

Original Paper

Performance of ChatGPT-3.5 and ChatGPT-4 in the Taiwan National Pharmacist Licensing Examination: Comparative Evaluation Study

Ying-Mei Wang^{1,2,3,4}, MBA; Hung-Wei Shen^{1,2,4}, MBA; Tzeng-Ji Chen^{5,6,7}, Dr Med; Shu-Chiung Chiang^{1,8}, PhD; Ting-Guan Lin^{2,4}, BS

¹Department of Medical Education and Research, Taipei Veterans General Hospital Hsinchu Branch, Hsinchu, Taiwan

²Department of Pharmacy, Taipei Veterans General Hospital Hsinchu Branch, Hsinchu, Taiwan

³School of Medicine, National Tsing Hua University, Hsinchu, Taiwan

⁴Hsinchu County Pharmacists Association, Hsinchu, Taiwan

⁵Department of Family Medicine, Taipei Veterans General Hospital Hsinchu Branch, Hsinchu, Taiwan

⁶Department of Family Medicine, Taipei Veterans General Hospital, Taipei, Taiwan

⁷Department of Post-Baccalaureate Medicine, National Chung Hsing University, Taichung, Taiwan

⁸Institute of Hospital and Health Care Administration, School of Medicine, National Yang Ming Chiao Tung University, Taipei, Taiwan

Corresponding Author:

Ying-Mei Wang, MBA
Department of Medical Education and Research
Taipei Veterans General Hospital Hsinchu Branch
81, Section 1, Zhongfeng Road, Zhudong
Hsinchu, 310
Taiwan
Phone: 886 03-5962134 ext 127
Email: teresawang@vhct.gov.tw

Abstract

Background: OpenAI released versions ChatGPT-3.5 and GPT-4 between 2022 and 2023. GPT-3.5 has demonstrated proficiency in various examinations, particularly the United States Medical Licensing Examination. However, GPT-4 has more advanced capabilities.

Objective: This study aims to examine the efficacy of GPT-3.5 and GPT-4 within the Taiwan National Pharmacist Licensing Examination and to ascertain their utility and potential application in clinical pharmacy and education.

Methods: The pharmacist examination in Taiwan consists of 2 stages: basic subjects and clinical subjects. In this study, exam questions were manually fed into the GPT-3.5 and GPT-4 models, and their responses were recorded; graphic-based questions were excluded. This study encompassed three steps: (1) determining the answering accuracy of GPT-3.5 and GPT-4, (2) categorizing question types and observing differences in model performance across these categories, and (3) comparing model performance on calculation and situational questions. Microsoft Excel and R software were used for statistical analyses.

Results: GPT-4 achieved an accuracy rate of 72.9%, overshadowing GPT-3.5, which achieved 59.1% ($P < .001$). In the basic subjects category, GPT-4 significantly outperformed GPT-3.5 (73.4% vs 53.2%; $P < .001$). However, in clinical subjects, only minor differences in accuracy were observed. Specifically, GPT-4 outperformed GPT-3.5 in the calculation and situational questions.

Conclusions: This study demonstrates that GPT-4 outperforms GPT-3.5 in the Taiwan National Pharmacist Licensing Examination, particularly in basic subjects. While GPT-4 shows potential for use in clinical practice and pharmacy education, its limitations warrant caution. Future research should focus on refining prompts, improving model stability, integrating medical databases, and designing questions that better assess student competence and minimize guessing.

JMIR Med Educ 2025;11:e56850; doi: [10.2196/56850](https://doi.org/10.2196/56850)

Keywords: artificial intelligence; ChatGPT; chat generative pre-trained transformer; GPT-4; medical education; educational measurement; pharmacy licensure; Taiwan; Taiwan national pharmacist licensing examination; learning model; AI; Chatbot; pharmacist; evaluation and comparison study; pharmacy; statistical analyses; medical databases; medical decision-making; generative AI; machine learning

Introduction

Background

With the advent of the artificial intelligence (AI) era, applications of AI in the medical field have increased with ChatGPT (OpenAI) being the most notable examples. ChatGPT is a large language model based on a generative pretrained transformer developed by OpenAI. ChatGPT-3.5 (GPT-3.5) was the first publicly accessible version, while ChatGPT-4 (GPT-4) was the subscription version. GPT-4 surpasses GPT-3.5 in advanced reasoning, almost nearing human-level performance in professional and academic examinations [1,2]. For instance, GPT-4 ranked in the top 10% of scores on a law examination, whereas GPT-3.5 ranked in the bottom 10% [3]. Additionally, GPT-3.5 resolved 90% of false-belief tasks, achieving the level of a 7-year-old child, whereas GPT-4 resolved 95% of these tasks [4]. Following its launch, ChatGPT has been extensively studied and discussed in both the medical and educational fields [5]. The most widely recognized performance of GPT-3.5 has been on the United States Medical Licensing Examination (USMLE) [6,7]; however, GPT-3.5's performance did not meet expectations in other examinations [8-11]. Gradually, Nori et al [12] observed that the accuracy of GPT-4 was higher than that of the GPT-3.5 on the USMLE, and further studies confirmed that GPT-4 outperforms GPT-3.5 [13-16]. However, there has been limited research on its performance in pharmacy examinations.

In the field of pharmacy, GPT-3.5 has exhibited commendable performance in clinical toxicology and pharmacology [17,18], although it has not passed the National Pharmacist Licensing Examination (NPLE) in Taiwan [19]. However, GPT-4 has outperformed GPT-3.5 in drug information [20] and China's Pharmacist Licensing Examination [21]. Generative AI models, a large language model, has been applied in drug development and novel drug design [22-24], pharmacovigilance [25,26], pharmacokinetic model development [27], pharmacy education, and research writing [28,29].

Goal of the Study

According to previous studies, GPT-3.5 failed to pass the NPLE, indicating its limitations in pharmacy education. Based on these findings, we hypothesized that GPT-4 would outperform GPT-3.5 in this context, demonstrating greater proficiency. To test this hypothesis, this study compared the performance of GPT-3.5 and GPT-4 on Taiwan's NPLE. Additionally, we conducted a comprehensive assessment of their performance across various question types, with a focus on pharmacy-related tasks such as pharmacokinetic calculation and clinical decision-making scenarios. This analysis aims to determine the practical applications of GPT-4 in pharmacy education and establish guidelines for its optimal use in this field.

Methods

Background

The NPLE in Taiwan is divided into 2 stages. The first stage focuses on 3 basic subjects: pharmacology and pharmaceutical chemistry, pharmaceutical analysis and pharmacognosy (including traditional Chinese medicine), and pharmaceuticals and biopharmaceuticals. The second stage focuses on 3 clinical subjects: dispensing and clinical pharmacy, pharmacotherapy, and pharmacy administration and pharmacy law. The first and second stages of the examination have 240 and 210 multiple-choice questions, respectively. Pharmacy students typically complete the first-stage exam after completing their third year of university coursework. They become eligible for the second-stage exam only after passing the first examination, completing their internships and obtaining their graduation certificates. After passing the second-stage examination, candidates receive their pharmacist certificate, allowing them to practice as a pharmacist legally.

Data Source

This study used the 2-stage NPLE questions released by the Ministry of Examination in February 2023, with each subject exam lasting for 1 hour. The version of NPLE used in this study was the most recent available at the time of research. We used both GPT-3.5 (free version) and GPT-4 (licensed version). No temperature settings were applied. Examination questions were manually fed into GPT-4 and GPT-3.5 sequentially. To simulate student responses, complete questions were entered into the models without tailored prompts. One question was input at a time, and the responses were recorded for analysis. Since GPT-3.5 cannot process images and image functionality of GPT-4 was unavailable during the analysis, only text-based questions were used. Questions containing graphics, such as chemical structures, tables, symbols, and formulas were excluded. Both models were presented with the same set of questions under identical conditions. Due to the limitations on the number of times the model could be used and required cooling time between queries, all questions were answered sequentially and not timed to avoid any potential bias introduced by time constraints.

Study Design

The study was divided into 3 parts; the first part compared the accuracy of GPT-4 and GPT-3.5, as well as in different subjects. The second part compared the accuracy of GPT-4 and GPT-3.5 across different question types. These questions were categorized into 4 types: memory-based questions (1 correct word answer out of 4 options, low-level thinking; Figure 1), judgment questions (1 correct statement out of 4, medium-level thinking; Figure 2), reverse questions (1 incorrect statement out of 4, medium to high-level thinking;

Figure 3), and comprehension questions (multiple-choice or matching types, high-level thinking; Figure 4). One pharmacist classified the questions according to these established categories and the second pharmacist reviewed the classifications. In the event of disagreement, a third pharmacist was consulted for the final decision. All pharmacists had

over 10 years of experience in medical center hospitals or community teaching hospitals. The third part compared the accuracy of GPT-4 and GPT-3.5 for calculation-based and case scenario questions (Figure 5). Model testing for this study was conducted from May 10 to July 20, 2023.

Figure 1. Template of a memory-based question (choose 1 correct word from 4 options, requiring low-level thinking).

Model: GPT-4

c0 10. 抗凝血 (anticoagulant) 口服藥物rivaroxaban，其主要作用在於抑制下列何種凝血因子？

- A.factor Xa
- B.factor VIIa
- C.factor IV
- D.factor IIa

Figure 2. Template of a judgment question (choose 1 correct statement from 4 options, requiring medium-level thinking).

Model: GPT-4

c0 2. 下列何者是partial agonist的性質？

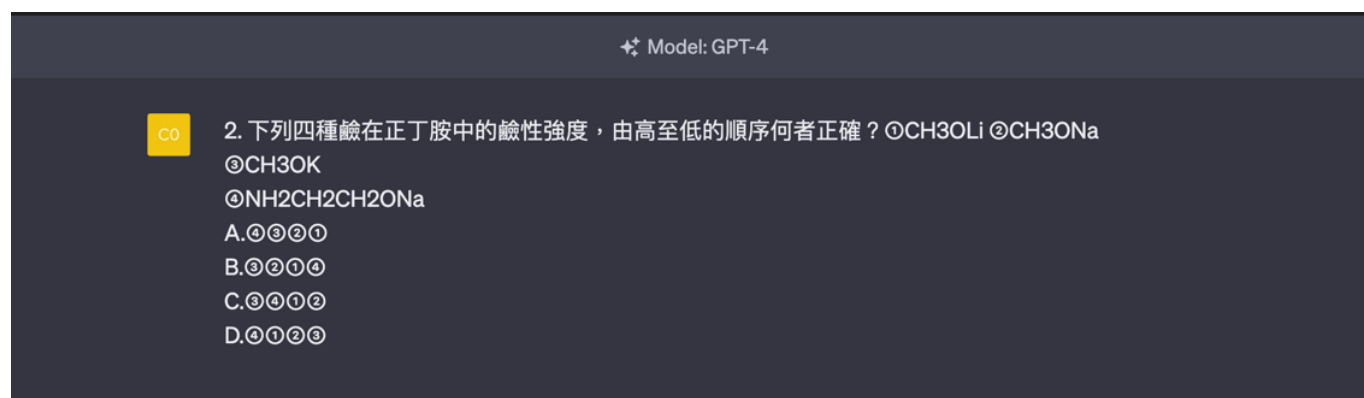
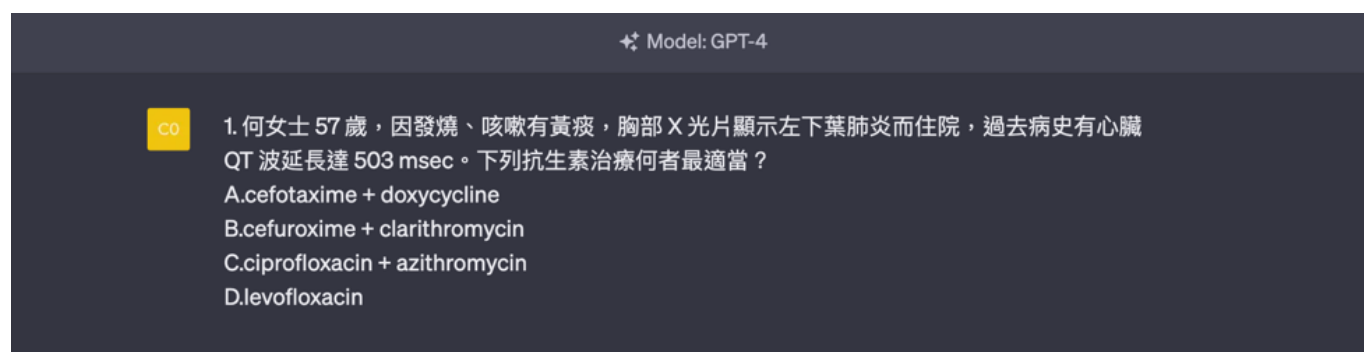
- A.藥理作用會受antagonist拮抗
- B.作用位置是在受體的allosteric site
- C.不具有intrinsic activity
- D.增加濃度可以達到full agonist的作用

Figure 3. Template of a reverse question (choose 1 incorrect statement from 4 options, requiring medium- to high- level thinking).

Model: GPT-4

c0 11. 下列有關降血脂fibrates類藥物之敘述，何者錯誤？

- A.具有加強抗凝血藥物 (anticoagulant) 之作用
- B.fenofibrate可與statin一起併用
- C.適用於肝臟及腎臟缺損之病人
- D.可導致膽結石 (gallstones) 形成

Figure 4. Template of a comprehension questions (multiple-choice or matching types, requiring high- level thinking).**Figure 5.** Template of a case scenario question.

Statistical Analysis

Microsoft Excel 2019 was used to compare the accuracy rates of the 2 models. χ^2 tests were used to compare the overall accuracy rates of answers obtained using GPT-3.5 and GPT-4. McNemar tests were used to compare the consistency in answers between GPT-3.5 and GPT-4, and for the calculation-based and situational question types using R software (version 4.2.2; R Foundation for Statistical Computing).

Ethical Considerations

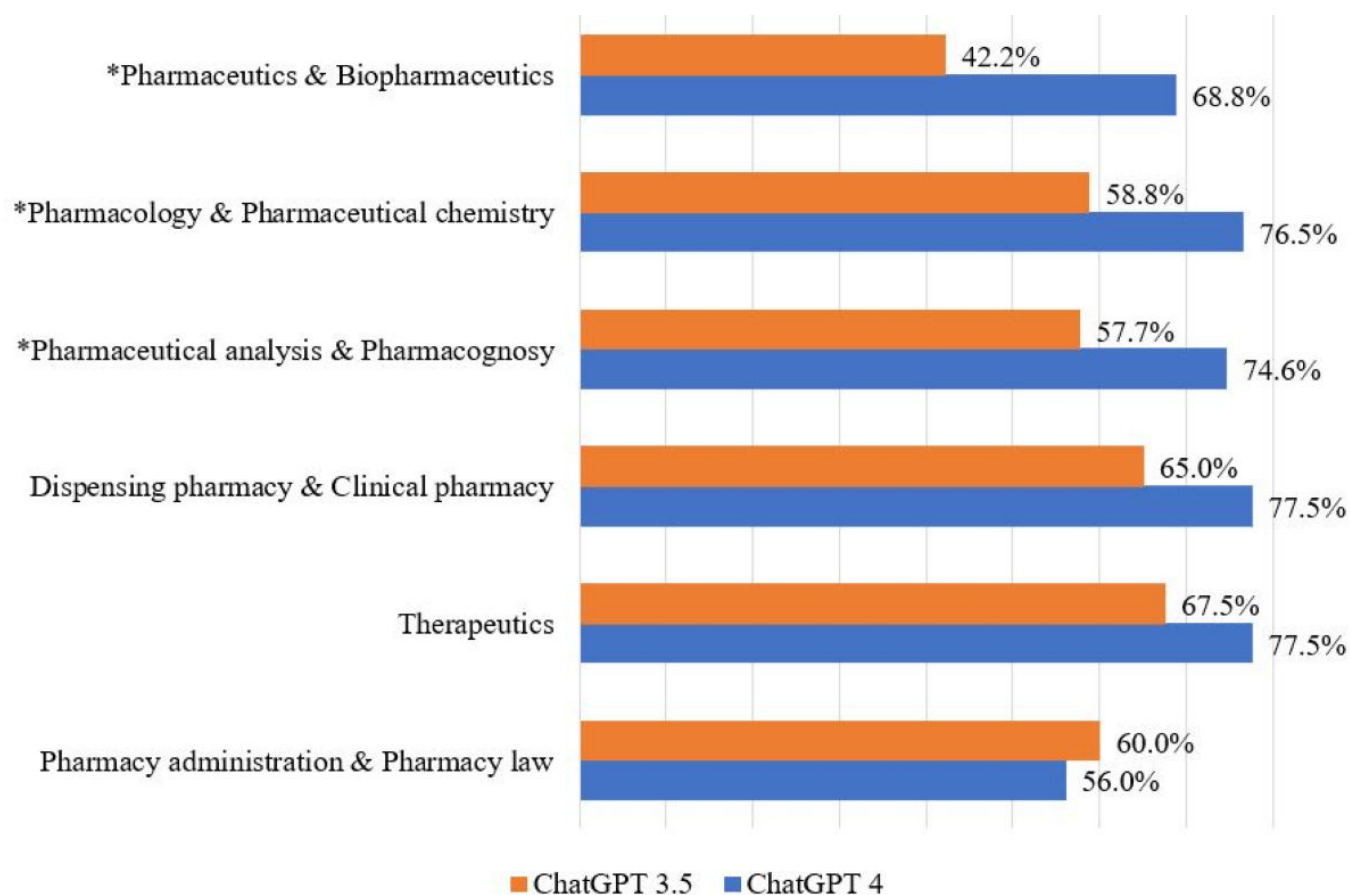
This study involved comparing the performance of ChatGPT-4 and ChatGPT-3.5 in the pharmacist licensing examination. It did not involve human participants. As per the guidelines of the 'Human Research Cases Exempted from Ethics Review Board' issued by the Ministry of Health and Welfare, Taiwan, this study was exempted from Ethics Review Board analysis.

Results

Accuracy in Different Subjects

In total, 203 and 210 questions were included for analysis from the first- and second-stage examinations, respectively,

after excluding 37 questions containing graphical elements (N=413) (Figure 6). GPT-4 had an overall accuracy of 72.9% (301/413), easily passing the test (60% threshold) and outperforming GPT-3.5 which achieved an accuracy of 59.1% (244/413; $P<.001$). In terms of accuracy by stage, GPT-4's overall accuracy was significantly higher than that of GPT-3.5 (73.4% vs 53.2% or 149/203 vs 108/203; $P<.001$) in basic subjects of the first stage. GPT-4 also significantly outperformed GPT-3.5 in each of the 3 basic subjects. In the clinical subjects of the second stage, GPT-4's accuracy was higher but not statistically significant than that of GPT-3.5 (72.4% vs 64.8% or 152/210 vs 136/210; $P=.096$). In pharmacy administration and pharmacy law, GPT-4's accuracy was lower than that of GPT-3.5 (56% vs 60% or 28/50 vs 30/50; $P=.96$). Among individual subjects, significant differences were observed in pharmacology and pharmaceutical chemistry ($P=.02$), pharmaceutical analysis and pharmacognosy ($P=.02$), and pharmaceuticals and biopharmaceuticals ($P=.002$). No significant differences were noted in dispensing pharmacy and clinical pharmacy ($P=.07$), pharmacotherapeutics ($P=.10$), and pharmacy administration and pharmacy law ($P=.48$).

Figure 6. Accuracy comparison of ChatGPT-3.5 and ChatGPT-4 across different subjects. * $P<.05$.

The overall consistency among answers significantly differed between the 2 models (68%, $P<.001$), with GPT-4 showing consistent correct answers in 49.4% ($n=204$) of cases and consistent incorrect answers in 18.6% ($n=77$) of cases (Table 1).

Table 1. Performance comparison of consistency between ChatGPT-3.5 and ChatGPT-4.

ChatGPT-3.5 responses	GPT-4	
	Correct answers, n (%)	Incorrect answers, n (%)
Correct answer	204 (49.4)	38 (9.2)
Incorrect answer	94 (22.8)	77 (18.6)

Accuracy in Different Question Types

Among the 413 examination questions analyzed, memory-based questions were the most common ($n=254$, 61.5%), followed by judgment questions ($n=82$, 19.9%), reverse questions ($n=46$, 11.1%), and comprehension questions

($n=31$, 7.5%). GPT-4 and GPT-3.5 did not differ significantly in terms of accuracy of answers between question types ($P=.461$ vs $P=.18$; Table 2). GPT-4 is significantly better than GPT-3.5 in memory-based questions ($P<.001$) and comprehension-based questions ($P=.03$).

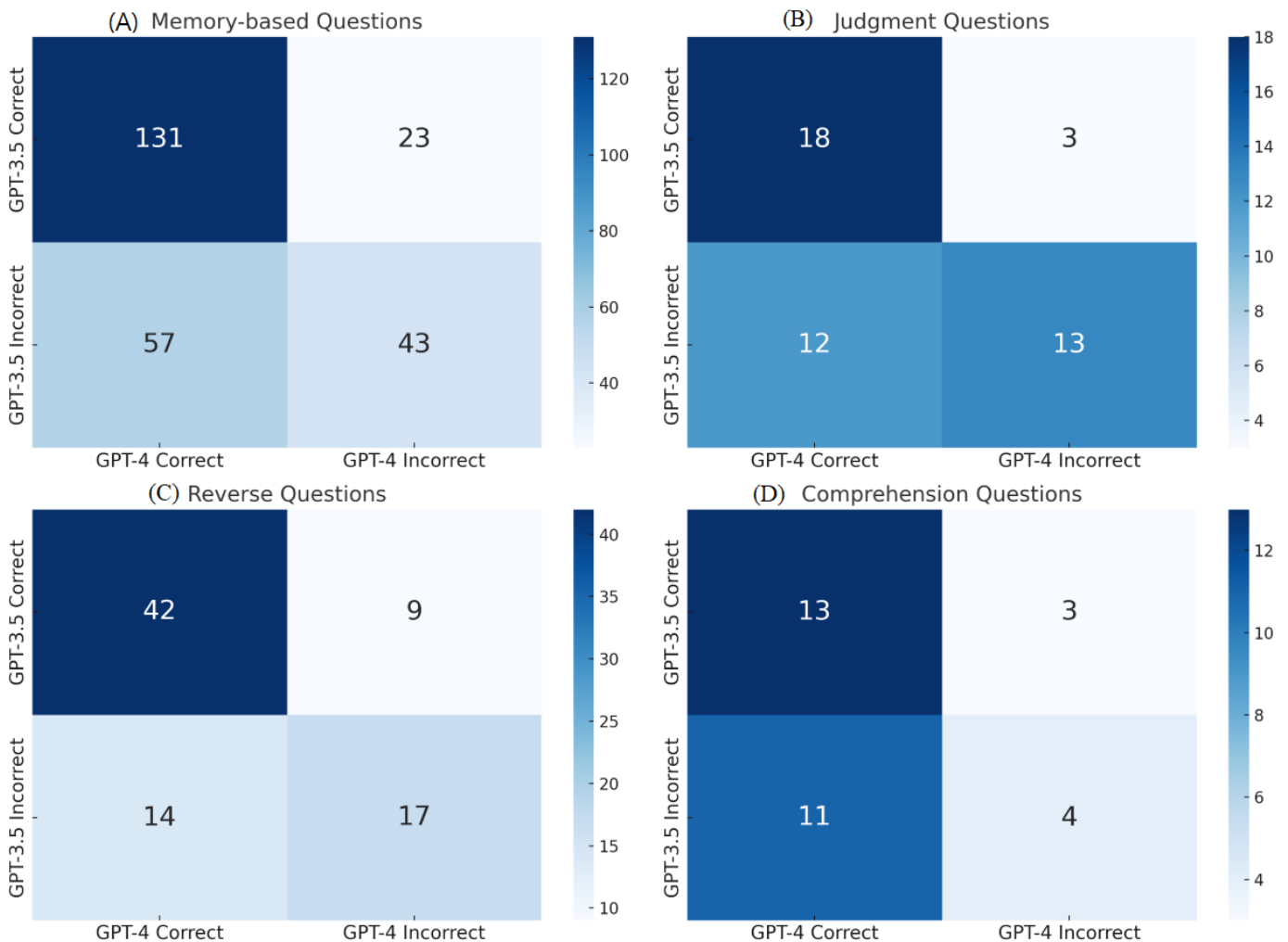
Table 2. Accuracy comparison of ChatGPT-3.5 and ChatGPT-4 by question type.

Question type	GPT-3.5 Correct answers, n (%)	GPT-4 Correct answers, n (%)	Total, n (%)	P value
Memory-based questions	155 (61)	188 (74)	254 (61.5)	<.001 ^a
Judgment questions	21 (45.7)	30 (65.2)	46 (11.1)	.06
Reverse questions	51 (62.6)	56 (68.3)	82 (19.9)	.41
Comprehension questions	16 (51.6)	24 (77.4)	31 (7.5)	.03 ^a

^a $P<.05$.

Figure 7 shows the performance comparison of GPT-3.5 and GPT-4 across question types. The data provided insights into the relative strengths and weaknesses of each model.

Figure 7. Performance comparison of GPT-3.5 and GPT-4 across question types (A) memory-based, (B) judgement, (C) reverse, and (D) comprehension. The heatmaps display the number of answers, with darker shades indicating higher counts of correct responses and highlighting model performance.



Further analysis of the discrepancies between the models revealed no significant difference in questions answered incorrectly by GPT-3.5 but correctly by GPT-4 ($n=94$) and vice versa ($n=38$) across the 4 question types ($P=.27$ vs $P=.95$).

For calculation-based questions, GPT-4 showed higher accuracy than that of GPT-3.5 (80% vs 40%, $P=.03$), with the most pronounced difference in pharmaceuticals and biopharmaceuticals subjects. In scenario-based questions, GPT-4 also outperformed GPT-3.5 in terms of accuracy (63% vs 44.4%, $P=.41$), though the difference was nonsignificant.

Discussion

Principal Findings

This study demonstrates that GPT-4 significantly outperformed GPT-3.5 in the Taiwan NPLE, surpassing the passing threshold, especially in basic pharmacy subjects. These subjects, which have only a 13.82% passing rate among human students, are particularly challenging. GPT-4 excelled in areas such as pharmacology, pharmaceutical chemistry, pharmaceutical analysis, and pharmaceuticals, consistently providing correct answers and comprehensive explanations.

Although GPT-4 also performed better than GPT-3.5 in clinical subjects such as dispensing pharmacy and therapeutics, the performance gap was narrower in these areas.

In specific subjects like pharmacodynamics, pharmacokinetics, and drug-related topics in the autonomic nervous system, GPT-4 consistently provided accurate responses, where GPT-3.5 often faltered. Additionally, GPT-4 exhibited superior accuracy in bioavailability, dosing, and pharmacokinetic calculations. However, GPT-4's accuracy dropped in topics like herbal medicines and pharmacy law, emphasizing the need for further model refinement in these areas [30].

Comparison with Literature

Previous studies have established that GPT-4 consistently outperforms GPT-3.5 in various medical exams, including the Australian Medical Licensing Examination [31], Canadian Radiology Examination [15], Turkish Medical Examination [32], and Japanese Medical Licensing Examination [33]. In many of these examinations, GPT-4 consistently achieved scores above 70% [34-36]. This study aligns with those findings, showing GPT-4's superior performance in the Taiwan NPLE. Unlike prior research that focused on real-world clinical applications [37-43], this

study comprehensively assessed the models across various pharmacy domains.

A study by Choi [44] reported that GPT-3.5 performed well on memory-based questions but struggled with problem-solving, whereas GPT-4 demonstrated better performance in comprehension and judgment tasks. Similarly, a radiology study suggested that GPT-4 outperformed GPT-3.5 on higher-order thinking questions but not on lower-order questions [15]. These findings slightly differ from the results of our study, where GPT-3.5 exhibited higher accuracy in both memory-based (low-level thinking) and reverse (mid-level thinking) questions. However, GPT-4 surpassed GPT-3.5 across all question types, particularly in comprehension (high-level thinking) and memory-based (low-level thinking) questions. In judgment, reverse, and comprehension questions—tasks that demand more advanced reasoning

—GPT-4 demonstrated superior accuracy with fewer errors compared to GPT-3.5. Additionally, GPT-4's ability to correct errors made by GPT-3.5 reinforces its potential as a more reliable model for pharmacy-related assessments.

Further, GPT-4 significantly outperformed GPT-3.5 in calculation questions. While GPT-3.5 provided step-by-step explanations but often guessed the final answer—a phenomenon known as hallucination—due to insufficient training—GPT-4 exhibited stronger logical reasoning (Figure 8) with over 80% accuracy. However, it still made errors in 20% of cases, indicating the need for needed during its use [21,45]. In clinical applications, modifying prompts has been shown to improve GPT's accuracy [46]. For integrated analysis questions, GPT-4's performance was slightly better than GPT-3.5, consistent with findings from a nursing licensure examination in Japan [14].

Figure 8. Template of the questions that GPT-4 exhibited stronger logical reasoning.

Model: GPT-4

76. 張先生體重 60 kg，若每隔 8 小時口服投與 20 mg 某藥物 ($F = 0.7$, $V_d = 20\text{ L}$)，其半衰期為 48 小時，當生理狀態沒變，改為每隔 12 小時口服投與 40 mg，則欲達 95% 穩定狀態濃度至少約投藥若干次？

A.5
B.10
C.15
D.20

在藥物動力學中，我們通常認為需要大約 5 個半衰期才能達到穩定狀態濃度，這是一種簡化的模型，但在大多數情況下都適用。在這個問題中，某藥物的半衰期為 48 小時。

所以，張先生需要投藥 $5 * 48\text{ 小時} = 240\text{ 小時}$ 才能達到穩定狀態濃度。

然後，我們要把這個時間轉換成投藥次數。如果他每隔 12 小時口服一次，那麼他需要投藥 $240 / 12 = 20$ 次。

所以，答案是 D.20。

Implications for Education

The study highlights GPT-4's potential as an educational tool, particularly in pharmacy education. GPT-4 can offer extensive practice opportunities for pharmacy students across both basic and clinical subjects, providing both correct answers and detailed explanations [18,47] to enhance understanding. Given the lower passing rates among pharmacy students in basic subjects among that were challenging, GPT-4 could assist in individualized learning. Its strength in comprehension and integrated analysis questions makes it a valuable resource for fostering critical thinking skills.

Despite its advancements over GPT-3.5, GPT-4's occasional inconsistencies suggest that model stability is not yet perfect. Questions correctly answered by GPT-3.5 were not always consistently answered by GPT-4. Nevertheless, GPT-4's accuracy, approaching 80% suggests that it can serve as an effective learning supplement, provided educators guide students in minimizing potential errors. For instance, specifying clearer prompts, such as "Please do not add your own opinions", may help mitigate hallucinations and enhance its use in educational settings.

In addition, educators should consider adjusting the format of examinations by replacing memory-based questions with comprehension questions, which can reduce the chances of guessing and better assess students' true intelligence.

Limitations

The primary limitation of this study is the time frame during which the models were tested (ie, from May 10 to July 20, 2023), which may affect the reproducibility of the results if retested in the future. Additionally, both GPT-3.5 and GPT-4 struggled with recognizing structural diagrams, limiting their performance in areas such as pharmaceutical chemistry and pharmacognosy. These limitations, consistent with previous research, highlight the need for cautious application of GPT models in fields that require visual recognition [11,48,49]. Additionally, the models showed poorer performance in subjects with less available training data and specific medical knowledge such as pharmacy law and traditional medicine, indicating potential biases in the models' training. We suggest that future efforts in model development should focus on incorporating more diverse and comprehensive data to reduce such biases.

Conclusions

This study demonstrates that GPT-4 outperforms GPT-3.5 in the Taiwan NPLE, particularly in pharmacy expertise, calculation ability, and situational case studies, with a notable advantage in basic subjects. It is recommended that GPT-4 be applied in clinical pharmacy practice (ie, patient education, drug consultation) and pharmacy education, particularly to support self-directed learning. However, given its limitations, caution is advised when integrating GPT-4 into clinical settings and educational programs. Future research should focus on refining prompts, improving model stability, integrating medical databases, and enhancing comprehensive questions to evaluate student competence more effectively while minimizing the chance of guessing correct answers.

Acknowledgments

This work was supported by Taipei Veterans General Hospital Hsinchu Branch (2024-VHCT-P-0008) and the authors would like to thank Wallace Academic Editing (<https://www.editing.tw/>) for English language editing.

Conflicts of Interest

None declared.

References

1. ChatGPT: optimizing language models for dialogue. OpenAI. URL: <https://chatgpt.r4wand.eu.org/> [Accessed 2023-03-03]
2. Research index. OpenAI. URL: <https://openai.com/research/gpt-4> [Accessed 2023-08-03]
3. OpenAI. GPT-4 technical report. arXiv. Preprint posted online on Mar 4, 2024. [doi: [10.48550/arXiv.2303.08774](https://doi.org/10.48550/arXiv.2303.08774)]
4. Kosinski M. Evaluating large language models in theory of mind tasks. Proc Natl Acad Sci U S A. Nov 5, 2024;121(45):e2405460121. [doi: [10.1073/pnas.2405460121](https://doi.org/10.1073/pnas.2405460121)] [Medline: [39471222](https://pubmed.ncbi.nlm.nih.gov/39471222/)]
5. Wang YM, Chen TJ. ChatGPT surges ahead: GPT-4 has arrived in the arena of medical research. J Chin Med Assoc. Sep 1, 2023;86(9):784-785. [doi: [10.1097/JCMA.0000000000000955](https://doi.org/10.1097/JCMA.0000000000000955)] [Medline: [37406215](https://pubmed.ncbi.nlm.nih.gov/37406215/)]
6. Gilson A, Safranek CW, Huang T, 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. Feb 8, 2023;9:e45312. [doi: [10.2196/45312](https://doi.org/10.2196/45312)] [Medline: [36753318](https://pubmed.ncbi.nlm.nih.gov/36753318/)]
7. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Dig Health. Feb 2023;2(2):e0000198. [doi: [10.1371/journal.pdig.0000198](https://doi.org/10.1371/journal.pdig.0000198)] [Medline: [36812645](https://pubmed.ncbi.nlm.nih.gov/36812645/)]
8. 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. [doi: [10.3352/jeehp.2023.20.1](https://doi.org/10.3352/jeehp.2023.20.1)] [Medline: [36627845](https://pubmed.ncbi.nlm.nih.gov/36627845/)]
9. 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. Apr 2023;185:109732. [doi: [10.1016/j.resuscitation.2023.109732](https://doi.org/10.1016/j.resuscitation.2023.109732)] [Medline: [36775020](https://pubmed.ncbi.nlm.nih.gov/36775020/)]
10. Weng TL, Wang YM, Chang S, Chen TJ, Hwang SJ. ChatGPT failed Taiwan's Family Medicine Board Exam. J Chin Med Assoc. Aug 1, 2023;86(8):762-766. [doi: [10.1097/JCMA.0000000000000946](https://doi.org/10.1097/JCMA.0000000000000946)] [Medline: [37294147](https://pubmed.ncbi.nlm.nih.gov/37294147/)]
11. 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. Dec 2023;3(4):100324. [doi: [10.1016/j.xops.2023.100324](https://doi.org/10.1016/j.xops.2023.100324)] [Medline: [37334036](https://pubmed.ncbi.nlm.nih.gov/37334036/)]
12. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. arXiv. Preprint posted online on Apr 12, 2023. [doi: [10.48550/arXiv.2303.13375](https://doi.org/10.48550/arXiv.2303.13375)]
13. Rosoł M, Gašior JS, Łaba J, Korzeniewski K, Młyńczak M. Evaluation of the performance of GPT-3.5 and GPT-4 on the Polish Medical Final Examination. Sci Rep. Nov 22, 2023;13(1):20512. [doi: [10.1038/s41598-023-46995-z](https://doi.org/10.1038/s41598-023-46995-z)] [Medline: [37993519](https://pubmed.ncbi.nlm.nih.gov/37993519/)]

14. Kaneda Y, Takahashi R, Kaneda U, et al. Assessing the performance of GPT-3.5 and GPT-4 on the 2023 Japanese Nursing Examination. *Cureus*. Aug 2023;15(8):e42924. [doi: [10.7759/cureus.42924](https://doi.org/10.7759/cureus.42924)] [Medline: [37667724](https://pubmed.ncbi.nlm.nih.gov/37667724/)]
15. Bhayana R, Bleakney RR, Krishna S. GPT-4 in radiology: improvements in advanced reasoning. *Radiology*. Jun 2023;307(5):e230987. [doi: [10.1148/radiol.230987](https://doi.org/10.1148/radiol.230987)] [Medline: [37191491](https://pubmed.ncbi.nlm.nih.gov/37191491/)]
16. Oh N, Choi GS, Lee WY. ChatGPT goes to the operating room: evaluating GPT-4 performance and its potential in surgical education and training in the era of large language models. *Ann Surg Treat Res*. May 2023;104(5):269-273. [doi: [10.4174/ast.2023.104.5.269](https://doi.org/10.4174/ast.2023.104.5.269)] [Medline: [37179699](https://pubmed.ncbi.nlm.nih.gov/37179699/)]
17. Sabry Abdel-Messih M, Kamel Boulos MN. ChatGPT in clinical toxicology. *JMIR Med Educ*. Mar 8, 2023;9:e46876. [doi: [10.2196/46876](https://doi.org/10.2196/46876)] [Medline: [36867743](https://pubmed.ncbi.nlm.nih.gov/36867743/)]
18. Nisar S, Aslam MS. Is ChatGPT a good tool for T&CM students in studying pharmacology? SSRN. Preprint posted online on Jan 17, 2023. [doi: [10.2139/ssrn.4324310](https://doi.org/10.2139/ssrn.4324310)]
19. Wang YM, Shen HW, Chen TJ. Performance of ChatGPT on the Pharmacist Licensing Examination in Taiwan. *J Chin Med Assoc*. Jul 1, 2023;86(7):653-658. [doi: [10.1097/JCMA.0000000000000942](https://doi.org/10.1097/JCMA.0000000000000942)] [Medline: [37227901](https://pubmed.ncbi.nlm.nih.gov/37227901/)]
20. He N, Yan Y, Wu Z, et al. Chat GPT-4 significantly surpasses GPT-3.5 in drug information queries. *J Telemed Telecare*. Jun 22, 2023;1357633X231181922. [doi: [10.1177/1357633X231181922](https://doi.org/10.1177/1357633X231181922)] [Medline: [37350055](https://pubmed.ncbi.nlm.nih.gov/37350055/)]
21. Li D, Yu J, Hu B, Xu Z, Zhang M. ExplainCPE: A free-text explanation benchmark of Chinese Pharmacist Examination. arXiv. Preprint posted online on Oct 26, 2023. [doi: [10.48550/arXiv.2305.12945](https://doi.org/10.48550/arXiv.2305.12945)]
22. Vert JP. How will generative AI disrupt data science in drug discovery? *Nat Biotechnol*. Jun 2023;41(6):750-751. [doi: [10.1038/s41587-023-01789-6](https://doi.org/10.1038/s41587-023-01789-6)] [Medline: [37156917](https://pubmed.ncbi.nlm.nih.gov/37156917/)]
23. Blanco-González A, Cabezón A, Seco-González A, et al. The role of AI in drug discovery: challenges, opportunities, and strategies. *Pharmaceuticals (Basel)*. Jun 18, 2023;16(6):891. [doi: [10.3390/ph16060891](https://doi.org/10.3390/ph16060891)] [Medline: [37375838](https://pubmed.ncbi.nlm.nih.gov/37375838/)]
24. Savage N. Drug discovery companies are customizing ChatGPT: here's how. *Nat Biotechnol*. May 2023;41(5):585-586. [doi: [10.1038/s41587-023-01788-7](https://doi.org/10.1038/s41587-023-01788-7)] [Medline: [37095351](https://pubmed.ncbi.nlm.nih.gov/37095351/)]
25. Wang H, Ding YJ, Luo Y. Future of ChatGPT in pharmacovigilance. *Drug Saf*. Aug 2023;46(8):711-713. [doi: [10.1007/s40264-023-01315-2](https://doi.org/10.1007/s40264-023-01315-2)] [Medline: [37306853](https://pubmed.ncbi.nlm.nih.gov/37306853/)]
26. Carpenter KA, Altman RB. Using GPT-3 to build a lexicon of drugs of abuse synonyms for social media pharmacovigilance. *Biomolecules*. Feb 18, 2023;13(2):387. [doi: [10.3390/biom13020387](https://doi.org/10.3390/biom13020387)] [Medline: [36830756](https://pubmed.ncbi.nlm.nih.gov/36830756/)]
27. Cloesmeijer ME, Janssen A, Koopman SF, Cnossen MH, Mathôt RAA, consortium S. ChatGPT in pharmacometrics? Potential opportunities and limitations. *Br J Clin Pharmacol*. Jan 2024;90(1):360-365. [doi: [10.1111/bcp.15895](https://doi.org/10.1111/bcp.15895)] [Medline: [37621112](https://pubmed.ncbi.nlm.nih.gov/37621112/)]
28. 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*. Mar 29, 2023;3(1):e103. [doi: [10.52225/narra.v3i1.103](https://doi.org/10.52225/narra.v3i1.103)]
29. Zhu Y, Han D, Chen S, Zeng F, Wang C. How can ChatGPT benefit pharmacy: a case report on review writing. Preprints. Preprint posted online on Feb 20, 2023. [doi: [10.20944/preprints202302.0324.v1](https://doi.org/10.20944/preprints202302.0324.v1)]
30. Hsu HY, Hsu KC, Hou SY, Wu CL, Hsieh YW, Cheng YD. Examining real-world medication consultations and drug-herb interactions: ChatGPT performance evaluation. *JMIR Med Educ*. Aug 21, 2023;9:e48433. [doi: [10.2196/48433](https://doi.org/10.2196/48433)] [Medline: [37561097](https://pubmed.ncbi.nlm.nih.gov/37561097/)]
31. 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*. Sep 4, 2023;219(5):237. [doi: [10.5694/mja2.52061](https://doi.org/10.5694/mja2.52061)] [Medline: [37528548](https://pubmed.ncbi.nlm.nih.gov/37528548/)]
32. Kılıç ME. AI in medical education: A comparative analysis of GPT-4 and GPT-3.5 on turkish medical specialization exam performance. medRxiv. Preprint posted online on Jul 12, 2023. [doi: [10.1101/2023.07.12.23292564](https://doi.org/10.1101/2023.07.12.23292564)]
33. 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*. Jun 29, 2023;9:e48002. [doi: [10.2196/48002](https://doi.org/10.2196/48002)] [Medline: [37384388](https://pubmed.ncbi.nlm.nih.gov/37384388/)]
34. Guerra GA, Hofmann H, Sobhani S, et al. GPT-4 artificial intelligence model outperforms ChatGPT, medical students, and neurosurgery residents on neurosurgery written board-like questions. *World Neurosurg*. Nov 2023;179:e160-e165. [doi: [10.1016/j.wneu.2023.08.042](https://doi.org/10.1016/j.wneu.2023.08.042)] [Medline: [37597659](https://pubmed.ncbi.nlm.nih.gov/37597659/)]
35. Lewandowski M, Łukowicz P, Świetlik D, Barańska-Rybak W. ChatGPT-3.5 and ChatGPT-4 dermatological knowledge level based on the Specialty Certificate Examination in dermatology. *Clin Exp Dermatol*. Jun 25, 2024;49(7):686-691. [doi: [10.1093/ced/llad255](https://doi.org/10.1093/ced/llad255)] [Medline: [37540015](https://pubmed.ncbi.nlm.nih.gov/37540015/)]
36. Khorshidi H, Mohammadi A, Yousem DM, et al. Application of ChatGPT in multilingual medical education: How does ChatGPT fare in 2023's Iranian Residency Entrance Examination. *Inform Med Unlocked*. 2023;41:101314. [doi: [10.1016/j.imu.2023.101314](https://doi.org/10.1016/j.imu.2023.101314)]

37. Huang X, Estau D, Liu X, Yu Y, Qin J, Li Z. Evaluating the performance of ChatGPT in clinical pharmacy: a comparative study of ChatGPT and clinical pharmacists. *Br J Clin Pharmacol*. Jan 2024;90(1):232-238. [doi: [10.1111/bcp.15896](https://doi.org/10.1111/bcp.15896)] [Medline: [37626010](https://pubmed.ncbi.nlm.nih.gov/37626010/)]
38. Jairoun AA, Al-Hemyari SS, Shahwan M, Humaid Alnuaimi GR, Zyoud SH, Jairoun M. ChatGPT: threat or boon to the future of pharmacy practice? *Res Social Adm Pharm*. Jul 2023;19(7):975-976. [doi: [10.1016/j.sapharm.2023.03.012](https://doi.org/10.1016/j.sapharm.2023.03.012)] [Medline: [37061346](https://pubmed.ncbi.nlm.nih.gov/37061346/)]
39. Juhi A, Pipil N, Santra S, Mondal S, Behera JK, Mondal H. The capability of ChatGPT in predicting and explaining common drug-drug interactions. *Cureus*. Mar 2023;15(3):e36272. [doi: [10.7759/cureus.36272](https://doi.org/10.7759/cureus.36272)] [Medline: [37073184](https://pubmed.ncbi.nlm.nih.gov/37073184/)]
40. Davies NM. Adapting artificial intelligence into the evolution of pharmaceutical sciences and publishing: technological darwinism. *J Pharm Pharm Sci*. 2023;26:11349. [doi: [10.3389/jpps.2023.11349](https://doi.org/10.3389/jpps.2023.11349)] [Medline: [37034476](https://pubmed.ncbi.nlm.nih.gov/37034476/)]
41. Kleebayoon A, Wiwanitkit V. Performance and risks of ChatGPT used in drug information: comment. *Eur J Hosp Pharm*. Dec 27, 2023;31(1):85-86. [doi: [10.1136/ejpharm-2023-003864](https://doi.org/10.1136/ejpharm-2023-003864)] [Medline: [37339863](https://pubmed.ncbi.nlm.nih.gov/37339863/)]
42. Mohammed M, Kumar N, Zawiah M, et al. Psychometric properties and assessment of knowledge, attitude, and practice towards ChatGPT in pharmacy practice and education: a study protocol. *J Racial Ethn Health Disparities*. Aug 2024;11(4):2284-2293. [doi: [10.1007/s40615-023-01696-1](https://doi.org/10.1007/s40615-023-01696-1)] [Medline: [37428357](https://pubmed.ncbi.nlm.nih.gov/37428357/)]
43. Abu-Farha R, Fino L, Al-Ashwal FY, et al. Evaluation of community pharmacists' perceptions and willingness to integrate ChatGPT into their pharmacy practice: a study from Jordan. *J Am Pharm Assoc*. Nov 2023;63(6):1761-1767. [doi: [10.1016/j.japh.2023.08.020](https://doi.org/10.1016/j.japh.2023.08.020)]
44. Choi W. Assessment of the capacity of chatgpt as a self-learning tool in medical pharmacology: a study using mcqs. *BMC Med Educ*. Nov 13, 2023;23(1):864. [doi: [10.1186/s12909-023-04832-x](https://doi.org/10.1186/s12909-023-04832-x)] [Medline: [37957666](https://pubmed.ncbi.nlm.nih.gov/37957666/)]
45. Snoswell CL, Falconer N, Snoswell AJ. Pharmacist vs machine: pharmacy services in the age of large language models. *Res Social Adm Pharm*. Jun 2023;19(6):843-844. [doi: [10.1016/j.sapharm.2023.03.006](https://doi.org/10.1016/j.sapharm.2023.03.006)] [Medline: [36907776](https://pubmed.ncbi.nlm.nih.gov/36907776/)]
46. Meskó B. Prompt engineering as an important emerging skill for medical professionals: tutorial. *J Med Internet Res*. Oct 4, 2023;25:e50638. [doi: [10.2196/50638](https://doi.org/10.2196/50638)] [Medline: [37792434](https://pubmed.ncbi.nlm.nih.gov/37792434/)]
47. Krumborg JR, Mikkelsen N, Damkier P, et al. ChatGPT: first glance from a perspective of clinical pharmacology. *Basic Clin Pharma Tox*. Jul 2023;133(1):3-5. [doi: [10.1111/bcpt.13879](https://doi.org/10.1111/bcpt.13879)]
48. Fergus S, Botha M, Ostovar M. Evaluating academic answers generated using ChatGPT. *J Chem Educ*. Apr 11, 2023;100(4):1672-1675. [doi: [10.1021/acs.jchemed.3c00087](https://doi.org/10.1021/acs.jchemed.3c00087)]
49. Massey PA, Montgomery C, Zhang AS. Comparison of ChatGPT-3.5, ChatGPT-4, and orthopaedic resident performance on orthopaedic assessment examinations. *J Am Acad Orthop Surg*. Dec 1, 2023;31(23):1173-1179. [doi: [10.5435/JAAOS-D-23-00396](https://doi.org/10.5435/JAAOS-D-23-00396)] [Medline: [37671415](https://pubmed.ncbi.nlm.nih.gov/37671415/)]

Abbreviations

GPT-3.5: ChatGPT-3.5

GPT-4: ChatGPT-4

NPLE: National Pharmacist Licensing Examination

USMLE: United States Medical Licensing Examination

Edited by Blake Lesselroth; peer-reviewed by Ghulam Farid, Suodi Zhai, Wang Yu Jen; submitted 28.01.2024; final revised version received 26.09.2024; accepted 17.12.2024; published 17.01.2025

Please cite as:

Wang YM, Shen HW, Chen TJ, Chiang SC, Lin TG

Performance of ChatGPT-3.5 and ChatGPT-4 in the Taiwan National Pharmacist Licensing Examination: Comparative Evaluation Study

JMIR Med Educ 2025;11:e56850

URL: <https://mededu.jmir.org/2025/1/e56850>

doi: [10.2196/56850](https://doi.org/10.2196/56850)

© Ying-Mei Wang, Hung-Wei Shen, Tzeng-Ji Chen, Shu-Chiung Chiang, Ting-Guan Lin. Originally published in JMIR Medical Education (<https://mededu.jmir.org>), 17.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on <https://mededu.jmir.org/>, as well as this copyright and license information must be included.