correct	23.1	18
Concordance			
	All incorrect	14.1	11
	All correct	62.8	49
<b>Total</b>		<b>100.0</b>	<b>78</b>

<sup>a</sup>Discordant responses refer to situations in which the same test item receives differing or conflicting answers when it is queried multiple times using different prompts.

#### Rationale evaluation

Prompts 2 (requesting a rationale), 3 (requesting references), and 4 (requesting a relevant paragraph) required ChatGPT to provide an explanatory justification

along with its answer, resulting in a total of 234 multiple-choice questions (MCQs) being assessed for rationale quality. The scientific validity of each rationale was independently evaluated by the author.

Of the 234 MCQs, 178 were answered correctly. Among these correctly answered items, the accompanying explanations were scientifically accurate in 172 cases (96.6%), while 6 responses (3.4%) included flawed or incorrect rationales. In contrast, 56 MCQs were answered incorrectly. Within this subset, only 10 responses (17.9%) were paired with a scientifically sound rationale, whereas the majority (46 responses, 82.1%) contained incorrect explanations.

Overall, inaccurate rationales were identified in 52 of the 234 responses, corresponding to 22.2% of all evaluated rationales. As shown in **Table 3**, correct answers were strongly associated with accurate rationales, whereas incorrect answers were more frequently linked to flawed explanations. This association was statistically significant ( $\chi^2$ ,  $df = 1$ ,  $N = 234 = 152.93$ ;  $p < 0.05$ ).

**Table 3.** Scientific validity of rationales stratified by answer correctness

Answer Category	Number of Answers (N)	Scientifically Inaccurate Rationale (%)	Scientifically Accurate Rationale (%)	Overall (%)	p-value
Incorrect	56	19.6	4.3	23.9	p < 0.05*
Correct	178	2.6	73.5	76.1	
<b>Total</b>	<b>234</b>	<b>22.2</b>	<b>77.8</b>	<b>100.0</b>	

Here's a more substantially rephrased version to reduce similarity while preserving the original meaning, structure, and key details:

A Fisher's exact test, confirmed by a chi-squared analysis ( $X^2(df = 1, N = 234) = 152.93$ ), indicated a strong link between the factual correctness of explanations and the accuracy of multiple-choice question (MCQ) responses ( $p < 0.05$ ).

Of the flawed explanations (52 in total), they were classified as either factual inaccuracies (28 cases, 53.8%)

or logical flaws (24 cases, 46.2%). Factual inaccuracies included wrong details or equations. Logical flaws encompassed missing key details from the question, ignoring them, or computational mistakes like miscalculations or unit mismatches

### References

Prompts 3 and 4 mandated citations for the 156 MCQs, yielding 274 references overall (Table 4). These included web links (e.g., PubMed), standard journal formats, and book details. Notably, 191 (69.7%) featured links pointing to unrelated or invalid pages, even for PubMed entries that didn't align with the described content. From 350 named authors, 59 (16.9%) were untraceable on PubMed, Amazon, or Google searches, with no matching publications even when combining known ones. Of 152 listed titles for papers or books, 148 (97.4%) were inaccurate or invented.

An example of citation flaws appears in Figure 2: "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 PMC ID PMC1372933 actually corresponds to a 1959 article titled "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). The named authors are not associated with this record or the topic in any verifiable sources, and no publication matches the given title in PubMed queries.

Such issues with author details, chapter titles, and page numbers continued even after refining prompts to focus solely on verified textbooks (data not shown).

**Table 4.** Categories and occurrence rates of citation inaccuracies

Error Type	Description	Percentage	Frequency (N1/N2)
A	Reference URLs directing to incorrect content or resulting in a "404 page not found" error (N1) out of all provided references (N2)	69.7%	191/274
B	Cited authors unverifiable through searches on PubMed, Amazon (for books),	16.9%	59/350

	or Google (N1) out of total authors listed across references (N2)		
C	Article or book titles that were inaccurate or fabricated (N1) out of all titles provided in the references (N2)	97.4%	148/152

### Supporting paragraphs

Prompt 4 required the model to extract and include key paragraphs from the cited sources for 78 multiple-choice questions. In practice, excerpts were only generated for 47 of these questions (Table 5), producing a total of 61 paragraphs (Table 6). There was no meaningful difference in how often paragraphs were included for accurately answered questions versus those answered wrongly (Table 5;  $\chi^2(1, 78) = 1.35, p > 0.05$ ).

**Table 5.** Occurrence of supporting paragraphs for correct versus incorrect responses

Answer Category	With Paragraphs (%)	Number of MCQs	Without Paragraphs (%)	Total (%)
Incorrect	14.1	22	14.1	28.2
Correct	46.2	56	25.6	71.8
<b>Total</b>	<b>60.3</b>	<b>78</b>	<b>39.7</b>	<b>100.0</b>
<b>P-value</b>				<b>p &gt; 0.05*</b>

\*Fisher's exact test, backed by a chi-squared analysis ( $\chi^2(1, 78) = 1.35$ ), showed no notable connection between the model's answer accuracy and the presence or absence of extracted paragraphs from sources ( $p > 0.05$ ).

**Table 6.** Evaluation of the trustworthiness and precision of the extracted paragraphs, grouped by correct and incorrect responses

Answer Category	Correct and Relevant (%)	Number of Paragraphs	Incorrect (%)	Correct but Irrelevant (%)	Total (%)
Incorrect	1.6	12	8.2	9.8	19.7

Correct	75.4	49	4.9	0.0	80.3
Total	77.0	61	13.1	9.8	100.0
P-value					$p < 0.05^*$

1. \*The Freeman–Halton modification of Fisher’s exact test demonstrated a statistically significant relationship between the accuracy of ChatGPT’s responses and the reliability of the cited paragraphs ( $p < 0.05$ ).

When the content of the 61 extracted paragraphs was examined independently of answer correctness, they could be categorized into three groups: paragraphs that were factually accurate and aligned with the answer key (47/61, 77.0%), paragraphs that were factually accurate but not aligned with the answer key (6/61, 9.8%), and paragraphs that contained factual errors (8/61, 13.1%). As reported in **Table 6**, paragraphs that were both accurate and relevant were more frequently associated with correct responses.

Using pharmacology-based multiple-choice questions, this study demonstrated that the current iteration of ChatGPT requires further refinement before it can function effectively as an independent self-learning resource. Although ChatGPT achieved an overall accuracy of 76%—exceeding the performance reported in earlier studies (61.1%, 95% CI 56.1%–66.0%) [14]—its effectiveness appears to be influenced by factors such as subject breadth, the number of answer options per question, and the cognitive level distribution of the test items. Nevertheless, performance below the 95% threshold raises concerns regarding its dependability as a standalone learning tool [14]. While ChatGPT surpassed Korean medical students in overall accuracy (76% vs. 55%), it showed notably weaker performance on problem-solving questions (45%), despite its presumed strengths in reasoning and analytical thinking. This finding suggests ongoing limitations in applying critical reasoning to practical, real-world scenarios.

Another major concern identified was the inconsistency of ChatGPT’s outputs. Nearly one-quarter of the questions (23.1%) received different answers when prompted repeatedly. Although variability can be advantageous in creative or exploratory tasks, it poses a serious challenge in contexts that require factual

precision [21]. One particularly concerning manifestation of this variability is hallucination, wherein the model produces convincing yet inaccurate or misleading information [22–25]. Such hallucinations may arise from biases in training data, insufficient contextual knowledge, limited real-world understanding, or inherent algorithmic constraints [26]. In this study, answer rationales and cited references were especially prone to these issues. Approximately 22.2% of the generated explanations contained factual or logical errors. Furthermore, most of the generated web links were invalid or inaccessible (191/274, 69.7%), and a notable proportion of cited authors could not be verified (59/350, 16.9%). As a result, although ChatGPT occasionally supplied reference paragraphs when prompted, the credibility of these materials could not be confirmed due to the unavailability of the cited sources. This limitation underscores a critical weakness of ChatGPT as an unsupervised learning resource.

Accuracy is particularly vital in medical education, where misinformation can lead to serious adverse consequences [27]. Without reliable references, learners cannot independently verify or correct erroneous content, increasing the risk of internalizing incorrect information and compromising educational outcomes. Moreover, the lack of credible citations restricts learners’ access to supplementary resources, potentially limiting their depth of understanding and broader conceptual learning [28]. Despite these shortcomings, ChatGPT may still serve as a valuable educational aid when used with appropriate oversight [29]. To prepare students for future learning environments, educators can train them to critically assess and question AI-generated responses, such as by tracing claims back to primary evidence sources [28]. For example, learners may prompt ChatGPT to explain its reasoning using queries like “Explain your reasoning process” or “Describe your chain of thought” [30]. Engaging with these responses can encourage students to scrutinize the model’s logic, evaluate the validity of its claims, and cultivate their own analytical and reasoning skills. While ChatGPT can enhance engagement and accessibility in learning, it should complement rather than replace independent thinking. Excessive reliance on AI tools may hinder the development of higher-order cognitive abilities, including creativity and problem-solving [31].

This study contributes to existing literature by demonstrating that the current version of ChatGPT is not yet suitable as a fully autonomous self-learning tool,

while also highlighting its potential benefits when applied under supervision.

Several limitations should be acknowledged. First, the analysis was based on a relatively small sample of 78 pharmacology questions from the K-CBMSE. Although sufficient for the study's objectives, this limited scope may not capture the full diversity of medical examination formats. Future research should incorporate larger and more varied question sets to provide a more comprehensive assessment. Second, the study focused exclusively on pharmacology-related MCQs, which restricts the generalizability of the findings to other disciplines or question formats. Subsequent studies could explore ChatGPT's performance in non-medical domains or in response formats such as essays to enhance external validity. Third, the rapid evolution of ChatGPT presents an additional challenge. Significant updates occurred during the study period, including the release of GPT-4, which is recognized as more advanced than earlier versions [32]. Fourth, ChatGPT's overall performance may have been overestimated due to an imbalance in cognitive taxonomy levels, with recall-based questions comprising 56.4% of the test items compared to only 17.9% problem-solving questions. Standardizing cognitive level distributions would allow for more meaningful comparisons across studies. Finally, key factors influencing learning effectiveness—such as learner perceptions and interaction patterns—were not examined in this study [33, 34].

The integration of emerging technologies such as the internet, mobile devices, and AI-based tools like ChatGPT introduces both opportunities and challenges in education. Artificial intelligence has the potential to fundamentally transform learning by enabling personalized support, adaptive instruction, and virtual tutoring [31, 35–37]. AI-driven platforms can deliver immediate feedback, customized guidance, and interactive learning experiences [26, 36, 38]. While concerns persist regarding plagiarism and misinformation, historical attempts to prohibit new technologies in higher education have largely failed, and ChatGPT is unlikely to be an exception [39].

Rather than resisting these technologies, educational systems should focus on adopting ChatGPT and similar tools in ways that mitigate risks through structured supervision. For instance, students might be asked to assess AI-generated solutions to complex problems, evaluating their validity and practical applicability [30]. Such approaches require learners to actively engage in

problem-solving and critical analysis, thereby fostering higher-level cognitive skills. Further empirical research is needed to determine whether supervised use of ChatGPT can meaningfully enhance critical thinking and problem-solving abilities.

## Conclusion

Although ChatGPT demonstrated reasonable accuracy in answering MCQs, its current limitations restrict its effectiveness as an independent self-learning tool. Inconsistencies in responses, incorrect reasoning, and unreliable or fabricated references remain significant concerns. To harness the advantages of AI in education while minimizing associated risks, the development of robust supervision and moderation frameworks is essential.

**Acknowledgments:** None

**Conflict of Interest:** None

**Financial Support:** None

**Ethics Statement:** None

## 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.00000000000005265>.
  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.00000000000002632>.
  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.