Original Research Article

DIGITAL HEALTH

DIGITAL HEALTH
Volume 10: 1–8

DOI: 10.1177/20552076241233144
journals.sagepub.com/home/dhj

Sage

Performance of ChatGPT on Stage 1 of the Taiwanese medical licensing exam

Chao-Hsiung Huang[1,*], Han-Jung Hsiao[2,*], Pei-Chun Yeh[2,*], Kuo-Chen Wu[2,3] and Chia-Hung Kao[2,4,5,6]

Abstract

Introduction: Since its release by OpenAI in November 2022, numerous studies have subjected ChatGPT to various tests to evaluate its performance in medical exams. The objective of this study is to evaluate ChatGPT's accuracy and logical reasoning across all 10 subjects featured in Stage 1 of Senior Professional and Technical Examinations for Medical Doctors (SPTEMD) in Taiwan, with questions that encompass both Chinese and English.

Methods: In this study, we tested ChatGPT-4 to complete SPTEMD Stage 1. The model was presented with multiple-choice questions extracted from three separate tests conducted in February 2022, July 2022, and February 2023. These questions encompass 10 subjects, namely biochemistry and molecular biology, anatomy, embryology and developmental biology, histology, physiology, microbiology and immunology, parasitology, pharmacology, pathology, and public health. Subsequently, we analyzed the model's accuracy for each subject.

Result: In all three tests, ChatGPT achieved scores surpassing the 60% passing threshold, resulting in an overall average score of 87.8%. Notably, its best performance was in biochemistry, where it garnered an average score of 93.8%. Conversely, the performance of the generative pre-trained transformer (GPT)-4 assistant on anatomy, parasitology, and embryology was not as good. In addition, its scores were highly variable in embryology and parasitology.

Conclusion: ChatGPT has the potential to facilitate not only exam preparation but also improve the accessibility of medical education and support continuous education for medical professionals. In conclusion, this study has demonstrated ChatGPT's potential competence across various subjects within the SPTEMD Stage 1 and suggests that it could be a helpful tool for learning and exam preparation for medical students and professionals.

Keywords

ChatGPT, Taiwanese medical licensing exam, artificial intelligence, educational measurement, OpenAI

Submission date: 15 August 2023; Acceptance date: 25 January 2024

[1]School of Medicine, China Medical University, Taichung
[2]Artificial Intelligence Center, China Medical University Hospital, China Medical University, Taichung
[3]Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taipei
[4]Graduate Institute of Biomedical Sciences, School of Medicine, College of Medicine, China Medical University, Taichung
[5]Department of Nuclear Medicine and PET Center, China Medical University Hospital, Taichung
[6]Department of Bioinformatics and Medical Engineering, Asia University, Taichung

[*]Chao-Hsiung Huang, Han-Jung Hsiao, and Pei-Chun Yeh contributed equally and shared the first author.

Corresponding author:
Chia-Hung Kao, Graduate Institute of Biomedical Sciences and School of Medicine, College of Medicine, China Medical University, No. 2, Yuh-Der Road, Taichung 404.
Emails: 010040@tool.caaumed.org.tw; dr.kaochiahung@gmail.com

Introduction

In recent years, the progress in technology has profoundly transformed the landscape of medical education and clinical practice. The advent of artificial intelligence, neural networks, and deep learning has bestowed upon us powerful tools for data interpretation and precise diagnosis. Among these useful tools are the large language models, exemplified by ChatGPT,[1] intricate neural networks boasting tens of billions of parameters. Ever since ChatGPT's public release in November 2022, its astounding ability to generate natural, coherent sentences with internal consistency has captured widespread attention.

As a result of this captivating attribute, the applications of ChatGPT have begun to emerge across various fields of expertise, and the realm of medicine is no exception. The impact of ChatGPT on the medical domain is remarkable, as it aids medical professionals in academic writing, patient interaction, and data analysis.[2] Additionally, this innovative technology has the potential to enhance medical education by assisting students to learn the crucial skills of communication, problem-solving, and critical thinking.[3] In terms of undergraduate education, ChatGPT could be a powerful tool for facilitating evidence-based decision-making and could help students prepare for multiple-choice examinations and objective structured clinical examinations (OSCEs).

To gain deeper insights into ChatGPT's competence in the domain of medical communication and education, a series of studies was undertaken, with the primary goal of investigating its capabilities. During these studies, ChatGPT underwent a diverse set of medical tests. These tests encompassed a wide range of assessments, starting from entrance exams like the Medical College Admission Test[4] and extending to expert-level tests across various medical specialties, including but not limited to radiology,[5] anesthesiology,[6] plastic surgery,[7] obstetrics, and gynecology.[8] In the expert-level tests, ChatGPT was able to yield results comparable to the results of the human examinees. While most of these evaluations were done with multiple-choice questions, the work of Li and colleagues[8] evaluated ChatGPT's performance through virtual OSCE in obstetrics and gynecology. The fact that ChatGPT outscored human candidates in their experiment further reflects the potential ChatGPT possessed in physician–patient communication.

There were also studies based on licensing exams. In the work of Kung and colleagues,[9] where the questions from the United States Medical Licensing Examination (USMLE) in 2022 were used, ChatGPT yielded passing performance with moderate accuracy. Its performance was generally better on questions requiring lower order thinking and was worse on questions involving calculation and application of concept. Apart from that, the model was able to generate insightful explanations with a confident tone, irrespective of their accuracy.

Given that the training data for ChatGPT was predominantly in English, several studies were carried out to validate the model's ability to maintain its accuracy when confronted with non-English medical examinations. Sahin and colleagues[10] found that generative pre-trained transformer (GPT)-4 achieved an accuracy of 78.77% on the Turkish Neurosurgical Society Proficiency Board Exams, surpassing the 62.02% average accuracy of human candidates. Oztermeli et al.[11] demonstrated GPT-4's success on the Medical Specialty Exam in Turkish, with an accuracy of 64.2% in basic science questions and 62.9% in clinical questions, both exceeding passing thresholds. As for Spanish-based exams, Carrasco et al.[12] reported that ChatGPT achieved an accuracy of 51.4% on the Médico Interno Residente Exam, slightly below the median of human candidates but adequate for passing. Flores-Cohalia et al.[13] found GPT-4's accuracy to be 86% on the Peruvian National Licensing Medical Exam, which was also written in Spanish, outperforming the 55% average accuracy of human candidates. The study conducted by Bonetti et al.[14] demonstrated GPT-3's capability to pass the Italian-written Residency Admission National Examination with a score of 122/140, reaching the 98.8th percentile of human examinees. Overall, ChatGPT was capable of achieving passing performances in exams written in various non-English languages and could often outperform human examinees.

On the other hand, ChatGPT's performance on the exams written in Chinese was limited. Weng et al.[15] observed GPT-4's challenges with the Taiwanese Family Medicine Board Exam, achieving an accuracy of 41.6%, below the passing threshold of 60%, though its performance in certain question types was acceptable. Wang et al.[16] reported accuracy for GPT-4 on the Chinese National Medical Licensing Examination (NMLE), with 47.0% in NMLE 2020, 45.8% in NMLE 2021, and 36.5% in NMLE 2022, all falling short of the passing threshold and were inferior to the performance of Chinese medical students. Kao et al.[17] evaluated GPT-3 on the internal medicine section of the Taiwanese Staged Senior Professional and Technical Examinations for Medical Doctors (SPTEMD), reporting a variable accuracy range of 30.4% to 53.8%, which was below the 60% passing threshold. Huang et al.[18] explored GPT-3.5's performance on the Taiwanese Registered Nurse License Exam, achieving an accuracy between 51.6% and 63.75%, with concerns raised about potentially misleading or inaccurate explanations. Another study regarding the Taiwanese Pharmacist Licensing Examination[19] yielded a non-passing result. It's important to highlight that within this study, ChatGPT demonstrated improved performance when confronted with the translated version of the exam in English. Based on these studies, it is reasonable to conclude that ChatGPT does exhibit a disadvantage when answering questions written in Chinese.

Table 1. Subject distribution of the question set.

Subject	Feb 2022	Jul 2022	Feb 2023
Anatomy	31	31	31
Embryology[a]	5	5	5
Histology	10	11	11
Physiology	27	26	26
Biochemistry[b]	27	27	27
Microbiology[c]	28	28	28
Parasitology	7	7	7
Public health	17	15	15
Pharmacology	23	25	25
Pathology	25	25	25
Total	200	200	200

[a]Embryology and Development Biology; [b]Biochemistry and Molecular Biology; [c]Microbiology and Immunology

In this study, we seek to delve deeper into ChatGPT's capabilities in a multilingual environment, specifically by subjecting it to Stage 1 of SPTEMD in Taiwan. This exam encompasses a combination of Chinese and English languages, with Chinese being the predominant medium of communication. Regarding the subjects covered in the exam, it comprises 10 distinct areas as defined by the Ministry of Examination. These subjects include biochemistry and molecular biology, anatomy, embryology and developmental biology, histology, physiology, microbiology and immunology, parasitology, pharmacology, pathology, and public health. By subjecting ChatGPT to this complex examination, researchers aim to evaluate its performance and adaptability in comprehending and responding accurately to medical content presented in a multilingual context. We believe the findings from this investigation will shed light on ChatGPT's potential role in medical education in diverse linguistic settings. If ChatGPT could successfully pass the Stage 1 exam, it might serve as a useful self-study tool for medical students studying these subjects.

Method

Generative pre-trained transformer

Developed and released by OpenAI, ChatGPT is a large language model. In terms of design, ChatGPT is structured as a massive neural network with tens of billions of parameters, enabling it to perform complex language-related tasks effectively. The model's architecture comprises multiple layers of transformers, facilitating long-range dependencies and context retention, which is pivotal in generating coherent and contextually consistent responses.

We used GPT-4[20] in this study. The date of its training data was up to September 2021. No additional pre-training was done in the experiment. We didn't allow the model to search any internal or external database in this experiment.

Staged senior professional and technical examinations for medical doctors

The Staged SPTEMD[21] is an exam that Taiwanese medical students are required to pass to obtain the licenses of medical doctors. It was set by an examination committee appointed by the Ministration of Examination, the government branch overseeing the licensing exams in Taiwan. The content of the SPTEMD follows the syllabus and textbook lists declared by the Ministration of Examination.

The SPTEMD consists of two stages, and both are held twice yearly. Stage 1 assesses the examinee's understanding of important concepts of science basic to the practice of medicine. It includes 10 subjects: biochemistry and molecular biology, anatomy, embryology and developmental biology, histology, physiology, microbiology and immunology, parasitology, pharmacology, pathology, and public health. On the other hand, Stage 2 assesses the examinee's capabilities of handling practical clinical scenarios. Since the main focus of this study is the Stage 1 exam, the details of the Stage 2 exam will not be discussed here.

The Stage 1 exam consists of 200 multiple-choice questions. Each question contains a description and four options, of which only one is correct. It is worth noticing that the questions were not evenly distributed among the subjects. Some subject, such as anatomy, pharmacology, and pathology, typically takes around 25 to 30 questions in each test; on the other hand, subjects such as embryology and histology usually take only 5 to 10 questions.

While the questions of the Stage 1 Exam were mainly written in Chinese, English phrases were often used for clarification since many medical terms, such as anatomic structures or medications, lack unified Chinese translation. As a result, English characters constituted roughly 41% to 48% of these questions.

In this study, questions extracted from three tests, respectively, held in February 2022, July 2022, and February 2023 were used in the experiment. These test questions are publicly available on the official website of the Ministration of Examination for open access by the general public. Since each test contains 200 questions, we obtained a total of 600 questions for the experiment. We didn't implement any modification to the questions in this extraction process. Since the training data of GPT-4 is

only up to September 2021, it is reasonable to assume that the version of ChatGPT used in this study had not been trained with these questions. We included image-based questions since GPT-4 is capable of processing images. We described the distribution of subjects in the resulting question set in Table 1.

Experiment design

We created a GPT-4 assistant using the configuration shown in Figure 1. We provided the following instructions to the assistant: "Your main responsibility is to help individuals who have graduated from medical schools in Taiwan and are preparing for the first stage of written examinations required to become a practicing doctor. Your assistance will focus on providing study materials, explanations, practice questions, and relevant information to help these graduates understand and recall medical knowledge effectively. It is important to provide accurate and reliable medical information and comply with the latest guidelines and standards in Taiwan's medical field. If the query is unclear or beyond your expertise, seek clarification or advise the user accordingly. Your answers should be informative, clear, and directly consistent with the topics and format of the Taiwanese Phase 1 Physical Examination."

The researchers entered the 600 questions manually into the GPT-4 assistant and collected their replies. For image-based questions, we would first upload the image and then enter the question. We showed an example of an image-based question in Figure 2. We entered 20 questions in one conversation. For every conversation, we created a new GPT-4 assistant to reduce memory bias. When entering the questions, we provided no additional information or instruction in the prompt aside from the question itself. The "zero-shot" method was used and the GPT-4 assistant was allowed only one attempt per question.

The responses generated by the model were subsequently compared with the standard answers, and the corresponding scores were computed. The scores were then analyzed using Microsoft Excel. Notably, while adhering to the principle of permitting only one correct answer per question, certain ambiguities identified by the examination committee allowed for multiple options to be deemed correct in specific instances. In such scenarios, the model's response was deemed accurate if it encompassed any of the sanctioned choices. For example, the 66th question in the test held in February 2022 was as follows (the original text was in Chinese):

"Which of the following is the primary factor determining the secretion and excretion of potassium ions in the distal renal tubule?

(A) Presence or absence of antidiuretic hormone
(B) Quantity of positively charged ions within the tubular lumen
(C) Urine flow rate in the proximal renal tubule
(D) Excretion through specific potassium ion channels."

Originally, the standard answer for this question was option D. However, the examination committee later decided that option B was also an acceptable answer after discussion. Thus, both options B and D were considered correct in this case.

Result

Table 2 illustrates the number of correct questions for each test. Overall, the model achieved a score of 176 (88.00%) in the February 2022 test, 171 (85.60%) in the July 2022 test, and 178 (89.00%) in the February 2023 test.

In more detail, ChatGPT achieved accuracy above 90% in biochemistry, histology, pharmacology, and microbiology. The average scores of ChatGPT are shown in Table 3. Notably, its best performance was in biochemistry, where it garnered an average score of 93.8%. On the other hand, ChatGPT's average score on anatomy (81.7%), parasitology (81.0%), and embryology (80.0%) were not as good. It is worth noting that the model exhibited less consistent performance in subjects with fewer than 10 questions per test, specifically embryology and parasitology. In embryology, the average score was 80.0%, with respective test scores of 100.0%, 60.0%, and 80.0%. Similarly, in parasitology, the model's average score was 80.95%, accompanied by test scores of 85.71%, 57.14%, and 100.0%.

In summary, ChatGPT answered 525 questions correctly out of the 600 questions from Stage 1 of Taiwanese Staged Senior Professional Examinations for Medical Doctors used in this experiment, achieving an average accuracy of 87.5%. The model produced the highest accuracy of 93.8% in biochemistry. Conversely, the performance of the GPT-4 assistant on anatomy, parasitology, and embryology was not as good. In addition, its scores were highly variable in embryology and parasitology.

Discussion

In the evaluation across three tests (February 2022, July 2022, and February 2023), ChatGPT consistently achieved scores of 88.00%, 85.60%, and 89.00%, respectively. Demonstrating proficiency above 90% in biochemistry, histology, pharmacology, and microbiology, the model's average score was 93.8% in biochemistry. However, it exhibited comparatively lower performance in anatomy (81.7%), parasitology (81.0%), and embryology (80.0%). The model's consistency in subjects with fewer than 10 questions per test, specifically embryology and parasitology, varied. In embryology, the average score was 80.0%, with test scores of 100.0%, 60.0%, and 80.0%. Similarly, in parasitology, the model's average score was 80.95%, accompanied by test scores of 85.71%, 57.14%,

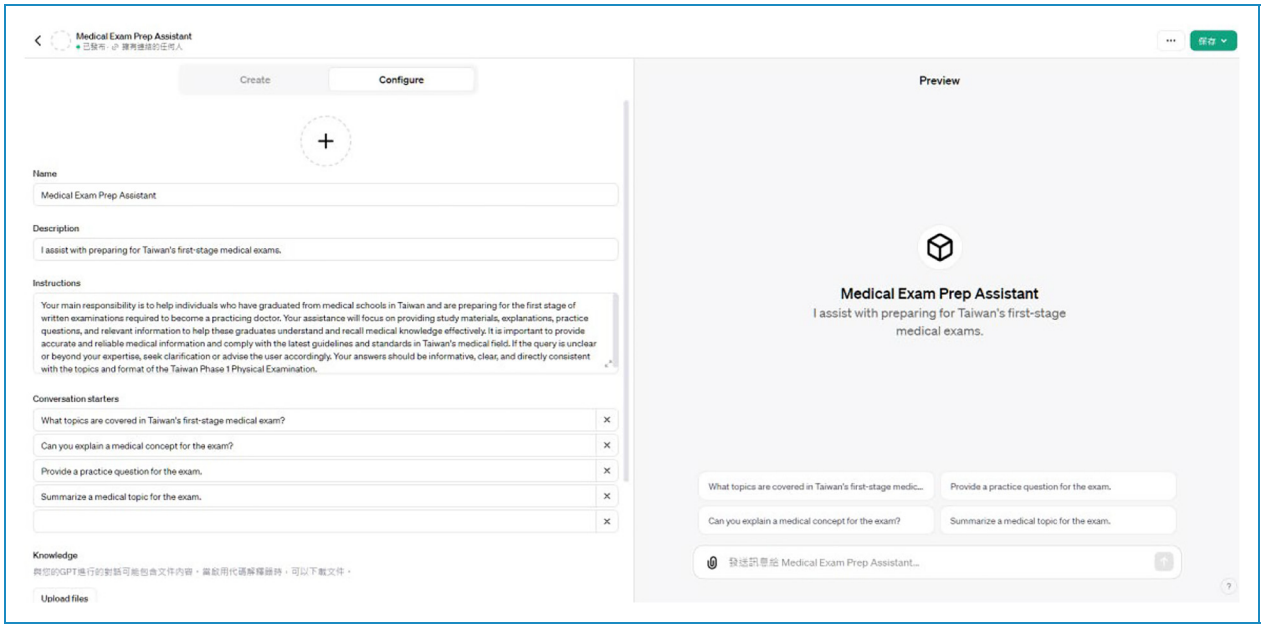

Figure 1. The configuration of the GPT-4 assistant. GPT: generative pre-trained transformer.

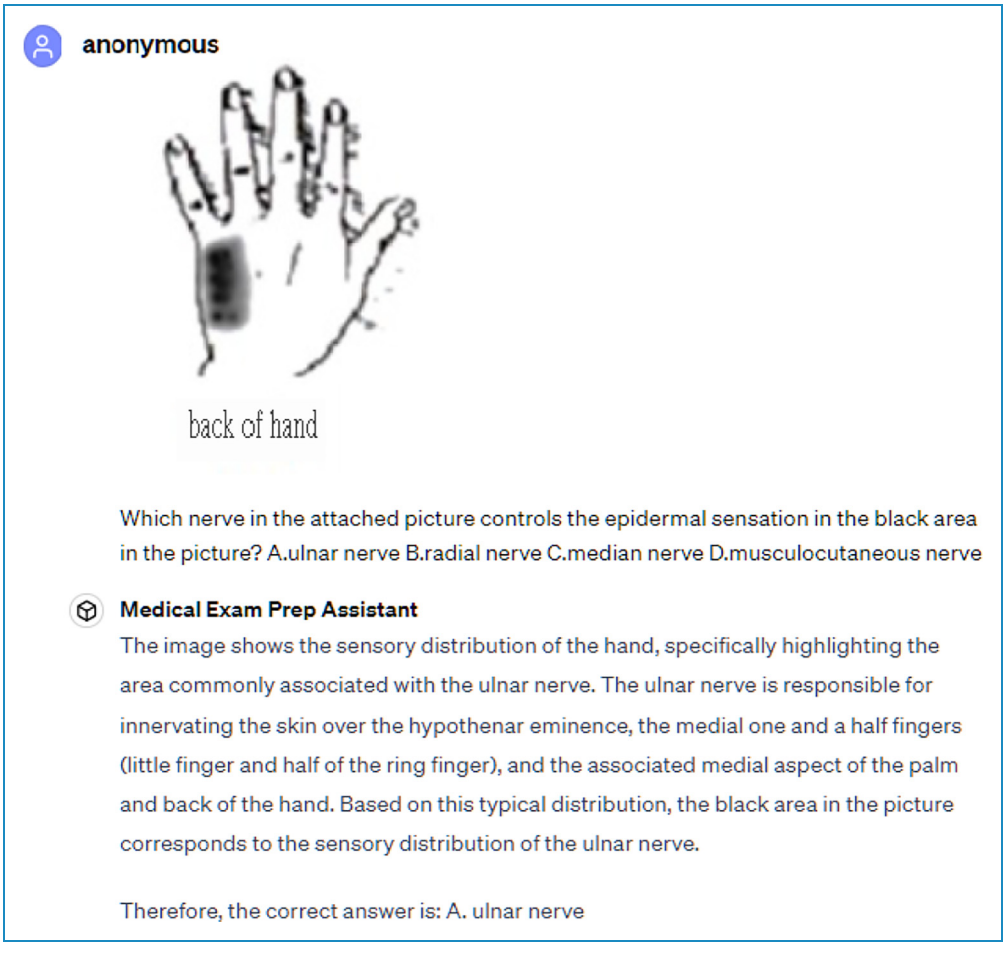

Figure 2. An example of an image-based question.

Table 2. Test results.

	Feb 2022			Jul 2022			Feb 2023		
Subject	Total	Correct	%	Total	Correct	%	Total	Correct	%
Anatomy	27	27	100.0	31	26	83.9	31	25	80.7
Embryology[a]	10	9	90.0	5	3	60.0	5	4	80.0
Histology	23	20	86.9	11	10	90.9	11	11	100.0
Physiology	28	24	85.7	26	22	84.6	26	22	84.6
Biochemistry[b]	17	16	94.1	27	23	85.2	27	26	96.3
Microbiology[c]	25	24	96.0	28	26	92.9	28	26	92.9
Parasitology	27	22	81.5	7	4	57.1	7	7	100.0
Public health	7	6	85.7	15	13	86.7	15	12	80.0
Pharmacology	31	25	80.7	25	24	96.0	25	24	96.0
Pathology	5	5	100.0	25	20	80.0	25	21	84.0
Total	200	178	89.0	200	171	85.5	200	178	89.0

[a]Embryology and Development Biology; [b]Biochemistry and Molecular Biology; [c]Microbiology and Immunology

Table 3. Average scores of each subject.

Subject	Total	Correct	%
Biochemistry[b]	81	76	93.8
Microbiology[c]	84	76	90.5
Anatomy	93	76	81.7
Pharmacology	73	68	93.2
Physiology	79	66	83.5
Pathology	75	65	86.7
Public health	47	41	87.2
Histology	32	30	93.8
Parasitology	21	17	81.0
Embryology[a]	15	12	80.0
Total	600	527	87.8

[a]Embryology and Development Biology; [b]Biochemistry and Molecular Biology; [c]Microbiology and Immunology

and 100.0%. Overall, ChatGPT correctly answered 525 out of 600 questions from Stage 1 of SPTEMD in Taiwan, resulting in an average accuracy of 87.5%. The model exhibited strengths in certain subjects while showing areas for improvement, particularly in anatomy, parasitology, and embryology.

The demonstrated capability of ChatGPT to pass the Taiwanese Licensing Exam for Medical Doctors suggests its potential role in medical education across three aspects. Firstly, as a supplementary tool, ChatGPT could augment exam preparation for medical students. Institutions may consider integrating ChatGPT or similar AI models into educational platforms, providing learners with interactive and personalized experiences, thereby optimizing the learning process and enabling focused improvement. One such example is Chatprogress,[22] a chatbot-based game developed at Paris Descartes University. Al Kahf and colleagues demonstrated a significant improvement in students' results when using Chatprogress to facilitate learning.

Secondly, the advent of high-performing AI models, demonstrated by ChatGPT's capability to pass a Chinese-written test, could further enhance the accessibility and inclusivity of medical education resources by lowering language barriers. Students in various geographic locations,

with disparate access to traditional educational resources, stand to benefit from the availability of AI-driven tools, promoting a more equitable distribution of educational opportunities.

Lastly, ChatGPT and similar AI models hold promise for supporting continuous learning among medical professionals beyond formal education. Serving as interactive resources, they offer practitioners a means to stay informed about medical advancements, updated guidelines, and best practices throughout their careers. In summary, ChatGPT's potential role in medical education spans exam preparation enhancement, increased accessibility, and support for continuous learning in the evolving landscape of medical knowledge.

While ChatGPT demonstrated satisfactory accuracy in passing Stage 1 of the SPTEMD in Taiwan within the scope of this study, its performance proved suboptimal in other Chinese-written examinations. A comparison with the findings of Wang and colleagues[16] revealed a noteworthy contrast. Despite both studies employing GPT-4 in their experiments and examining licensing exams for medical doctors, our results differed from theirs. As outlined in the introduction section, GPT-4 fell short in the Chinese NMLE in their study, exhibiting inferior performance compared to medical students. This discrepancy can be attributed to various factors, including variations in the subjects included, question format, and the medical education system. In essence, while ChatGPT achieved a passing performance in our study, it is essential to exercise caution in generalizing this conclusion to all Chinese-written examinations.

This study does possess certain limitations. Firstly, the lack of published statistics by the Ministry of Examination prevented us from establishing the evidence validity of SPTEMD and from directly comparing the model's performance with that of human examinees. Secondly, the restriction of only one attempt per question can potentially affect ChatGPT's performance. Thirdly, the uneven distribution of questions across subjects hindered a comprehensive analysis of the model's performance, particularly in subjects with lower question occurrences, such as embryology and parasitology. The statistical significance of results in these less frequent subjects was consequently difficult to ascertain. Lastly, the lack of comparison with other chatbots would limit the findings of this study to ChatGPT, and further investigations are called for evaluating the performance of other large language models on this exam.

Conclusion

ChatGPT has the potential to facilitate not only the preparation for exams but also improve the accessibility of medical education and support continuous education for medical professionals. As such, it could play a multifaceted role in enhancing medical education and practice. In conclusion, the study suggests that ChatGPT-4 could be harnessed to support medical students and professionals, contributing to their understanding of basic science and medical knowledge.

Acknowledgments: This study is supported in part by China Medical University Hospital (DMR-113-048, DMR-113-060, DMR-113-061). The funders had no role in the study design, data collection, analysis, the decision to publish, or the preparation of the manuscript. No additional external funding was received for this study.

Author contributions: The authors' individual contributions were as follows: Chao-Hsiung Huang and Chia-Hung Kao were responsible for the study design. Chao-Hsiung Huang, Han-Jung Hsiao, Pei-Chun Yeh, and Kuo-Chen Wu collected the data. All authors performed the statistical analyses, data interpretation, and article drafting. All authors provided some intellectual content. All authors approved this version of the manuscript for submission. All authors read and approved the final manuscript.

Declaration of conflicting interests: The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study is supported in part by China Medical University Hospital (DMR-112-072, DMR-112-073).

ORCID iD: Chia-Hung Kao https://orcid.org/0000-0002-6368-3676

Supplemental material: Supplemental material for this article is available online.

References

1. OpenAI. ChatGPT: optimizing language models for dialogue. [cited 2023 March 3], https://openai.com/blog/chatgpt/ (accessed 3 March 2023).
2. Sedaghat S. Early applications of ChatGPT in medical practice, education and research. *Clin Med (Lond)* 2023; 23: 278–279.
3. Seetharaman R. Revolutionizing medical education: can ChatGPT boost subjective learning and expression? *J Med Syst* 2023; 47: 61.
4. Bommineni VL, et al. Performance of ChatGPT on the MCAT: the road to personalized and equitable premedical learning. *MedRxiv* 2023. doi: 10.1101/2023.03.05.23286533.
5. Bhayana R, Krishna S and Bleakney RR. Performance of ChatGPT on a radiology board-style examination: insights into current strengths and limitations. *Radiology* 2023; 307: e230582.
6. Birkett L, Fowler T and Pullen S. Performance of ChatGPT on a primary FRCA multiple choice question bank. *Br J Anaesth* 2023; 131: e34–e35.

7. Humar P, et al. ChatGPT is equivalent to first-year plastic surgery residents: evaluation of ChatGPT on the plastic surgery in-service examination. *Aesthet Surg J* 2023; 43: NP1085–NP1089.
8. Li SW, et al. ChatGPT outscored human candidates in a virtual objective structured clinical examination in obstetrics and gynecology. *Am J Obstet Gynecol* 2023; 229: 172.e1–172.e12.
9. Kung TH, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health* 2023; 2: e0000198.
10. Sahin MC, et al. Beyond human in neurosurgical exams: ChatGPT's success in the Turkish neurosurgical society proficiency board exams. *Comput Biol Med* 2023; 169: 107807.
11. Oztermeli AD and Oztermeli A. ChatGPT performance in the medical specialty exam: an observational study. *Medicine (Baltimore)* 2023; 102: e34673.
12. Carrasco JP, et al. ¿Es capaz "ChatGPT" de aprobar el examen MIR de 2022? Implicaciones de la inteligencia artificial en la educación médica en España. *Revista Española de Educación Médica* 2023; 4: 12–18.
13. Flores-Cohaila JA, et al. Performance of ChatGPT on the Peruvian National Licensing Medical Examination: cross-sectional study. *JMIR Med Educ* 2023; 9: e48039.
14. Bonetti MA, et al. How does ChatGPT perform on the Italian residency admission national exam compared to 15,869 medical graduates? *Ann Biomed Eng* 2023. doi: 10.1007/s10439-023-03318-7. Online ahead of print.
15. Weng TL, et al. ChatGPT failed Taiwan's family medicine board exam. *J Chin Med Assoc* 2023; 86: 762–766.
16. Wang X, et al. Chatgpt performs on the Chinese national medical licensing examination. *J Med Syst* 2023; 47: 86.
17. Kao YS, Chuang WK and Yang J. Use of ChatGPT on Taiwan's examination for medical doctors. *Ann Biomed Eng* 2023. doi: 10.1007/s10439-023-03308-9. Online ahead of print.
18. Huang HM. Performance of ChatGPT on registered nurse license exam in Taiwan: a descriptive study. *Healthcare (Basel)* 2023; 11: 2855.
19. Wang YM, Shen HW and Chen TJ. Performance of ChatGPT on the pharmacist licensing examination in Taiwan. *J Chin Med Assoc* 2023; 86: 653–658.
20. OpenAI. GPT-4 Technical Report. 2023; arXiv (Cornell University). doi: 10.48550/arxiv.2303.08774.
21. Kao MC. Overview of the history of the examination system for medical doctors in Taiwan. *Taiwan Med J* 2012; 55: 38–49.
22. Al Kahf S, et al. Chatbot-based serious games: a useful tool for training medical students? A randomized controlled trial. *PLoS One* 2023; 18: e0278673.