J Chin Med Assoc

ORIGINAL ARTICLE

While GPT-3.5 is unable to pass the Physician Licensing Exam in Taiwan, GPT-4 successfully meets the criteria

Tsung-An Chen[a], Kuan-Chen Lin[a], Ming-Hwai Lin[a,b], Hsiao-Ting Chang[a,b], Yu-Chun Chen[a,b,c,d,e,*], Tzeng-Ji Chen[f,g,*]

[a]Department of Family Medicine, Taipei Veterans General Hospital, Taipei, Taiwan, ROC; [b]School of Medicine, National Yang Ming Chiao Tung University, Taipei, Taiwan, ROC; [c]Big Data Center, Taipei Veterans General Hospital, Taipei, Taiwan, ROC; [d]Institute of Hospital and Health Care Administration, National Yang Ming Chiao Tung University, Taipei, Taiwan, ROC; [e]Department of Family Medicine, Taipei Veterans General Hospital Yuli Branch, Hualien, Taiwan, ROC; [f]Department of Family Medicine, Taipei Veterans General Hospital Hsinchu Branch, Hsinchu, Taiwan, ROC; [g]Department of Post-Baccalaureate Medicine, National Chung Hsing University, Taichung, Taiwan, ROC

Abstract

Background: This study investigates the performance of ChatGPT-3.5 and ChatGPT-4 in answering medical questions from Taiwan's Physician Licensing Exam, ranging from basic medical knowledge to specialized clinical topics. It aims to understand these artificial intelligence (AI) models' capabilities in a non-English context, specifically traditional Chinese.

Methods: The study incorporated questions from the Taiwan Physician Licensing Exam in 2022, excluding image-based queries. Each question was manually input into ChatGPT, and responses were compared with official answers from Taiwan's Ministry of Examination. Differences across specialties and question types were assessed using the Kruskal–Wallis and Fisher's exact tests.

Results: ChatGPT-3.5 achieved an average accuracy of 67.7% in basic medical sciences and 53.2% in clinical medicine. Meanwhile, ChatGPT-4 significantly outperformed ChatGPT-3.5, with average accuracies of 91.9% and 90.7%, respectively. ChatGPT-3.5 scored above 60.0% in seven out of 10 basic medical science subjects and three of 14 clinical subjects, while ChatGPT-4 scored above 60.0% in every subject. The type of question did not significantly affect accuracy rates.

Conclusion: ChatGPT-3.5 showed proficiency in basic medical sciences but was less reliable in clinical medicine, whereas ChatGPT-4 demonstrated strong capabilities in both areas. However, their proficiency varied across different specialties. The type of question had minimal impact on performance. This study highlights the potential of AI models in medical education and non-English languages examination and the need for cautious and informed implementation in educational settings due to variability across specialties.

Keywords: Accuracy; Artificial intelligence; ChatGPT; Medical education; Taiwan's Physician Licensing Exam

1. INTRODUCTION

In the rapidly evolving field of artificial intelligence (AI), emergence of sophisticated language models such as Generative Pre-trained Transformer (ChatGPT-3.5 and ChatGPT-4), developed by OpenAI, has sparked significant interest and debate, particularly in their application to specialized fields such as medicine.[1,2] In November 2022, OpenAI introduced ChatGPT-3.5, a significant language model trained using 175 billion parameters, capable of producing diverse content including dialogs, language translations, and article writing. Subsequently, on March 14, 2023, OpenAI introduced a more advanced version, ChatGPT-4, as a paid service. With an expanded parameter set and enhanced training, this version is expected to exhibit superior reasoning and comprehension abilities compared with its predecessor.[3]

Since its launch, ChatGPT has sparked extensive discussions within the medical community.[4] Its application in medical education has captured much attention.[5] Benefits of using it in medical education include personalized learning, clinical reasoning, and instant feedback.[6] However, some concerns persist regarding its application in medical education, including bias and inaccuracy.[7] Thus, the medical knowledge of this AI model and its accuracy are crucial to ensure reliability.[7,8] Several studies have assessed the accuracy of ChatGPT in answering medical questions across various conditions.[9] However, research has found that its performance was influenced by factors such as specialty,[10–14] language,[15–17] and generation of a GPT model.[16,18,19] While studies have predominantly focused on English-based assessments and subspecialities,[20–22] further research on its performance in

*Address correspondence. Dr. Tzeng-Ji Chen, Department of Family Medicine, Taipei Veterans General Hospital Hsinchu Branch, 81, Section 1, Zhongfeng Road, Zhudong Township, Hsinchu 310, Taiwan, ROC. E-mail address: tjchen@vhct.gov.tw (T.-J. Chen), Dr. Yu-Chun Chen, Department of Family Medicine, Taipei Veterans General Hospital Yuli Branch, 91, Xinxing Street, Yuli Township, Hualien 981, Taiwan, ROC. E-mail address: yuchn.chen@gmail.com (Y.-C. Chen).

Author contributions: Dr. Yu-Chun Chen and Dr. Tzeng-Ji Chen contributed equally to this work.

Conflicts of interest: Dr. Tzeng-Ji Chen and Dr. Yu-Chun Chen, editorial board members at the Journal of the Chinese Medical Association, have no roles in the peer review process of or decision to publish this article. The other authors declare that they have no conflicts of interest related to the subject matter or materials discussed in this article.

Journal of Chinese Medical Association. (2025) 88: 352-360.

Received September 3, 2023; accepted April 13, 2024.

doi: 10.1097/JCMA.0000000000001225

interpreting and responding to exam questions formulated in traditional Chinese should generate valuable insights into the functionality of language models in non-English languages.

This study aims to explore the performance of ChatGPT-3.5 and ChatGPT-4 across a spectrum of medical specialties and different types of questions as presented in the Taiwan Physician Licensing Exam. Our primary goal is to deepen our understanding of how these models perform when tasked with responding to medical examination questions in traditional Chinese, a departure from predominantly English-focused assessments in existing studies. By evaluating its capabilities in a comprehensive range of medical disciplines and diverse question formats within the unique traditional Chinese linguistic context, the study seeks to substantially contribute to the ongoing exploration of language models' applicability in medical education and examination settings in traditional Chinese.

2. METHODS

2.1. Background of Taiwan's National Physician Licensing Exam

In Taiwan, medical students must undergo two stages of national examinations before earning their physician licenses. The first stage focuses on basic medical science, which occurs between their fourth and fifth years in medical school and is divided into two sections: (1) Medical Science I, comprising biochemistry, anatomy, embryology, histology, and physiology and (2) II, encompassing microbiology and immunology, parasitology, pharmacology, pathology, and public health. Each section consists of 100 single-select multiple-choice questions with four options each. The perfect score is 100 points, with a passing score of 60.0% or above. The second focuses on clinical medicine, which takes place after graduation and consists of multiple sections: (1) Medical Science III, covering internal medicine and family medicine; (2) IV, which includes pediatrics, dermatology, neurology, and psychiatry; (3) V, focusing on surgery, orthopedics, and urology; and (4) VI, comprising anesthesiology, ophthalmology, otolaryngology, obstetrics and gynecology, and rehabilitation. Each section has 80 single-select multiple-choice questions with four options each. Again, the perfect score is 100 points, with a passing threshold of 60.0% or higher.

2.2. Data source

This study aims to rigorously evaluate the performance of ChatGPT-3.5 and ChatGPT-4, two advanced language models developed by OpenAI, using questions from the Taiwan Physician Licensing Examination conducted in 2022. The choice of questions was deliberate; the data used for training ChatGPT extend to September 2021; moreover, at the time of the experiment (August 10–20, 2023), ChatGPT-4 had no Internet access. This approach was adopted to avoid any potential pretraining bias and provide a fair assessment of the models' capabilities.

The examination questions and answers were obtained from the official website of the Ministry of Examination.[23] The questions were written in Chinese, but the drug names and professional terminologies were presented in English. Additionally, image-based questions were excluded from the analysis.

2.3. Examination structure and question selection

The Taiwan Physician Licensing Examination is a comprehensive test encompassing a wide array of subjects in basic medical science and clinical medicine. Exam questions are predominantly multiple-choice. For the purpose of this study, we selected a diverse set of questions from various medical specialties to thoroughly evaluate the models' proficiency. However, image-based and diagrammatic questions were not included as they are outside the scope of ChatGPT's text-based processing capabilities.

2.4. Query formation and model interaction

Each selected exam question was reformulated into a Chinese query format that caters to ChatGPT, directing it to choose the most accurate answer along with justifications. To minimize possible biases from inconsistencies in multiple queries, each question was presented to it just once. Additionally, every query was initiated in a new session, effectively preventing the retention of information from prior sessions. Each query began with an additional instruction to prompt the model to start over (Fig. 1, a translated version in the Supplementary File, http://links.lww.com/JCMA/A319). Efforts were made to ensure that each question's essence was preserved while making them comprehensible and answerable by the models. The questions were manually input into ChatGPT-3.5 and ChatGPT-4, with their respective responses being meticulously recorded. This approach was carefully employed to ensure each question was evaluated objectively, maintaining integrity and reliability of the assessment. ChatGPT was prompted to provide reasoning for its answers, and the responses, along with their rationales generated by both models, were reviewed by two physicians, Dr. Chen TA and Dr. Lin KC. This was done to ensure an in-depth evaluation of the logic underlying its answers.

2.5. Assessing the accuracy of ChatGPT responses

Questions with images were excluded. After each query was entered into ChatGPT, it was prompted to answer the question. The answers were then compared with the official answers released by the Ministry of Examination.[24] For multiple correct answers, a response from ChatGPT was considered accurate if it corresponded to any of them. The accuracy rate was calculated as the number of correctly answered questions divided by the total number of questions after excluding those with images. The entire process took place from August 10–20, 2023.

2.6. Categorization and analysis of question types

Questions were sequentially categorized into four types: clinical vignette questions, single-choice multiple-item questions, positive-choice questions, and negative-choice questions (Fig. 2, a translated version in the Supplementary File, http://links.lww.com/JCMA/A319). A question was classified as a clinical vignette if it covered a clinical scenario. For the rest of the questions, those presenting a list of items and requiring respondents to select a choice from the list were termed single-choice multiple-item questions. Finally, we ascertained the question's polarity based on its phrasing: those ending with "Which of the following options is correct?" or similar expressions were classified as positive-choice questions whereas those ending with "Which of the following options is incorrect?" or equivalent statements were deemed negative-choice questions. The entire process was conducted from August 10–20, 2023.

2.7. Statistical analysis

Responses from ChatGPT-3.5 and ChatGPT-4 were compared against the official answers. Accuracy rates were calculated for each model across different specialties and question types. The Kruskal–Wallis test and Fisher's exact test were conducted to compare mean scores and accuracy rates between models across various domains and question types. A p value <0.05 was considered statistically significant. Data visualization and statistical analysis were conducted using Microsoft Excel 2019 and IBM SPSS Statistics 25.

Fig. 1 A structured interaction with ChatGPT in Chinese. The upper segment displays the query, a combination of a question from Taiwan's National Physician Licensing Exam, and additional instructions. The lower segment reveals ChatGPT's response comprising an answer and its rationale.

3. RESULTS

This study incorporated questions from the physician licensing exam held in February 2022. After excluding questions containing images, 99 questions were examined for Medical Sciences I, 99 for II, 76 for III, 69 for IV, 74 for V, and 72 for VI (Table 1). Table 2 displays the distribution of questions across various specialties.

3.1. Overall performance of ChatGPT models

ChatGPT-4 (91.1%) showed a statistically significant improvement over ChatGPT-3.5 (58.0%, $p < 0.001$). Moreover, ChatGPT-3.5 was more adept at basic medical sciences (67.7%) compared with clinical medicine (53.3%; $p = 0.002$). Conversely, ChatGPT-4 was highly efficient across both domains, with an accuracy of 91.9% in basic medicine and 90.7% in clinical medicine, indicating no significant difference in proficiency ($p = 0.647$). While ChatGPT-3.5 struggled to meet the required accuracy benchmark for the second stage of the exam, ChatGPT-4's accuracy rates were consistently higher and surpassed the benchmarks across all subjects in the physician licensing exam of 2022. ChatGPT-3.5's accuracy rates for Medical Sciences I and II were 65.7% (65/99) and 69.7% (69/99), respectively, with an average of 67.7%, which is sufficient to pass the first stage of the exam. However, its accuracy rates for Medical Science III, IV, V, and VI were 64.5% (49/76), 53.6% (37/69), 43.2% (32/74), and 51.4% (37/72), respectively, resulting in an average of 53.2%, which does not meet the threshold for the second stage of the exam. In contrast, ChatGPT-4's accuracy rates were 89.9% (89/99) and 93.9% (93/99) in Medical Sciences I and II, respectively, with an average of 91.9%. For Medical Science III, IV, V, and VI, its respective accuracy rates were 94.7% (72/76), 91.3% (63/69), 83.8% (62/74), and 93.1% (67/72), with an average of 90.7%, enough to pass the second stage of the exam (Table 1).

3.2. Accuracy of ChatGPT across medical specialties

During the same test, both ChatGPT-3.5 and ChatGPT-4 showed variances in accuracy across medical specialties, with ChatGPT-4 consistently achieving higher accuracy rates. The top three highest-scoring specialties for ChatGPT-3.5 were embryology (100.0%), biochemistry (81.5%), and pathology (79.2%), while the three lowest-scoring subjects were urology (12.5%), histology (30.0%), and ophthalmology and orthopedics (both 33.3%). Meanwhile, ChatGPT-4 achieved a 100.0% accuracy rate in five subjects: otolaryngology, public health, family medicine, dermatology, and anesthesiology. Its lowest-scoring subjects were urology (75.0%), embryology (80.0%), and surgery (84.2%). For ChatGPT-3.5, 7 of 10 specialties of basic medical sciences had achieved accuracy rates exceeding 60.0%; in the 14 specialties of clinical medicine, only three achieved accuracy rates above 60.0%. ChatGPT-4 exceeded 60% in every specialty. Table 2 and Fig. 3 present further comparisons of accuracy rates between ChatGPT-3.5 and ChatGPT-4 in other specialties.

3.3. Performance of ChatGPT across question types

Analysis of accuracy rates of ChatGPT-3.5 and ChatGPT-4 across various question types revealed distinct performance levels though such differences did not achieve statistical significance. ChatGPT-3.5 had an accuracy rate of 62.0% (132/213) for positive-choice questions, 58.5% (127/217) for negative-choice questions, 52.7% (29/55) for clinical vignette questions, and 25.0% (1/4) for single-choice multiple-item questions. However, differences in performance across these question types were not statistically significant ($p = 0.307$). ChatGPT-4 demonstrated an accuracy rate of 91.6% (195/213) for positive-choice questions, 92.6% (201/217) for negative-choice questions, 85.5% (47/55) for clinical vignette questions, and 75.0% (3/4)

1. Clinical vignette question

42歲周女士因為類風濕性關節炎，自1年前開始接受腫瘤壞死因子拮抗劑（tumor necrosis factor-alpha inhibitor）adalimumab每2週40 mg注射治療。大概3週前，周女士發現臉上、手臂伸側、上背部及頸部後側開始出現紅疹，驗血發現抗核抗體（antinuclear antibodies, ANA）以及抗雙股DNA抗體（anti-dsDNA antibodies）由用藥前的陰性轉為陽性，下列處置何者最適當？

A.停掉adalimumab，改用etanercept

B.加上prednisolone 1 mg/kg/day，繼續使用adalimumab

C.停掉adalimumab，改用rituximab

D.停掉adalimumab，改用infliximab

2. Single-choice multiple-item question

相較於正常換氣（normal ventilation），過度換氣（hyperventilation）時可見：①肺泡內氧分壓（Po_2）顯著上升 ②肺泡內二氧化碳分壓（Pco_2）顯著下降 ③體動脈血中二氧化碳總含量（total content of CO_2）顯著下降

A.僅①②

B.僅②③

C.僅①③

D.①②③

3. Positive choice question

某病人患有B細胞無法形成漿細胞（plasma cells）之罕見免疫疾病，則其最有可能缺乏下列那一個細胞激素（cytokine）？

A.type 1 interferons

B.type 2 interferons

C.interleukin 2

D.colony-stimulating factors

4. Negative choice question

王大明車禍失血導致平均動脈壓（mean arterial pressure）下降，下列何者與此一現象之發生最不相關？

A.血液總體積大幅下降

B.分布到心臟的副交感神經活性減少

C.周邊靜脈壓減少及靜脈回流減少

D.心室舒張末期容積明顯減少

Fig. 2 Classification of Taiwan's National Physician Licensing Exam questions. Illustrated are four types: clinical vignette, single-choice multiple-item, positive-choice, and negative-choice questions. Clinical vignette questions contain clinical scenarios. Single-choice multiple-item questions require choosing from multiple items in a list. Positive- and negative-choice queries are determined by phrasings such as "Which is correct?" and "Which is incorrect?," respectively.

for single-choice multiple-item questions. Statistical significance was not achieved in this case either (p = 0.245; Table 3, Fig. 4). Nevertheless, the accuracy rate of ChatGPT-4 surpassed that of ChatGPT-3.5 for different types of questions.

4. DISCUSSION

This study investigated the performance of ChatGPT-3.5 and ChatGPT-4 across various specialties in the Taiwan Physician Licensing Exam of 2022. Our findings indicated that ChatGPT-3.5 could pass the first stage of the exam but failed the second. Meanwhile, ChatGPT-4 was not only capable of passing the first and second stages but also scored significantly higher than ChatGPT-3.5. A substantial variance was observed in the models' accuracy rates across different medical specialties, but ChatGPT-4's performance surpassed that of ChatGPT-3.5 in almost every specialty. Furthermore, we found that different types of questions had minimal impact on the accuracy of ChatGPT.

This study demonstrated that while ChatGPT-3.5 could achieve passing scores in basic medical sciences, it failed to meet the criteria for clinical medicine at the second stage. ChatGPT-3.5 performed better in basic medical science specialties, achieving an accuracy rate of 60.0% or higher in all but histology, parasitology, and physiology. This result echoes those of previous studies using ChatGPT-3.5 in medical examinations. Gilson et al[25] observed that ChatGPT model's accuracy rate in USMLE step 1 was higher than in step 2. Talan and Kalinkara[26] further revealed

Table 1

Comparative analysis of numbers, correct responses, and accuracy rates for ChatGPT-3.5 and ChatGPT-4 across various classes in Taiwan's National Physician License Exam, 2022

Class	Total numbers[a]	Correct answers: GPT-3.5	Accuracy rates: GPT-3.5	Correct answers: GPT-4	Accuracy rates: GPT-4	*p*
First stage						
Medical Science I	99	65	65.7	89	89.9	-
Medical Science II	99	69	69.7	93	93.9	-
Average			67.7		91.9	
Second stage						
Medical Science III	76	49	64.5	72	94.7	-
Medical Science IV	69	37	53.6	63	91.3	-
Medical Science V	74	32	43.2	62	83.8	-
Medical Science VI	72	37	51.4	67	93.1	-
Average			53.2		90.7	
Total average			58.0		91.1	0.027

[a]Image-based questions were excluded.

Table 2

Comparative analysis of numbers, correct responses, and accuracy rates for ChatGPT-3.5 and ChatGPT-4 across specialties in Taiwan's National Physician Licensing Exam, 2022

Specialties	Total numbers[a]	Correct answers: GPT-3.5	Accuracy rates: GPT-3.5 (%)	Correct answers: GPT-4	Accuracy rates: GPT-4 (%)
Medical Science I					
Embryology	5	5	100.0	4	80.0
Physiology	27	15	55.6	23	85.2
Anatomy	30	20	66.7	27	90.0
Histology	10	3	30.0	9	90.0
Biochemistry	27	22	81.5	26	96.3
Medical Science II					
Parasitology	7	3	42.9	6	85.7
Pathology	24	19	79.2	21	87.5
Pharmacology	25	19	76.0	24	96.0
Microbiology and Immunology	28	19	67.9	27	96.4
Public Health	15	9	60.0	15	100.0
Medical Science III					
Internal Medicine	67	44	65.7	63	94.0
Family Medicine	9	5	55.6	9	100.0
Medical Science IV					
Psychiatrics	17	9	52.9	15	88.2
Pediatrics	31	19	61.3	28	90.3
Neurology	15	6	40.0	14	93.3
Dermatology	6	3	50.0	6	100.0
Medical Science V					
Urology	8	1	12.5	6	75.0
Surgery	57	28	49.1	48	84.2
Orthopedics	9	3	33.3	8	88.9
Medical Science VI					
Rehabilitation	14	7	50.0	12	85.7
Ophthalmology	9	3	33.3	8	88.9
Gynecology and Obstetrics	33	18	54.6	31	93.9
Otolaryngology	6	4	66.7	6	100.0
Anesthesiology	10	5	50.0	10	100.0

[a]Image-based questions were excluded.

that ChatGPT's performance scores in anatomy exceeded those of undergraduate students. Although a study showed that ChatGPT-3.5's performance in parasitology was inferior to that of medical students in Korea, it still exhibited an accuracy rate of 60.8% as opposed to the medical students' 89.6%.[27] However, it is important to acknowledge that these comparisons may not be fair when medical students from different countries are used as benchmarks for educational systems and assessment methodologies vary significantly across countries.

In terms of performance in clinical medicine, ChatGPT-3.5 showed much variability, achieving accuracy rates above 60.0% in only three of 14 specialties.[9] Its highest accuracy rate was 65.7% (44/67) in internal medicine, while its lowest was 12.5% (1/8) in urology. Although some studies have shown that ChatGPT-3.5's performance did not significantly vary between basic medical and clinical sciences,[28,29] other research on specific specialties revealed inconsistent results. According to Ali et al,[14] ChatGPT-3.5 passed examinations in

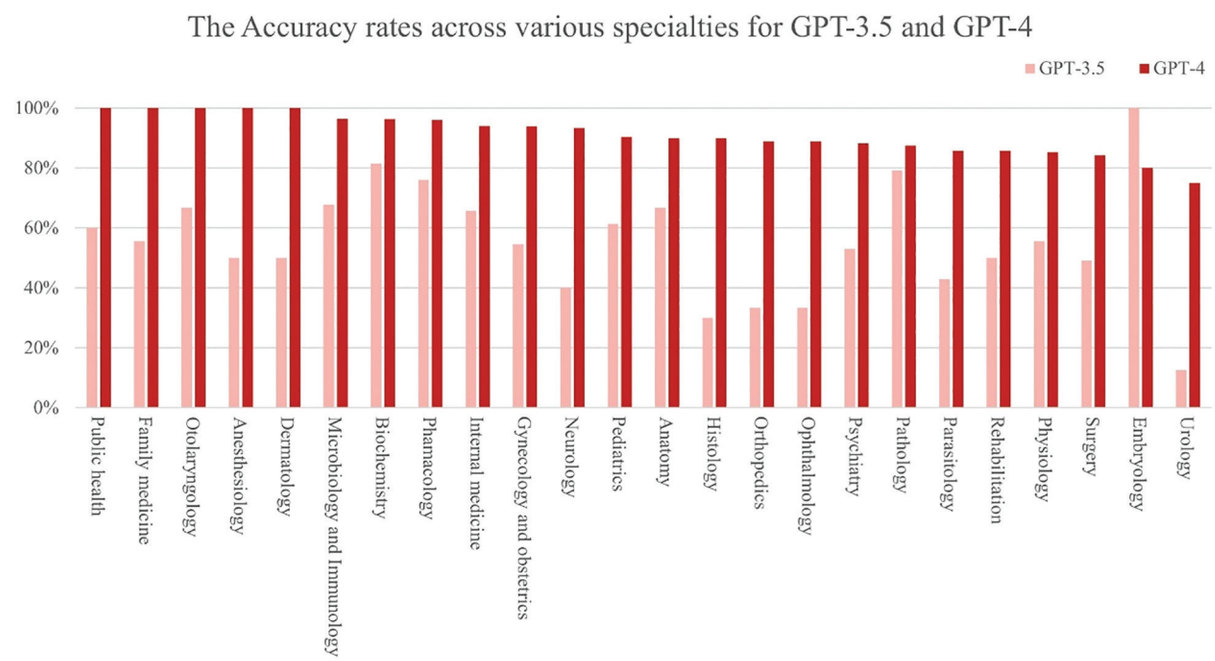

Fig. 3 Accuracy rates across various specialties for ChatGPT-3.5 (light red) and ChatGPT-4 (dark red).

the neurosurgery field. Skalidis et al[30] further demonstrated that ChatGPT could pass the core examinations in European cardiology; however, other studies have shown varying results. Antaki et al[10] revealed that ChatGPT-3.5 had an accuracy rate below 60% in ophthalmology. Additionally, Yeo et al[18] found a significant gap in ChatGPT-3.5's performance between English and non-English questions. Wang et al[15] observed that ChatGPT-3.5 was unable to pass the Chinese National Medical Licensing Examination. Furthermore, a study by Kao et al[22] showed that ChatGPT-3's performance was limited in the field of internal medicine. Besides language, other factors might affect the field, such as epidemiology, medical policies, and laws.[15]

Our results clearly exhibited that ChatGPT-4 demonstrated significantly higher accuracy rates in basic medical science and clinical medicine than ChatGPT-3.5. It scored higher than ChatGPT-3.5 across all specialties except embryology. These findings are consistent with the literature; for instance, Kleinig et al[19] observed that ChatGPT-4's performance in the Australian Medical Council licensing examination surpassed that of ChatGPT-3.5.[19] Even in other languages, ChatGPT-4 exhibited higher accuracy in answering questions and better performance than ChatGPT-3.5.[16–18] This finding is consistent with OpenAI's own claims of ChatGPT-4 demonstrating higher accuracy than its predecessor.[3] The potential integration of ChatGPT in medical education and examination systems seems promising but also poses challenges. While the models showed proficiency in various domains, they also highlighted areas for improvement. These findings underscore the need for continual model development, especially in managing the nuanced and evolving nature of medical knowledge.

Our results suggested that performance across different question types was not significantly different; ChatGPT-4 exhibited higher accuracy rates across all question types compared with ChatGPT-3.5. Oztermeli and Oztermeli[28] as well as Hoch have pointed out that ChatGPT performs better in single-select multiple-choice questions compared with multi-select multiple-choice questions.[31] In contrast, Weng et al[21] indicated that in the context of single-select multiple-choice questions, classification of the question does not significantly affect accuracy rates, which is consistent with our study because there are no multi-select questions in Taiwan's National Physician Licensing Exam. However, it is noteworthy that multi-select multiple-choice questions inherently present greater difficulty than single-select multiple-choice questions, which is reflected in lower accuracy rates. The variance in question types suggests that while ChatGPT can be a valuable educational resource, it cannot replace traditional methods and human expertise in medical training and assessment as yet.[32,33]

Although ChatGPT-4 surpassed ChatGPT-3.5 in almost every specialty, there were still certain questions that ChatGPT-3.5 answered accurately while ChatGPT-4 did not. According to Kleinig et al,[19] ChatGPT-3.5 and ChatGPT-4 might provide different answers to the same question at different times; however, this self-inconsistency did not affect the superiority of ChatGPT-4. In addition, Chen et al[34] found that the performance of ChatGPT-3.5 and ChatGPT-4 changes over time, which means that the same query to ChatGPT can lead to the generation of different responses at different times. Both could lead to variability in ChatGPT's answers. Our findings highlighted several areas for future AI development in the medical field. Enhancing the models' consistency, expanding their understanding of diverse medical specialties, and improving their adaptability to different languages and cultural settings are crucial. Additionally, exploring potential ethical implications and establishing guidelines for AI use in medical settings will be crucial as these technologies become more integrated into healthcare systems.[6,35] Medprompt, a recent prompt architecture that employs techniques including chain of thought, self-consistency, and choice shuffling ensemble, may further address this issue.[36]

This study sheds light on the performance of AI models in non-English-language contexts, particularly traditional Chinese. The results indicate that while ChatGPT can handle linguistic diversity to some extent, challenges persist in the accurate

Table 3

Comparative analysis of numbers, correct responses, and accuracy rates for ChatGPT-3.5 and ChatGPT-4 across question types in Taiwan's National Physician License Exam, 2022

Question types	Total numbers[a]	Correct answers: GPT-3.5	Accuracy rates: GPT-3.5 (%)	*p*[b]	Correct answers: GPT-4	Accuracy rates: GPT-4 (%)	*p*[b]
Overall							
Positive choice	213	132	62.0	0.307	195	91.6	0.245
Negative choice	217	127	58.5		201	92.6	
Clinical vignette	55	29	52.7		47	85.5	
Single-choice multiple-item	4	1	25.0		3	75.0	
Medical Science I							
Positive choice	73	46	63.0	0.112	66	90.4	0.941
Negative choice	24	19	79.2		21	87.5	
Clinical vignette	1	0	0.0		1	100.0	
Single-choice multiple-item	1	0	0.0		1	100.0	
Medical Science II							
Positive choice	52	40	76.9	0.114	48	92.3	<0.001
Negative choice	41	27	65.9		40	97.6	
Clinical vignette	5	2	40.0		5	100.0	
Single-choice multiple-item	1	0	0.0		0	0.0	
Medical Science III							
Positive choice	21	13	61.9	0.888	20	95.2	0.273
Negative choice	42	27	64.3		41	97.6	
Clinical vignette	12	8	66.7		10	83.3	
Single-choice multiple-item	1	1	100.0		1	100.0	
Medical Science IV							
Positive choice	23	14	60.9	0.304	21	91.3	0.673
Negative choice	32	18	56.3		30	93.8	
Clinical vignette	14	5	35.7		12	85.7	
Single-choice multiple-item	0	0	-		0	-	
Medical Science V							
Positive choice	17	10	58.8	0.078	17	100.0	0.200
Negative choice	43	16	37.2		34	79.1	
Clinical vignette	13	6	46.2		10	76.9	
Single-choice multiple-item	1	0	0.0		1	100.0	
Positive choice	27	9	33.3	0.027	23	85.2	0.069
Negative choice	35	20	57.1		35	100.0	
Clinical vignette	10	8	80.0		9	90.0	
Single-choice multiple-item	0	0	-		0	-	

[a]Image-based questions were excluded.

[b]Fisher's exact test was used to compare accuracy rates across question types.

interpretation and response to language-specific nuances. This finding is compatible with studies in other non-English countries.[18,22,37–39] It underscores the importance of developing AI models that are inclusive and adaptable to various linguistic and cultural contexts.

This study has certain limitations. First, it involved entering Chinese exam questions into ChatGPT. This linguistic focus provides a unique perspective but may not fully capture the models' capabilities in other languages or cultural contexts. It remains unclear whether ChatGPT's errors are attributable to language comprehension or conceptual understanding. Second, some specialties feature a limited number of questions, so ChatGPT's accuracy rate in these specialties might be not representative. Third, we used exam questions from 2022 because of pretrained data considerations, which might introduce biases. While such an approach minimizes the risk of the models having prior exposure to the questions, it limits the generalizability of the findings. The models' performance over a broader range of years and question sets might yield different insights. Fourth, during our research period (August 10–20, 2023), ChatGPT-4 was unable to read images, so we excluded questions containing those. This exclusion is a significant limitation given the importance of visual data in medical diagnostics and education. Fifth, we could not determine how ChatGPT operates exactly, including its parameters, self-inconsistency, and behavior change over time, which might lead to inconsistent results with time. Additionally, the study's scope was limited to two versions of the ChatGPT model, which may not fully represent the entire spectrum of AI capabilities in medical contexts. Finally, this study did not assess the inconsistency between repeated querying. A more sophisticated framework is needed to resolve such scenarios.

In conclusion, this study found that ChatGPT-3.5 is proficient in answering questions on basic medical sciences but falls short in accurately responding to clinical medicine questions. Meanwhile, ChatGPT-4 demonstrates not only competence in basic and clinical medical sciences but also the ability to handle questions in traditional Chinese, highlighting its linguistic versatility. However, its proficiency may vary across different specialties. Question type did not significantly affect ChatGPT's performance in answering medical questions. Discrepancies in performance across various medical specialties emphasize the need for a cautious and informed implementation of these technologies in educational settings.

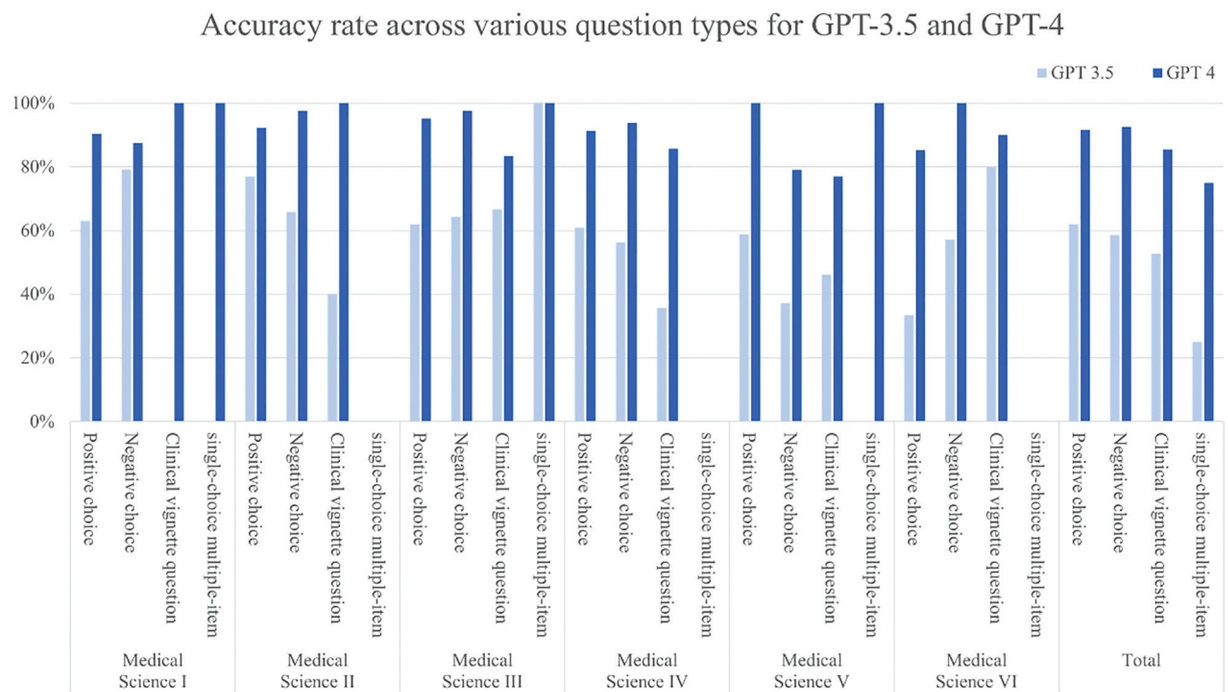

Fig. 4 Accuracy rates across various question types for ChatGPT-3.5 (light blue) and ChatGPT-4 (deep blue). No significant difference is observed between the two in terms of their performance across different question types.

APPENDIX A. SUPPLEMENTARY DATA

Supplementary data related to this article can be found at http://links.lww.com/JCMA/A319.

REFERENCES

1. Harris E. Study tests large language models' ability to answer clinical questions. *JAMA* 2023;330:496.
2. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N Engl J Med* 2023;388:1233–9.
3. Open AI. Online ChatGPT—Optimizing language models for dialogue. Available at https://online-chatgpt.com/. Accessed August 25, 2023
4. 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.
5. Arif TB, Munaf U, Ul-Haque I. The future of medical education and research: is ChatGPT a blessing or blight in disguise? *Med Educ Online* 2023;28:Artn 2181052.
6. Hswen Y, Abbasi J. AI will-and should-change medical school, says Harvard's Dean for medical education. *JAMA* 2023;330:1820–3.
7. Sallam M, Salim N, Barakat M, Al-Tammemi A. ChatGPT applications in medical, dental, pharmacy, and public health education: a descriptive study highlighting the advantages and limitations. *Narra J* 2023;3:e103.
8. Kung TH, Cheatham M, Medenilla A, Sillos C, Leon LD, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health* 2023;2:e0000198.
9. Levin G, Horesh N, Brezinov Y, Meyer R. Performance of ChatGPT in medical examinations: a systematic review and a meta-analysis. *Bjog* 2024;131:378–80.
10. Antaki F, Touma S, Milad D, El-Khoury J, Duval R. Evaluating the performance of ChatGPT in ophthalmology: An analysis of its successes and shortcomings. *Ophthalmol Sci* 2023;3:100324.
11. Fijačko N, Gosak L, Štiglic G, Picard CT, John Douma M. Can ChatGPT pass the life support exams without entering the American Heart Association course? *Resuscitation* 2023;185:109732.
12. Lum ZC. Can artificial intelligence pass the American Board of Orthopaedic Surgery examination? Orthopaedic residents versus ChatGPT. *Clin Orthop Relat Res* 2023;481:1623–30.
13. Sabry Abdel-Messih M, Kamel Boulos MN. ChatGPT in clinical toxicology. *JMIR Med Educ* 2023;9:e46876.
14. 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;93:1353–65.
15. Wang X, Gong Z, Wang G, Jia J, Xu Y, Zhao J, et al. ChatGPT performs on the Chinese national medical licensing examination. *J Med Syst* 2023;47:86.
16. Kasai J, Kasai Y, Sakaguchi K, Yamada Y, Radev D. Evaluating gpt-4 and chatgpt on Japanese medical licensing examinations. arXiv preprint arXiv:2303.18027.
17. Khorshidi H, Mohammadi A, Yousem DM, Abolghasemi J, Ansari G, Mirza-Aghazadeh-Attari M, et al. Application of ChatGPT in multilingual medical education: how does ChatGPT fare in 2023's Iranian residency entrance examination. *Inf Med Unlocked* 2023;41:101314.
18. Yeo YH, Samaan JS, Ng WH, Ma X, Ting PS, Kwak MS, et al. GPT-4 outperforms ChatGPT in answering non-English questions related to cirrhosis. *medRxiv* 2023:2023.05.04.23289482.
19. Kleinig O, Gao C, Bacchi S. This too shall pass: The performance of ChatGPT-3.5, ChatGPT-4 and New Bing in an Australian medical licensing examination. *Med J Aust* 2023;219:237.
20. Wang YM, Shen HW, Chen TJ. Performance of ChatGPT on the pharmacist licensing examination in Taiwan. *J Chin Med Assoc* 2023;86:653–8.
21. Weng TL, Wang YM, Chang S, Chen TJ, Hwang SJ. ChatGPT failed Taiwan's family medicine board exam. *J Chin Med Assoc* 2023;86:762–6.
22. Kao YS, Chuang WK, Yang J. Use of ChatGPT on Taiwan's examination for medical doctors. *Ann Biomed Eng* 2024;52:455–7.
23. Ministry of Examination. A platform for querying exam questions. Taiwan, Ministry of Examination R.O.C. Available at https://wwwq.moex.gov.tw/exam/wFrmExamQandASearch.aspx?y=2022&e=111110. Accessed April 23, 2023
24. Ministry of Examination. List for archived exam questions. Available at https://wwwq.moex.gov.tw/exam/wFrmExamQandASearch.aspx. Accessed August 10, 2023
25. 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? The implications of large language models for medical education and knowledge assessment. *JMIR Med Educ* 2023;9:e45312.
26. Talan T, Kalinkara Y. The role of artificial intelligence in higher education: ChatGPT assessment for anatomy course. *Int J Manag Inf Syst Comput Sci* 2023;7:33–40.
27. 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.
28. Oztermeli AD, Oztermeli A. ChatGPT performance in the medical specialty exam: an observational study. *Medicine (Baltim ore)* 2023;102:e34673.
29. Meo SA, Al-Masri AA, Alotaibi M, Meo MZS, Meo MOS. ChatGPT knowledge evaluation in basic and clinical medical sciences: multiple choice question examination-based performance. *Healthcare (Basel)* 2023;11:2046.
30. Skalidis I, Cagnina A, Luangphiphat W, Mahendiran T, Muller O, Abbe E, et al. ChatGPT takes on the European Exam in Core Cardiology: an artificial intelligence success story? *Eur Heart J Digit Health* 2023;4:279–81.
31. Hoch CC, Wollenberg B, Lüers JC, Knoedler S, Knoedler L, Frank K, et al. ChatGPT's quiz skills in different otolaryngology subspecialties: an analysis of 2576 single-choice and multiple-choice board certification preparation questions. *Eur Arch Otorhinolaryngol* 2023;280:4271–8.
32. Mihalache A, Huang RS, Popovic MM, Muni RH. ChatGPT-4: An assessment of an upgraded artificial intelligence chatbot in the United States Medical Licensing Examination. *Med Teach* 2023;46:366–72.
33. Li J, Gui L, Zhou Y, West D, Aloisi C, He Y. Distilling ChatGPT for explainable automated student answer assessment. arXiv:2305.12962.
34. Chen L, Zaharia M, Zou J. How is ChatGPT's behavior changing over time? Harvard Data Science Review, 6.
35. Karabacak M, Ozkara BB, Margetis K, Wintermark M, Bisdas S. The advent of generative language models in medical education. *JMIR Med Educ* 2023;9:e48163.
36. Nori H, Lee YT, Zhang S, Carignan D, Edgar R, Fusi N, et al. Can generalist foundation models outcompete special-purpose tuning? Case study in medicine. arXiv preprint arXiv:2311.16452.
37. Tong W, Guan Y, Chen J, Huang X, Zhong Y, Zhang C, et al. Artificial intelligence in global health equity: an evaluation and discussion on the application of ChatGPT, in the Chinese National Medical Licensing Examination. Article. *Front Med* 2023;10:71237432.
38. Tanaka Y, Nakata T, Aiga K, Etani T, Muramatsu R, Katagiri S, et al. Performance of generative pretrained transformer on the national medical licensing examination in Japan. *medRxiv* 2023:2023.04.17.23288603.
39. Fang C, Wu Y, Fu W, Ling J, Wang Y, Liu X, et al. How does ChatGPT4 preform on Non-English National Medical Licensing Examination? An evaluation in Chinese Language. *medRxiv* 2023:2023.05.03.23289443.