Evaluating the Cognitive Levels of Generative AI via Bloom's Taxonomy: A Cross-sectional Study

Kuan-Ju Huang, Cheng-Heng Liu, Chien-Chun Wu, Bor-Ching Sheu, Chih-Wei Yang

Submitted to: Journal of Medical Internet Research
on: May 15, 2024

Table of Contents

Evaluating the Cognitive Levels of Generative AI via Bloom’s Taxonomy: A Cross-sectional Study

Kuan-Ju Huang[1] MS; Cheng-Heng Liu[1] MD; Chien-Chun Wu[2] MSc; Bor-Ching Sheu[3] PhD; Chih-Wei Yang[1] PhD

[1]Department of Medical Education, National Taiwan University Hospital Taipei TW
[2]Institute for Interdisciplinary Studies, Brain and Cognitive Sciences, University of Amsterdam Amsterdam NL
[3]Department of Obstetrics and Gynecology, National Taiwan University Hospital Taipei TW

Corresponding Author:
Chih-Wei Yang PhD
Department of Medical Education, National Taiwan University Hospital
Department of Medical Education, National Taiwan University Hospital 7 Chung-Shan South Road
Taipei
TW

Abstract

Background: Generative AI has garnered awareness in the medical field, yet its potential is constrained by inherent limitations. By responding to inputs through predicting the next word from its memory-based archive, we aim to explore some of these constraints from a medical education and psychological perspective, utilizing Bloom’s taxonomy.

Objective: To assess AI's cognitive functions in the medical sector by examining its performance through medical licensing exams and applying Bloom's taxonomy.

Methods: Questions from the Taiwan Medical Licensing Examination (TMLE) (August 2022) and the third step of the United States Medical Licensing Examination (USMLE) (August 2022) were classified based on Bloom's taxonomy levels. The ChatGPT versions were tasked through individual prompts, with questions entered separately into ChatGPT-3.5 and ChatGPT-4 using different accounts. After each response, the chat logs were erased and reset to ensure the independence of each answer. Responses from ChatGPT-3.5 and ChatGPT-4, collected between January and February 2024, were analyzed. The questions from both exams were available online during the study period.

Results: Although the overall performance of ChatGPT-4 surpassed that of ChatGPT-3.5, the analysis of responses from both models across various cognitive levels revealed no significant correlation between their performance and the levels of Bloom's taxonomy. This lack of significance persisted even when considering the strength of ChatGPTs in their extensive databases classified under "remember," compared to other cognitive levels labeled as "non-remember."

Conclusions: In the medical field, ChatGPT models may utilize their "remember" function to answer all types of questions across all categories defined by Bloom's taxonomy. Further research is required focusing on different versions, medical specialties, and the level of difficulty assessed by individuals from various backgrounds.

(JMIR Preprints 15/05/2024:60565)
DOI: https://doi.org/10.2196/preprints.60565

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?
Please make my preprint PDF available to anyone at any time (recommended).
✓ **Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all u**
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.
2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?
✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Evaluating the Cognitive Levels of Generative AI via Bloom's Taxonomy: A Cross-sectional Study

Kuan-Ju Huang[1-3], Cheng-Heng Liu[4,5], Chien-Chun Wu[6], Bor-Ching Sheu[3], Chih-Wei Yang[4,5,7]

1. Department of Obstetrics and Gynecology, National Taiwan University Hospital, Yunlin Branch, Yunlin County, Taiwan
2. Graduate Institute of Clinical Medicine, College of Medicine, National Taiwan University, Taipei City, Taiwan
3. Department of Obstetrics and Gynecology, National Taiwan University Hospital, Taipei City, Taiwan
4. Department of Medical Education, National Taiwan University Hospital, Taipei, Taiwan
5. Department of Emergency Medicine, National Taiwan University Hospital, Taipei, Taiwan
6. Institute for Interdisciplinary Studies, Brain and Cognitive Sciences, University of Amsterdam, Amsterdam, The Netherlands
7. Department and Graduate Institute of Medical Education and Bioethics, College of Medicine, National Taiwan University, Taipei, Taiwan

Corresponding Author: Chih-Wei Yang

Department of Medical Education, National Taiwan University Hospital

7 Chung-Shan South Road, Taipei, Taiwan

Fax: +886-2-2311-4965

Tel: +886-2-23123456 ext. 261426

E-mail: cwyang100@ntu.edu.tw or cwyang0413@gmail.com

Manuscript word counts: 2299

Abstract

Background: Generative AI has garnered awareness in the medical field, yet its potential is constrained by inherent limitations. By responding to inputs through predicting the next word from its memory-based archive, we aim to explore some of these constraints from a medical education and psychological perspective, utilizing Bloom's taxonomy.

Objective: To assess AI's cognitive functions in the medical sector by examining its performance through medical licensing exams and applying Bloom's taxonomy.

Methods: Questions from the Taiwan Medical Licensing Examination (TMLE) (August 2022) and the third step of the United States Medical Licensing Examination (USMLE) (August 2022) were classified based on Bloom's taxonomy levels. The ChatGPT versions were tasked through individual prompts, with questions entered separately into ChatGPT-3.5 and ChatGPT-4 using different accounts. After each response, the chat logs were erased and reset to ensure the independence of each answer. Responses from ChatGPT-3.5 and ChatGPT-4, collected between January and February 2024, were analyzed. The questions from both exams were available online during the study period.

Results: Although the overall performance of ChatGPT-4 surpassed that of ChatGPT-3.5, the analysis of responses from both models across various cognitive levels revealed no significant correlation between their performance and the levels of Bloom's taxonomy. This lack of significance persisted even when considering the strength of ChatGPTs in their extensive databases classified under "remember," compared to other cognitive levels labeled as "non-remember."

Conclusions: In the medical field, ChatGPT models may utilize their "remember" function to answer all types of questions across all categories defined by Bloom's taxonomy. Further research is required focusing on different versions, medical specialties, and the level of difficulty assessed by individuals from various backgrounds.

Keywords: AI; artificial intelligence; Bloom taxonomy; ChatGPT; cognition; intuition; learning; theory of mind; thinking.

Introduction

Since its launch in late 2022, generative artificial intelligence (AI) has garnered significant interests due to its potential applications in medical education and routine clinical care[1, 2]. Upon evaluating the performance of generative AI in medical licensing exams across various specialties, it was revealed that generative AI has outperformed medical professionals, demonstrating superior outcomes[3]. Additionally, it has shown promising results in challenging or rare clinical cases[4-6]. However, recent meta-analyses have shown that the performance of representative generative AI models generally achieve results around a passing grade[7, 8]. Despite rigorous research and development efforts, the application of the cutting-edge instrument remains limited[2, 9, 10]. One of the significant challenges faced by AI is the occurrence of hallucinations, contributing to nonsensical or inaccurate outputs, necessitating thorough verification, especially in clinical settings[9, 11]. Through the utilization of rapid engineering, data training, or model building techniques, it is feasible to enhance AI's comprehension of inputs and generation of coherent outputs. Nonetheless, while simulating AI to mimic human behavior, it is crucial to examine whether machine learning processes resemble humans learning patterns[12].

Bloom's taxonomy is a cognitive framework that classify educational objectives into levels of complexity and specificity[13, 14]. It hypothesizes that humans learn hierarchically, progressing from "remember" and "understand" to "apply," "analyze," "evaluate," and "create" in the cognitive domain. While AI is believed to possess an endless capacity for memory within its training database, in contrast to the limited brain capacity of humans, its performance at higher levels of this hierarchy remains largely unexplored[12, 15]. This study aimed to assess whether the latest AI version is capable of cognitive processing and learning comparable to those of humans in the medical field, utilizing medical licensing exams and Bloom's taxonomy as benchmarks.

Methods

Recruitment

The study was conducted between January and February 2024. During the study period, two independent researchers (KJH and CWY) prospectively evaluated the cognitive levels of multiple-choice questions from the Taiwan Medical Licensing Examination (TMLE) (August 2022) and the third step of the United States Medical Licensing Examination (USMLE) (August 2022), both of which were accessible online. Question contents, including tables and laboratory results, were converted into text format (Multimedia Appendix 1). In addition, texts written in Traditional Chinese originally were translated into English prior to input. Also, questions with figures, and group questions were excluded. ChatGPT-3.5 and ChatGPT-4 were employed through distinct user accounts. Furthermore, each question initiated a new chat with ChatGPT, starting with a prompt, followed by the question and available choices. The responses and their explanations were then documented. The study was approved by National Taiwan University Hospital Ethics Center (202308059RINB).

Statistical Analysis

To assess the cognitive levels involved, we utilized Bloom's taxonomy for classification initially treating each level as a categorical variable. We analyzed the relationship between any cognitive level of a question and the accuracy of the corresponding answer. For instance, if answering a question correctly involved the domains "remember," "analyze," and "create" levels, all three levels were recorded. In cases of inconsistency between authors, all items were included in the analysis. Additionally, in order to mitigate this inconsistency, each level was further treated as a consecutive variable, with "remember" assigned a score of 1 and "create" a score of 6.

For each question, we employed the average of the highest levels for analysis. Given the AI's unrestricted "remember" capability and limited evidence at other levels, we distinguished between "remember" and the remaining five levels, terming them as "non-remember" (Multimedia Appendix 2). The data were presented in the form of counts and percentage, and logistic regression with multivariate analysis was employed to report odds ratios (OR) and their corresponding 95% confidence intervals (CI). A p-value of <0.05 was considered statistically significant. All statistical analyses were performed using Stata/SE 18.0 for Mac.

Results

The August 2022 editions of the TMLE and USMLE consisted of 80 and 137 questions, respectively. Questions that included figures or group questions—3 from the TMLE and 25 from the USMLE—were excluded from the analysis. These questions were categorized by their highest cognitive level, and there were 86 'Remember,' 10 'Understand,' 19 'Apply,' 22 'Analyze,' 21 'Evaluate,' and 31 'Create' questions." In addition, the study found an interobserver inconsistency of 5.82%, defined by a difference of ≥ 2 scores. ChatGPT-3.5 correctly answered 128 questions (67.72%), while ChatGPT-4 achieved 165 correct answers (87.30%). Figure 1 illustrates their performance based on Bloom's taxonomy.

Figure 1. ChatGPTs' general performance based on Bloom's taxonomy.

Upon categorization, the performance of ChatGPT models showed no correlation with the six levels of Bloom's taxonomy (Figure 2). After adjusting for the test's origin and the number of cognitive levels involved, the performance of ChatGPT models did not demonstrate any association with a particular cognitive level (ChatGPT-4, OR 0.95, 95% CI 0.63 – 1.42; ChatGPT-3.5, OR 1.01, 95% CI 0.74 – 1.36, Figure 2). Hence, the evidence does not support the hypothesis that the extensive database of ChatGPT models can improve performance in "remember" tasks compared to other cognitive levels termed as "non-remember" (ChatGPT-4, OR 0.77, 95% CI 0.31 – 1.93; ChatGPT-3.5, OR 0.93, 95% CI 0.48 – 1.83, Figure 2).

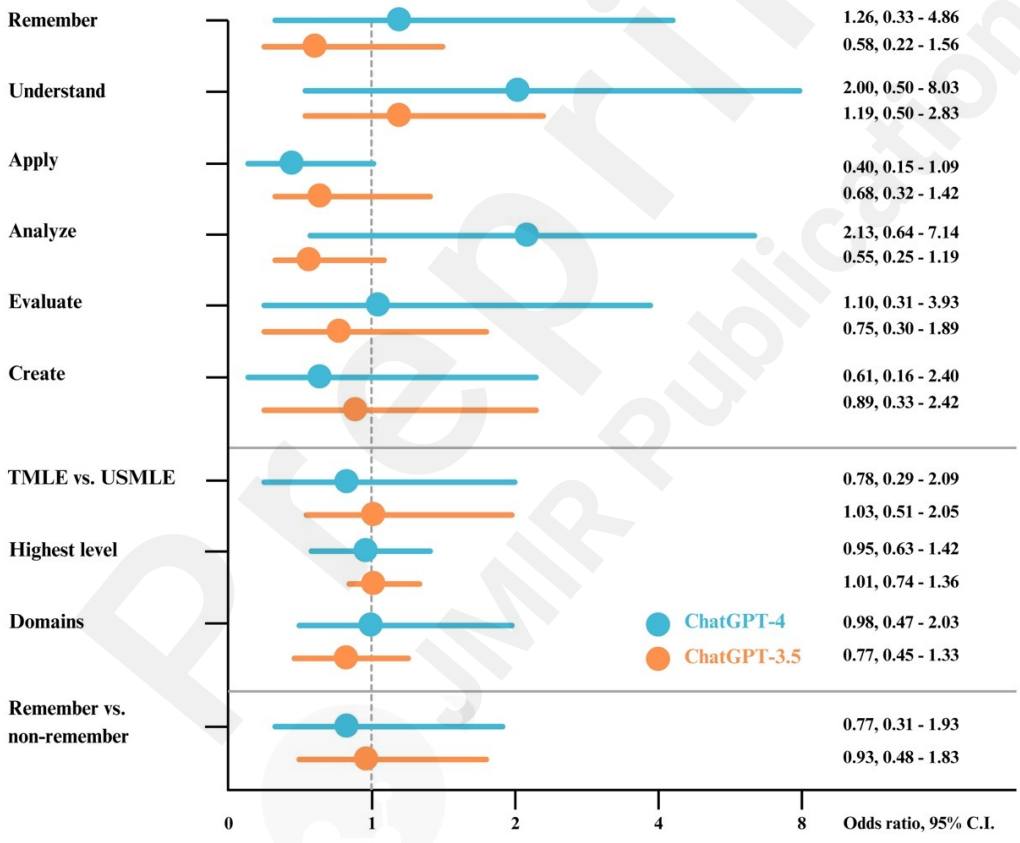

Figure 2 ChatGPTs' performance analyzed using six levels in Bloom's taxonomy.

Discussion

Principal Results

This study sheds light on the performance of AI deep-learning models in entry-level medical exams, from both medical education and psychology perspectives. Overall, ChatGPT-4 outperforms ChatGPT-3.5. However, the analysis reveals that both models effectively handled questions across all levels of Bloom's taxonomy in medical licensing exams, suggesting no direct correlation between their performance and cognitive levels required by these questions. Given the ongoing debate regarding the ability of ChatGPT models to replicate human thought processes, these findings offer valuable insights into the operational mechanisms of ChatGPTs through the inherent capacity for unlimited memory. Moreover, they illuminate their advancements in simulating human cognitive functions and the potential for further refining their application in the medical field[12].

Limitations

There are several limitations in interpreting the findings of this study. First, publicly available questions from medical licensing exams were utilized, raising uncertainty regarding the similarity of the knowledge level in the training database and the working models to other fields. Second, question difficulty was not factored into our analysis. Imbalanced difficulty distribution could potentially influence the results, with perception varying based on the knowledge level of the target audience. For example, a question might be easy for pulmonologists, moderate for medical students, but challenging for non-medical professionals, who could be part of the database training teams. Third, the specialty of these questions was not analyzed. Considering the hypothesis that AI has an unlimited memory capacity, it is inferred that its database covers all the required elements for a medical entrance exam. Additionally, our sample size might be constrained when divided by the diversity of medical specialties or subspecialties. Lastly, the current findings could be limited to the study period. With the emergence of new ChatGPT versions and the development of generative AI from other sources, the results of this study could evolve over time.

Comparison with Prior Work

Traditionally, the evaluation of learning objectives distinguishes between "remembering" and "understanding" as independent cognitive processes. "Remembering" entails the retention of knowledge, whereas "understanding" involves the application and utilization of memorized knowledge.

For instance, assessing knowledge about pneumonia with a statement such as “pneumonia is an acute infection of the pulmonary parenchyma” primarily targets memory, while asking to "differentiate between pneumonia and pneumonitis" requires "remembering," "understanding," "analyzing," and "evaluating" the two conditions. In contrast to humans, AI possesses a seemingly unlimited capacity for memory retention[15]. However, evidence supporting cognitive functions beyond "remembering" remains inconsistent. In the medical field, previous studies assessing ChatGPT using the USMLE and licensing exams for various specialties have shown that ChatGPTs possess knowledge at or above the passing level[7]. Conversely, ChatGPT models have surpassed medical professionals in addressing complex medical cases and challenging USMLE questions, which likely necessitate cognitive functions beyond “remembering[5, 16]."

ChatGPT models exhibited consistent performance across both foundational and advanced topics in studies spanning general medicine, nursing, radiology, and parasitology[17-21]. Intriguingly, a recent examination of ChatGPT's competency in a psychosomatic medicine exam uncovered that the majority of incorrect responses pertained to the "remember" and "understand" levels of Bloom's taxonomy[22]. This paradox highlights the models’ challenge in performing poorly at lower cognitive levels while proficiency at higher levels despite possessing knowledge merely at the passing-level[23].

In interpreting these findings, considering AI’s inherent capacity for unlimited memory, it seems plausible that the “remember” function facilitates its ability to address the majority of questions[12]. For example, in evaluating the “remember” level, if the requisite combination of elements is present in its database, AI can readily provide the most probable answers. When faced with complex questions that demand higher cognitive functions in humans, AI “remembers” how humans have previously solved similar questions, provided that there are comparable scenarios or key elements exist in its training database. As a result, the responses of AI appear to “analyze,” “evaluate,” and “create,” giving the impression that it has learned from human problem-solving techniques, rather than simply relying on foundational knowledge to formulate an appropriate answer. This apparent advanced cognitive function may mirror the model’s ability to extract and synthesize multiple human thought processes about a specific topic from its extensive database, thereby simulating the human cognitive process by predicting the next probable word based on these processed scripts.

This mechanism is analogous to the dual-process theory of human cognition, involving both

intuitive (System 1) and analytical (System 2) thinking[24, 25]. Generally, System 1 processes information quickly and intuitively, handling simple tasks such as memory recall. Conversely, System 2 engages in deliberate reasoning and manages complex tasks. ChatGPT models exhibit robust System 1 capabilities, with ongoing efforts aimed at enhancing their System 2 functions. Our hypothesis suggests that generative AI, leveraging foundational knowledge, might show a propensity for excelling in tasks that are suitable for System 1, while demonstrating comparatively weaker performance in tasks requiring System 2, due to inherent limitations in deep-learning models. However, this study did not identify a clear link between performance in tasks classified as "remember" (akin to System 1) versus "non-remember" (akin to System 2). Despite the lack of positive findings, a potential explanation exists for the observed results. As previously discussed, the human brain can streamline its response to complex questions through accumulated experience, effectively transitioning knowledge from System 2 to System 1. This process results in a cognitive state that could be referred to as "intuitive System 2," in which the brain efficiently employs memorized strategies to solve problems with minimal effort. This phenomenon could perhaps reflect the AI's ability to simulate higher cognitive functions by leveraging its vast database, suggesting a manifestation of "intuitive System 2" within AI models[12, 24, 25]. For instance, medical students typically rely on System 2 to differentiate between pneumonia and pneumonitis, which requires deliberate thought and analysis. In contrast, experienced pulmonologists may instinctively understand these differences and apply their “intuitive System 2” to generate rapid responses. ChatGPT models can simulate the process of transitioning from novice to expert by memorizing the problem-solving methods approached by medical students, and “learn” effectively to think like pulmonologists.

Nevertheless, if the database lacks precise medical knowledge on pneumonia and pneumonitis, is derived from non-expert sources, fails to accurately predict relevant information, or contains errors in data encoding, the AI may exhibit "hallucinations," encounter "blind spots," or creatively "think outside the box," deviating from conventional medical logic[1, 18, 26]. Insights into this dynamic are provided by a study evaluating ChatGPT models' advanced cognitive abilities in biology, utilizing both standard and hypothetical questions. The study found that ChatGPT models performed notably poorer on questions involving concepts of redefinition or invention[27]. Emerging evidence suggests that AI may exhibit its own cognitive process, as indirectly indicated by a trend of improved performance on questions at the lower levels of Bloom's taxonomy, particularly in disciplines such as neurology,

radiology, physiology, microbiology, and biochemistry[28-32]. In these investigations, the majority of questions assessed originated from internal materials or were inaccessible to ChatGPT models during the study periods. This restriction may have limited the models' ability to incorporate these questions into their database training, potentially affecting their ability to recall these specific questions and answers. Furthermore, compared to earlier generative AI models, Hagendorff et al. discovered that ChatGPT models exhibited enhanced responses and accuracy when presented with psychological questions related to human reasoning and decision-making, particularly with prompts suggesting the use of higher cognitive functions[12]. The variability in performance at higher cognitive levels in the aforementioned studies could imply the adoption of an "intuitive System 2" by the AI, or it may serve as evidence of the development of a nascent System 2 within the AI. In clinical counseling, ensuring accurate responses in complex decision-making is crucial. Before ChatGPTs can consistently produce logical and precise results, using customized ChatGPTs to restrict the input scripts to specific contents, such as definitions, guidelines, or real-time web access, and providing carefully designed prompts, followed by a review by medical professionals, could enhance the quality of answers, allow it to “think inside the box”, and reduce hallucination[1, 10, 16, 26, 33]. Technology is designed to fulfill human needs. Understanding cognitive functions provides us a new perspective on AI's psychological state, operational mechanisms, and limitations. As we develop technology that mimics human behavior and aids in human endeavors, it is our responsibility to guide generative AI in the medical field, with a focus on connecting people.

Conclusions

In the medical field, ongoing debate persists regarding AI’s capacity to mimic human thought processes. Prior to the arrival of the ideal generative AI, exploiting its inherent memory function could facilitate the development of customized applications for medical education or clinical practice, with guidance provided by specialized medical professionals. Further research is necessary for each version or major release of generative AI, assessing its performance across specialty hierarchies and adjusting for the diverse cognitive challenges encountered by individuals from various backgrounds.

Acknowledgment

The current study was supported by the National Taiwan University Hospital Yunlin Branch (113-S009) and National Taiwan University Hospital (113-S0265). This funding source had no role in the design of this study and had not any role during its execution, analyses, interpretation of the data, or decision to submit results.

Author contributions

KJH and CWY conceptualized the study. KJH designed the methodology and CWY analyzed the data. KJH drafted the original manuscript. CHL, CCW, and BCS reviewed and edited the manuscript. CCW visualized the final version of the study contents. CWY supervised the study.

Conflicts of Interest

None declared

Data availability

KJH and CWY have full access to the study data. Data is shared upon reasonable request.

Reference

1. Voelker R, Hswen Y. Clinical AI Tools Must Be Fed the Right Data, Stanford Health Care's Chief Data Scientist Says. JAMA. 2023 Dec 12;330(22):2137-9. PMID: 37966811. doi: 10.1001/jama.2023.19297.
2. Weidener L, Fischer M. Artificial Intelligence in Medicine: Cross-Sectional Study Among Medical Students on Application, Education, and Ethical Aspects. JMIR Med Educ. 2024 Jan 05;10:e51247. PMID: 38180787. doi: 10.2196/51247.
3. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023 Feb;2(2):e0000198. PMID: 36812645. doi: 10.1371/journal.pdig.0000198.
4. Kanjee Z, Crowe B, Rodman A. Accuracy of a Generative Artificial Intelligence Model in a Complex Diagnostic Challenge. JAMA. 2023 Jul 03;330(1):78-80. PMID: 37318797. doi: 10.1001/jama.2023.8288.
5. Eriksen AV, Möller S, Ryg J. Use of GPT-4 to Diagnose Complex Clinical Cases. NEJM AI. 2024;1(1):AIp2300031. doi: doi:10.1056/AIp2300031.
6. Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese Medical Licensing Examination: Comparison Study. JMIR Med Educ. 2023 Jun 29;9:e48002. PMID: 37384388. doi: 10.2196/48002.
7. Levin G, Horesh N, Brezinov Y, Meyer R. Performance of ChatGPT in medical examinations: A systematic review and a meta-analysis. BJOG. 2024 Feb;131(3):378-80. PMID: 37604703. doi: 10.1111/1471-0528.17641.
8. Wei Q, Yao Z, Cui Y, Wei B, Jin Z, Xu X. Evaluation of ChatGPT-generated medical responses: A systematic review and meta-analysis. J Biomed Inform. 2024 Mar;151:104620. PMID: 38462064. doi: 10.1016/j.jbi.2024.104620.
9. 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. PMID: 37215063. doi: 10.3389/frai.2023.1169595.

10. Wang X, Sanders HM, Liu Y, Seang K, Tran BX, Atanasov AG, et al. ChatGPT: promise and challenges for deployment in low- and middle-income countries. Lancet Reg Health West Pac. 2023 Dec;41:100905. PMID: 37731897. doi: 10.1016/j.lanwpc.2023.100905.
11. Adler-Milstein J, Redelmeier DA, Wachter RM. The Limits of Clinician Vigilance as an AI Safety Bulwark. JAMA. 2024 Mar 14. PMID: 38483397. doi: 10.1001/jama.2024.3620.
12. Hagendorff T, Fabi S, Kosinski M. Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. Nat Comput Sci. 2023 Oct;3(10):833-8. PMID: 38177754. doi: 10.1038/s43588-023-00527-x.
13. Bloom BS, Krathwohl DR. Taxonomy of Educational Objectives: The Classification of Educational Goals: Longmans, Green; 1956. ISBN: 9780582323865.
14. Krathwohl DR. A Revision of Bloom's Taxonomy: An Overview. Theory Into Practice. 2002 2002/11/01;41(4):212-8. doi: 10.1207/s15430421tip4104_2.
15. Chang BS. Transformation of Undergraduate Medical Education in 2023. JAMA. 2023 Oct 24;330(16):1521-2. PMID: 37698855. doi: 10.1001/jama.2023.16943.
16. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment. JMIR Med Educ. 2023 Feb 08;9:e45312. PMID: 36753318. doi: 10.2196/45312.
17. Su MC, Lin LE, Lin LH, Chen YC. Assessing question characteristic influences on ChatGPT's performance and response-explanation consistency: Insights from Taiwan's Nursing Licensing Exam. Int J Nurs Stud. 2024 Feb 08;153:104717. PMID: 38401366. doi: 10.1016/j.ijnurstu.2024.104717.
18. Acerbi A, Stubbersfield JM. Large language models show human-like content biases in transmission chain experiments. Proc Natl Acad Sci U S A. 2023 Oct 31;120(44):e2313790120. PMID: 37883432. doi: 10.1073/pnas.2313790120.
19. Morjaria L, Burns L, Bracken K, Ngo QN, Lee M, Levinson AJ, et al. Examining the Threat of ChatGPT to the Validity of Short Answer Assessments in an Undergraduate Medical Program. J Med Educ Curric Dev. 2023;10:23821205231204178. PMID: 37780034. doi: 10.1177/23821205231204178.
20. Kufel J, Paszkiewicz I, Bielówka M, Bartnikowska W, Janik M, Stencel M, et al. Will ChatGPT pass the Polish specialty exam in radiology and diagnostic imaging? Insights into strengths and limitations. Pol J Radiol. 2023;88:e430-e4. PMID: 37808173. doi: 10.5114/pjr.2023.131215.
21. 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. PMID: 36627845. doi: 10.3352/jeehp.2023.20.1.
22. Herrmann-Werner A, Festl-Wietek T, Holderried F, Herschbach L, Griewatz J, Masters K, et al. Assessing ChatGPT's Mastery of Bloom's Taxonomy Using Psychosomatic Medicine Exam Questions: Mixed-Methods Study. J Med Internet Res. 2024 Jan 23;26:e52113. PMID: 38261378. doi: 10.2196/52113.
23. Huang KJ. Evaluating GPT-4's Cognitive Functions Through the Bloom Taxonomy: Insights and Clarifications. J Med Internet Res. 2024 Apr 16;26:e56997. PMID: 38625725. doi: 10.2196/56997.
24. Tversky A, Kahneman D. Judgment under Uncertainty: Heuristics and Biases. Science. 1974 Sep 27;185(4157):1124-31. PMID: 17835457. doi: 10.1126/science.185.4157.1124.
25. Sloman SA. The empirical case for two systems of reasoning. Psychological Bulletin. 1996;119(1):3-22. doi: 10.1037/0033-2909.119.1.3.
26. Masters K, Benjamin J, Agrawal A, MacNeill H, Pillow MT, Mehta N. Twelve tips on creating and using custom GPTs to enhance health professions education. Med Teach. 2024 Jan 29:1-5. PMID: 38285894. doi: 10.1080/0142159X.2024.2305365.
27. Crowther GJ, Sankar U, Knight LS, Myers DL, Patton KT, Jenkins LD, et al. Chatbot responses suggest that hypothetical biology questions are harder than realistic ones. J Microbiol Biol Educ. 2023 Dec;24(3). PMID: 38107990. doi: 10.1128/jmbe.00153-23.
28. Luke WANV, Seow Chong L, Ban KH, Wong AH, Zhi Xiong C, Shuh Shing L, et al. Is ChatGPT 'ready' to be a learning tool for medical undergraduates and will it perform equally in different subjects? Comparative study of ChatGPT performance in tutorial and case-based learning questions in physiology and biochemistry. Med Teach. 2024 Jan 31:1-7. PMID: 38295769. doi: 10.1080/0142159X.2024.2308779.
29. Schubert MC, Wick W, Venkataramani V. Performance of Large Language Models on a Neurology Board-Style Examination. JAMA Netw Open. 2023 Dec 01;6(12):e2346721. PMID: 38060223. doi: 10.1001/jamanetworkopen.2023.46721.
30. Fergus S, Botha M, Ostovar M. EvaluatingAcademicAnswersGeneratedUsingChatGPT. Journal of Chemical Education. 2023 (100):4. doi: DOI: 10.1021/acs.jchemed.3c00087.
31. Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a Radiology Board-style Examination: Insights into Current Strengths and Limitations. Radiology. 2023 Jun;307(5):e230582. PMID: 37191485. doi: 10.1148/radiol.230582.

32. Das D, Kumar N, Longjam LA, Sinha R, Deb Roy A, Mondal H, et al. Assessing the Capability of ChatGPT in Answering First- and Second-Order Knowledge Questions on Microbiology as per Competency-Based Medical Education Curriculum. Cureus. 2023 Mar;15(3):e36034. PMID: 37056538. doi: 10.7759/cureus.36034.
33. Howell MD, Corrado GS, DeSalvo KB. Three Epochs of Artificial Intelligence in Health Care. JAMA. 2024 Jan 16;331(3):242-4. PMID: 38227029. doi: 10.1001/jama.2023.25057.

Abbreviations

AI artificial intelligence

TMLE Taiwan Medical Licensing Examination

USMLE The United States Medical Licensing Examination

Figure Legends

Figure 1 ChatGPTs' general performance based on Bloom's taxonomy.

Figure 2 ChatGPTs' performance analyzed using six levels in Bloom's taxonomy.

Multimedia Appendix

Multimedia Appendix 1. An Example of Prompting, Question Input, and Response in A Test Using Taiwan Medical Licensing Examination

Multimedia Appendix 2. Example of Question Type Classification by Bloom's Taxonomy (Question No. 46)

Supplementary Files

Multimedia Appendixes

An example of prompting, question input, and response in a test using Taiwan Medical Licensing Examination.
URL: http://asset.jmir.pub/assets/97efa57836ff7fa5621c8b00b17514f2.docx

An example of question type classification by Bloom's taxonomy (Question No. 46).
URL: http://asset.jmir.pub/assets/6e24f2b3f2a1cb6b06920cd6965bfb7e.docx