



# Comparative analysis of GPT-3.5 and GPT-4.0 in Taiwan's medical technologist certification: A study in artificial intelligence advancements

Wan-Hua Yang<sup>a,b,c,d</sup>, Yun-Hsiang Chan<sup>a</sup>, Cheng-Pin Huang<sup>a</sup>, Tzeng-Ji Chen<sup>e,f,g,\*</sup>

<sup>a</sup>Department of Pathology and Laboratory, Taipei Veterans General Hospital Hsinchu Branch, Hsinchu, Taiwan, ROC; <sup>b</sup>Department of Industrial Engineering and Management, Yuan Ze University, Taoyuan, Taiwan, ROC; <sup>c</sup>Department of Medical Laboratory Science and Biotechnology, Yuanpei University of Medical Technology, Hsinchu, Taiwan, ROC; <sup>d</sup>Department of Medical and Healthcare Business, Hsin Sheng Junior College of Medical Care and Management, Taoyuan, Taiwan, ROC; <sup>e</sup>Department of Family Medicine, Taipei Veterans General Hospital Hsinchu Branch, Hsinchu, Taiwan, ROC; <sup>f</sup>Department of Family Medicine, Taipei Veterans General Hospital, Taipei, Taiwan, ROC; <sup>g</sup>Department of Post-Baccalaureate Medicine, National Chung Hsing University, Taichung, Taiwan, ROC

## Abstract

**Background:** This study examines the comparative effectiveness of GPT-3.5 and GPT-4.0, in the certification of medical technologists (MT) in Taiwan, exploring their adeptness in processing complex medical language and their contributory role in the educational and communicative aspects of professional healthcare training.

**Methods:** This study used GPT-3.5 and GPT-4.0 to test the medical laboratory technician professional college entrance examination questions. The questions in different fields, including six subjects, such as Clinical Physiology and Pathology, Hematology, and Blood Bank, among others were answered one-on-one using two generative pretrained transformer (GPT) versions, simulating the situations during exam preparation.

**Results:** A total of 480 questions were analyzed and the results showed that both versions of the GPT met the certification standards. Version 4.0 was better than version 3.5 for all subjects, particularly in Clinical Biochemistry (score = 96.25) and Microbiology (score = 91.25). Outstanding performance compared to version 3.5, which had an average score of 65.42 and a maximum score of 77.5. Overall, version 4.0, which was significantly better than version 3.5 in both median and average scores, reflects a significant improvement in professional knowledge processing capabilities.

**Conclusion:** The GPT can provide valuable support for both the upstream and downstream processes of MT certification. Future research can further explore the application of GPT in different educational and certification contexts and improve the passing rate of medical personnel in the certification process. This study provides useful information for exploring the potential applications of GPT in certifying medical examiners. Furthermore, it provides new directions for future research in medical education.

**Keywords:** Artificial intelligence; Blood banks; Certification; Medical education; Taiwan

## 1. INTRODUCTION

With the rapid development of artificial intelligence (AI) technology and revolutionary progress in several fields, its potential in medical education and professional training has gradually received attention, particularly in the certification process of medical technologists (MT), the application of AI may open new paths for quality education, strengthening clinical skills,

and supporting professional development. This study evaluates the performance of the GPT-3.5 and GPT-4.0 versions launched by OpenAI in Taiwan's MT certification process. Furthermore, it explores how these AI models can assist medical professionals in mastering critical knowledge and skills. Through an in-depth analysis of GPT's ability to provide medical knowledge, case analysis, and technical guidance, this study aims to reveal the practical application and effect of AI in contemporary medical education, which is not only of great significance for promoting innovation in medical education methods but also provides an empirical basis for the further integration of AI technology in Taiwan and even the global healthcare system.

A generative pretrained transformer (GPT) is a large language model (LLM) powered by AI and developed by OpenAI. The latest versions of this model are GPT-3.5 and GPT-4.0, launched in March 2022 and April 2023, respectively. Both models are available and optimized for natural conversation in GPT web applications; however, GPT-4.0 is a paid monthly subscription.<sup>1</sup> Recently, several studies have highlighted that natural language processing models such as GPT-3.5 and GPT-4.0 have shown potential application capabilities in medical personnel license examinations. These models can automatically generate explanatory answers, provide detailed explanations, assist in knowledge

\* Address correspondence. Dr. Tzeng-Ji Chen, Department of Family Medicine, Taipei Veterans General Hospital Hsinchu Branch, 81, Section 1, Zhongfeng Road, Hsinchu 310, Taiwan, ROC. E-mail address: tjchen@vhct.gov.tw (T.-J. Chen).

Conflicts of interest: Dr. Tzeng-Ji Chen, an editorial board member at Journal of the Chinese Medical Association, had no role 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. (2024) 87: 525-530.

Received December 26, 2023; accepted March 3, 2024.

doi: 10.1097/JCMA.0000000000001092

Copyright © 2024, the Chinese Medical Association. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

retrieval, and be helpful for exam preparations. These assistive tools promise to improve candidates' learning and preparation processes. However, care must be taken to ensure the fairness and legitimacy of the exam. Taloni et al<sup>2</sup> compared GPT-4.0, GPT-3.5, and human performance on 1023 multiple-choice questions from the American Academy of Ophthalmology Basic and Clinical Sciences Course Self-Assessment Program. GPT-4.0 scored the highest at 82.4%, surpassing humans at 75.7%, and GPT-3.5 at 65.9%, indicating a significant difference in accuracy ( $p < 0.001$ ).<sup>2</sup> Takagi et al<sup>3</sup> evaluated GPT-3.5 and GPT-4.0 on the Japanese Medical Licensing Examination (JMLE), analyzing 254 questions to assess non-English clinical reasoning and medical knowledge. GPT-4 outperformed GPT-3.5 in accuracy, particularly in general, clinical, and complex questions, and met the JMLE's passing standards, demonstrating its reliability in non-English contexts.<sup>3</sup> Schubert et al<sup>4</sup> assessed LLMs, including GPT-4.0 and GPT-3.5, on the Neurology Board Examination. GPT-4.0 surpassed 3.5 and the average human score, highlighting its potential in clinical neurology and healthcare.<sup>4</sup> Oztermeli and Oztermeli<sup>5</sup> reviewed GPT-3.5's performance in medical specialty exams (MSEs) over 5 years, analyzing 1177 public MSE questions. The results varied, with success rates ranging from 54.3% to 70.9% and rankings ranging from 1787th out of 2214 to 4428th out of 21 476, indicating GPT-3.5's competence and limitations compared with field experts.<sup>5</sup>

Drawing from the insights gathered, we can understand the potential of GPT in assisting medical education. However, their knowledge base may only partially comply with medical standards and the latest research in a specific region (such as Taiwan). As there are no relevant reports on the performance of GPTs in MT Certification, and different GPT models may experience performance fluctuations when dealing with complex medical cases, updating and evaluating new versions is critical to the currency and appropriateness of educational resources. Therefore, this study aims to assess the application effect of GPT versions 3.5 and 4.0 in the certification of MTs in Taiwan and explore the potential of these advanced AI models in promoting educational innovation, improving the learning and training of medical professionals, and strengthening clinical decision support. Our results help us understand the actual application of AI in medical education and professional development in Taiwan, and provide an empirical basis for future technology integration and policy formulation.

## 2. METHODS

### 2.1. Background

With the rapid development of AI technology, its potential applications in medical education have gradually received attention. Particularly, it has significant advantages in assisting with learning and preparation for professional examinations. This study evaluates the performance of the GPT-3.5 and GPT-4.0 versions in answering MT examination questions and compares them with the average passing rate of human candidates.

### 2.2. Data sources

The data source used in this study is the Ministry of Examination, ROC (Taiwan) 2023 Second Senior Professional and Technical Examinations for Medical Technologists' question bank, which includes Clinical Physiology and Pathology, Hematology and Blood Bank, Molecular Biology, and Clinical Microscopy (including Parasitology), Microbiology and Clinical Microbiology (including Bacteria and Mold), and Biochemistry and Clinical Biochemistry, Immunology and Virology. Each subject had 80 questions, including various types of questions such as single-choice questions, diagram questions, and case reports,

resulting in a total of 480 questions and 800 points. The mean passing score was 60 points on average.

### 2.3. Study design

This study used an observational retrospective design method to explore the application of the two versions of the GPT in the medical field of national examinations. Each GPT version needs to complete the answers to these questions independently and is not affected by external assistance or additional information during the answering process. Evaluate the number of correct answers and the corresponding scores.

In addition, to evaluate the applicability of the GPT in this field more comprehensively, we compared these results with the published average passing rate of human candidates on the same test questions. This not only demonstrates the ability of GPT to apply medical knowledge but also provides an essential reference for the potential application of AI technology in medical education and practice in the future.

### 2.4. Statistical analysis

This study used Microsoft Excel 2016 to conduct a descriptive statistical analysis to explore the performance of the GPT between versions 3.5 and 4.0. We calculated the number of correct answers for each subject and the average score and used Python Language Reference, version 3.8.10 (Guido van Rossum, the Netherlands, 1989) for the statistical differences between versions. The  $p < 0.05$  are considered statistically significant. This analysis provides vital insights into the relative advantages and disadvantages of the different versions of medical-related questions.

In addition, to comprehensively evaluate the practical application of AI in the medical field and its future potential, this study also compared the performance of GPT with the published average passing rate of human candidates in relevant subjects, which helps to understand the role of GPT in the relative status of applications in the medical field and provides guidance for future research directions.

## 3. RESULTS

This study compared the performance of the GPT versions 3.5 and 4.0, in Taiwan's national examination and obtained preliminary results by answering 480 questions in six subjects. Table 1 shows that in all subjects except immunology and virology, GPT-4.0 had a statistically significant increase in correct answers compared to GPT-3.5. The  $p$  value for Immunology and Virology is slightly higher than the commonly used significance level (0.05), which implies that the difference in performance between GPT-3.5 and GPT-4.0 in this subject may not be statistically significant. Table 2 indicates that the average score of GPT-4.0 is 88.13 points (SD = 6.11), while the average score of GPT-3.5 is 65.42 points (SD = 8.68), both of which meet the passing standards. The average score of GPT-4.0 is higher, but its standard deviation is more minor, indicating that its performance is more consistent and stable. The  $p < 0.001$ , indicating a statistically significant difference between the GPT-3.5 and GPT-4.0.

Fig. 1 depicts the number of correct answers by GPT-4.0 and GPT-3.5 in each subject. Each bar represents the performance of the two versions on the corresponding subject, and the dotted line represents the number of questions passed. Fig 2 illustrates the score distribution of GPT-3.5 and GPT-4.0 in the Taiwan Medical Technologist National Examination (TMTNE). From the box plot it can be observed that the lowest score of GPT-3.5 is 53.75, the highest score is 77.5, the median (middle of the box horizontal line) is 71.88, and the average score ("X" mark) is

**Table 1**  
Performance of GPT versions 3.5 and 4.0 in various subject examinations

Subjects	Clinical Physiology and Pathology		Hematology and Blood Bank		Molecular Biology and Clinical Microscopy (including Parasitology)		Microbiology and Clinical Microbiology (including Bacteria and Mold)		Biochemistry and Clinical Biochemistry		Immunology and Virology	
	3.5	4.0	3.5	4.0	3.5	4.0	3.5	4.0	3.5	4.0	3.5	4.0
GPT version	3.5	4.0	3.5	4.0	3.5	4.0	3.5	4.0	3.5	4.0	3.5	4.0
Number of questions	80		80		80		80		80		80	
No. of correct answers	52	71	43	67	46	63	56	73	55	77	62	72
Score	65.0	88.75	53.75	83.75	57.5	78.75	70.0	91.25	68.75	96.25	77.5	90.0
Odds ratio <sup>a</sup> (95% CI)	0.23 (0.10-0.54)		0.23 (0.10-0.47)		0.37 (0.18-0.73)		0.22 (0.09-0.56)		0.09 (0.03-0.30)		0.38 (0.16-0.94)	
<i>p</i>	<0.001		<0.001		0.006		0.001		<0.001		0.052	

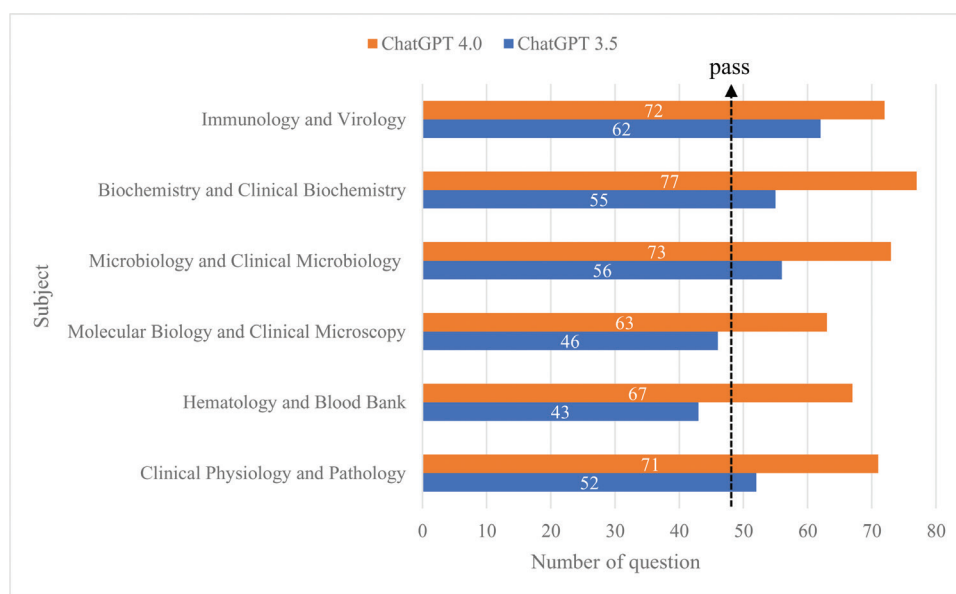
GPT = generative pretrained transformer.

<sup>a</sup>The odds ratio of GPT-3.5 and 4.0 to get the correct answers, according to each subject.

**Table 2**  
Statistical analysis of GPT versions' performance scores

GPT version	Total score	Average score (SD)	95% CI	<i>p</i>
GPT-3.5	392.5	65.42 (8.68)	56.31-74.53	<0.001
GPT-4.0	528.8	88.13 (6.11)	81.71-94.53	

GPT = generative pre-trained transformer.



**Fig. 1** Number of correct answers for GPT-3.5 and 4.0.

65.42, which implies that the average score is slightly lower than the median. In comparison, the GPT-4.0 has a minimum score of 78.75, a maximum score of 96.25, and a median score of 71.88. The number of digits is 89.38, and the average score is 88.13, which is slightly lower than the median. Overall, the GPT version 4.0 performed better than version 3.5, with higher median and average scores and upper bounds. This suggests that version 4.0 performs better on specific tasks related to the MT exam. Figs. 3 and 4 illustrate examples of GPT-4.0 and GPT-3.5 reactions to a question from Taiwan's national examination for MT.

#### 4. DISCUSSION

Weng et al<sup>6</sup> used the 2022 Taiwan Family Medicine Board examination questions. The accuracy rate of ChatGPT was

only 41.6%, which failed the exam. This study involved various questions including reverse, multiple-choice, and general medical knowledge. This may increase the difficulty and complexity of the exam.<sup>6</sup> Compared with the results of Weng et al,<sup>6</sup> the performance of this study using GPT-3.5 and GPT-4.0 in the TMTNE was significantly improved. Conversely, the examination in this study might be more focused on specific types of questions, which may help GPT-3.5 and GPT-4.0 perform better. The results suggest that GPT's performance can vary significantly across different types of tests.

Combining insights from Brin et al,<sup>7</sup> Gilson et al,<sup>8</sup> and our study, we observed advancements in the capabilities of GPT models in medical education. Brin et al<sup>7</sup> demonstrated GPT-4's high proficiency in the USMLE, particularly in communication and ethics, with an accuracy rate of 90%. Gilson et al<sup>8</sup>

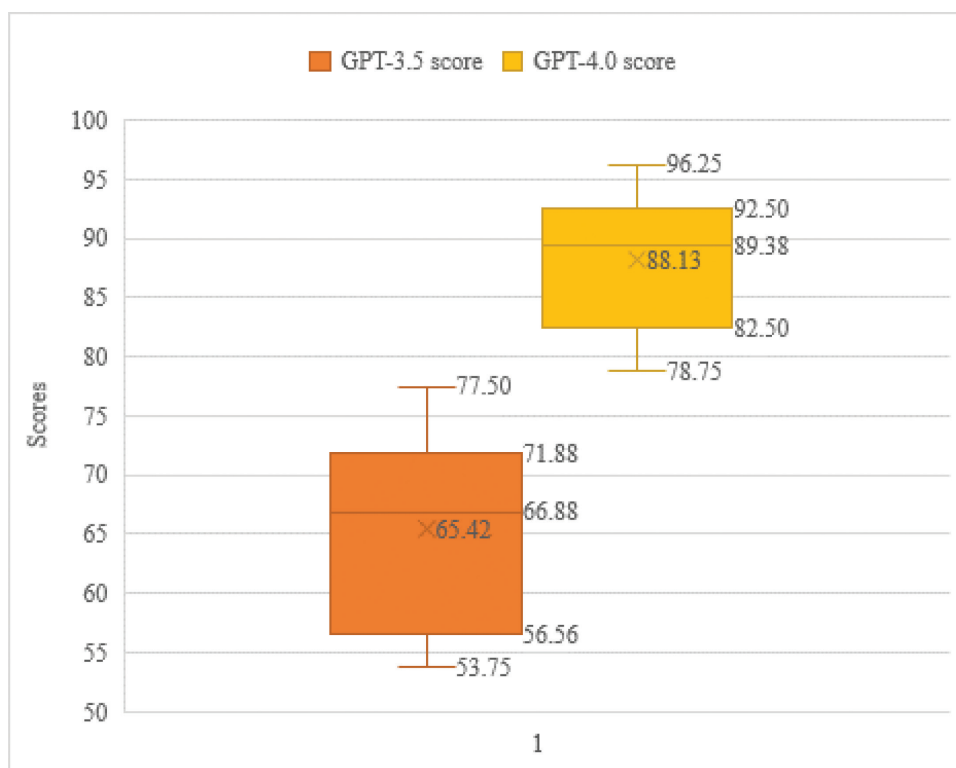


Fig. 2 Box plot for GPT version of score. GPT = generative pretrained transformer.

**You**

1. 感染parvovirus B19病毒的溶血性貧血病人會引發aplastic crisis，關鍵是紅血球母細胞表面有表達下列何種抗原，致使病毒可以進入紅血球母細胞？

A. P抗原  
 B. Rh抗原  
 C. H抗原  
 D. Lewis抗原

**ChatGPT**

感染parvovirus B19的溶血性貧血病患會引起aplastic crisis的原因是，該病毒能夠進入紅血球母細胞。這是因為紅血球母細胞表面表達了能夠與病毒結合的特定抗原。在您提供的選項中，正確答案是：

A. P抗原

Parvovirus B19病毒特別針對紅血球母細胞表面的P抗原，這允許病毒綁定並進入這些細胞。這種病毒與細胞的結合是造成紅血球生成受阻，進而導致aplastic crisis的關鍵機制。

Fig. 3 Example of GPT-4.0 reactions with a question from Taiwan's national examination for MT. MT = medical technologist.

corroborated the effectiveness of the GPT by comparing its performance with that of a third-year medical student. Our study further enriches this discussion by examining GPT versions 3.5 and 4.0 in Taiwan's national examination for MT. By analyzing 480 questions across six subjects, we observed a significant

enhancement in GPT-4.0's performance over GPT-3.5, in all areas except immunology and virology. Notably, GPT-4.0 achieved an average score of 88.13, with a minor standard deviation, indicating a higher performance level and greater consistency compared to GPT-3.5's average score of 65.42. This difference is



**You**

1. 感染parvovirus B19病毒的溶血性貧血病人會引發aplastic crisis，關鍵是紅血球母細胞表面有表達下列何種抗原，致使病毒可以進入紅血球母細胞？

A. P抗原  
B. Rh抗原  
C. H抗原  
D. Lewis抗原

**ChatGPT**

感染parvovirus B19病毒的溶血性貧血病人會引發aplastic crisis的原因是因為病毒可以感染紅血球母細胞。病毒可以進入紅血球母細胞是因為紅血球表面具有P抗原。

所以正確的答案是：

A. P抗原

**Fig. 4** Example of GPT-3.5 reactions with a question from Taiwan's national examination for MT. MT = medical technologist.

statistically significant ( $p < 0.001$ ), highlighting GPT-4.0's superiority. These studies indicate a promising trajectory for GPT models in medical education, particularly for exams. GPT-4.0, with its consistent and more accurate performance even in non-English contexts, holds the potential for broader applications. However, ethical considerations and responsible integration of AI into educational assessments remain critical. This technology should complement traditional learning methods, adhere to educational standards, and ensure a balanced and practical approach to medical education.

This study indicates that GPT-4.0 significantly improves accuracy in most subjects and performs exceptionally well in professional subjects such as "Biochemistry and Clinical Biochemistry" and "Microbiology and Clinical Microbiology." This difference may stem from technological advancements between the GPT versions, particularly in understanding and dealing with complex problems. GPT-4.0 performed better than GPT-3.5 in this study and showed a significant improvement compared with the results of Weng et al.<sup>6</sup> This reflects the iterative progress of AI technology and emphasizes its potential for application to professional knowledge. However, it also highlights that there exist some challenges in handling complex and diverse problems, particularly in highly specialized fields such as medical examinations. Although the application of AI technology in education and professional examinations has broad prospects, its limitations must be carefully considered.

Our study underscores AI's potential in academic and professional certification exams, as evidenced by the GPT's performance compared to the human passing rate of only 28.61% in the TMTNE.<sup>9</sup> While GPT-4.0 showed impressive results, surpassing GPT-3.5, and human candidates in various subjects; it also revealed some limitations, such as failing to achieve total scores in some areas. These limitations highlight the need for AI to complement, rather than replace, traditional learning methods and human judgment. Our analysis demonstrates GPT-4.0's significant advancement in natural language processing and knowledge comprehension, particularly in professional fields. Its success as an educational aid suggests its valuable applications in medical education, enhancing learning experiences, and providing detailed explanations. However, clinical judgment and professional intuition remain crucial.

In the 2023 Medical Technology Expert Certification, GPT-4.0 demonstrated its capabilities in this field. This achievement is primarily owing to its extensive data-training set and advanced algorithm technology. This remarkable performance may promote the broader application of the GPT in the medical industry and medical education. Compared with previous versions such as GPT-3.5, GPT-4.0 has significantly improved its information processing and retention capabilities, however, this may also bring new challenges, such as over-reliance on information from past conversations, thereby introducing bias. To reduce this impact, the same question was submitted to two different versions of the GPT, asking each question only once without providing additional information. In addition, it reduces the memory effects by logging out and in. Research shows that GPT-4.0 has significantly improved compared to GPT-3.5 in providing different answers to the same questions, and its answers are almost consistent. Based on the results of multiple studies, developers must continue to strive in the face of challenges and problems in AI. With technological advances, further improvements and innovations in this area are expected.

In response to the possible bias caused by the GPT memory effect, we propose suggestions, including the regular use of specific test datasets for bias assessment and correction. These datasets should contain diverse cases, particularly those that pose challenges to the systems. Simultaneously, the necessity of continuously updating and evaluating AI systems is emphasized to keep up with technological evolution and adaptation to new situations. This can help developers gain a deeper understanding of the system's performance and reduce possible biases and misunderstandings.

## ACKNOWLEDGMENTS

This study was supported by a grant (2024-VHCT-RD-P002) from Taipei Veterans General Hospital Hsinchu Branch.

## REFERENCES

1. OpenAI. 2023. Available at <https://openai.com/>. Accessed April 7, 2024.
2. Taloni A, Borselli M, Scarsi V, Rossi C, Coco G, Scoria V, et al. Comparative performance of humans versus GPT-4.0 and GPT-3.5 in the self-assessment program of American Academy of Ophthalmology. *Sci Rep* 2023;13:18562.

3. Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese Medical Licensing Examination: comparison study. *JMIR Med Educ* 2023;9:e48002.
4. Schubert MC, Wick W, Venkataramani V. Performance of large language models on a neurology board-style examination. *JAMA Network Open* 2023;6:e2346721.
5. Oztermeli AD, Oztermeli A. ChatGPT performance in the medical specialty exam: an observational study. *Medicine (Baltimore)* 2023;102:e34673.
6. 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.
7. Brin D, Sorin V, Vaid A, Soroush A, Glicksberg BS, Charney AW, et al. Comparing ChatGPT and ChatGPT-4 performance in USMLE soft skill assessments. *Sci Rep* 2023;13:16492.
8. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. *Med Educ* 2023;9:e45312.
9. Ministry of Examination, R.O.C. 2023. Available at [https://wwwc.moex.gov.tw/main/ExamReport/wFrmExamStatistics.aspx?menu\\_id=158](https://wwwc.moex.gov.tw/main/ExamReport/wFrmExamStatistics.aspx?menu_id=158). Accessed April 7, 2024.