Original Research Article

DIGITAL HEALTH

DIGITAL HEALTH
Volume 10: 1–11
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20552076241237678
journals.sagepub.com/home/dhj
Sage

Exploring the proficiency of ChatGPT-4: An evaluation of its performance in the Taiwan advanced medical licensing examination

Shih-Yi Lin[1,2], Pak Ki Chan[3], Wu-Huei Hsu[1,4] and Chia-Hung Kao[1,3,5,6]

Abstract

Background: Taiwan is well-known for its quality healthcare system. The country's medical licensing exams offer a way to evaluate ChatGPT's medical proficiency.

Methods: We analyzed exam data from February 2022, July 2022, February 2023, and July 2033. Each exam included four papers with 80 single-choice questions, grouped as descriptive or picture-based. We used ChatGPT-4 for evaluation. Incorrect answers prompted a "chain of thought" approach. Accuracy rates were calculated as percentages.

Results: ChatGPT-4's accuracy in medical exams ranged from 63.75% to 93.75% (February 2022–July 2023). The highest accuracy (93.75%) was in February 2022's Medicine Exam (3). Subjects with the highest misanswered rates were ophthalmology (28.95%), breast surgery (27.27%), plastic surgery (26.67%), orthopedics (25.00%), and general surgery (24.59%). While using "chain of thought," the "Accuracy of (CoT) prompting" ranged from 0.00% to 88.89%, and the final overall accuracy rate ranged from 90% to 98%.

Conclusion: ChatGPT-4 succeeded in Taiwan's medical licensing exams. With the "chain of thought" prompt, it improved accuracy to over 90%.

Keywords

ChatGPT-4, Taiwan's medical licensing exams, chain of thought, healthcare system

Submission date: 22 August 2023; Acceptance date: 20 February 2024

Introduction

Taiwan, a country that has ranked first consecutively in the Numbeo's mid-year "Global Health Care Index" which a testament to its commitment to providing accessible and high-quality medical services to its citizens, has been famous for its health care system.[1] Building upon its robust healthcare system, Taiwan further secured the top rank in the Nikkei COVID-19 Recovery Index during the pandemic which reflects Taiwan's adeptness in effectively navigating the challenges of the COVID-19 crisis through strategic measures and swift responses.[2,3]

The presence of highly trained and professional medical care staff stands as a cornerstone of Taiwan's healthcare system.[4] Taiwan boasts a robust medical education and

[1]Graduate Institute of Clinical Medical Science, College of Medicine, China Medical University, Taichung, Taiwan
[2]Division of Nephrology and Kidney Institute, China Medical University Hospital, Taichung, Taiwan
[3]Artificial Intelligence Center, China Medical University Hospital, Taichung, Taiwan
[4]Department of Chest Medicine, China Medical University Hospital, Taichung, Taiwan
[5]Department of Nuclear Medicine and PET Center, China Medical University Hospital, Taichung, Taiwan
[6]Department of Bioinformatics and Medical Engineering, Asia University, Taichung, Taiwan

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

training infrastructure that cultivates doctors who are well-prepared to meet the diverse healthcare needs of the population,[5–10] as well as their active involvement in research and collaboration with international medical networks enhances the overall quality and reputation of Taiwan's healthcare system.[11–13]

Considering the critical role of physicians in Taiwan's healthcare system, the Taiwan medical licensing examinations, essential for validating the competence of Taiwan's graduate medical students, are crucial in ensuring the high standards of medical professionals.[14,15] These examinations are meticulously designed to ensure that aspiring physicians possess the necessary knowledge and skills to provide quality patient care.[15]

Although several professions across different countries have recently leveraged their respective national licensing exams to assess the capabilities of ChatGPT, our study uniquely focuses on using the Taiwan medical licensing examinations to evaluate its proficiency in the specialized field of medicine.[16–21] With the Taiwan medical licensing examinations serving as a crucial tool for assessing the competence of medical students, we aim to rigorously evaluate ChatGPT-4's medical knowledge and reasoning skills by administering the challenging Taiwan advanced medical licensing examinations from 2022 and 2023. Furthermore, we intend to explore ChatGPT-4's capacity for generating a coherent "chain of thoughts" in response to medical queries which ChatGPT-4 initially misanswered. Our goal is to explore ChatGPT's understanding of medical practices and challenges unique to Taiwan, benchmark its capabilities in medicine against various international standards, and identify areas where ChatGPT excels or requires improvement in medical knowledge and reasoning in Taiwan. This approach will also help understand its potential contribution to the medical field in Taiwan and globally.

Methods

Due to concerns that ChatGPT's database, which extends up to 2021, might have been exposed to questions from earlier iterations of the Taiwan medical licensing examinations, we specifically chose to evaluate ChatGPT using the 2022 and 2023 examinations. This approach ensures that the AI system encounters entirely new questions, providing a more accurate assessment of its current capabilities. Our evaluation encompassed the full spectrum of exam content, including sections on imaging, ethics, electrocardiograms, calculations, and descriptive questions. To maintain the integrity of this assessment, we deliberately excluded exam questions from 2021 and earlier years, as they may have been within the scope of ChatGPT's preexisting knowledge base.

We utilized the examinations from February 2022, July 2022, February 2023, and July 2033. Each of these examinations consisted of four test papers, each paper containing 80 single-choice questions, predominantly written in traditional Chinese and are categorized into four distinct test papers: (a) Medicine (3): This test paper encompasses questions from internal medicine, family medicine, and a variety of other related disciplines. Additionally, it integrates clinical examples pertinent to the aforementioned subjects and questions on medical ethics. (b) Medicine (4): This segment focuses on pediatrics, dermatology, neurology, and psychiatry among other related disciplines. It also includes relevant clinical scenarios corresponding to the covered subjects, as well as sections dedicated to medical ethics. (c) Medicine (5): This paper primarily targets surgical disciplines, covering general surgery, orthopedics, urology, and other associated specialties. Alongside, the paper integrates clinical examples relating to the subjects and poses questions related to medical ethics. (d) Medicine (6): The final test paper incorporates questions from anesthesiology, ophthalmology, otolaryngology, obstetrics and gynecology, and rehabilitation, among other related subjects. corresponding to the mentioned questions could be broadly classified into two types: descriptive and picture interpretation. All questions, whether descriptive or imaging-based were directly input for ChatGPT answer. From June 12, 2023, to August 13, 2023, ChatGPT 4 was deployed to complete these exams. Of special concern, in July 2023, ChatGPT was unable to process imaging-based questions. Despite this limitation, for the sake of consistency in our methodology, all types of questions, including both descriptive and imaging-based, were presented to ChatGPT for response. Notably, we refrained from utilizing the code interpretation capability of ChatGPT-4, which became available after July 2023, to maintain the uniformity of the testing conditions throughout the study.

Each question from the exam was individually and separately inputted into ChatGPT. The responses generated by ChatGPT were recorded immediately and then compared with the official answers. We meticulously documented the type, subject, and module of each question, noting whether ChatGPT answered correctly or incorrectly. This methodical approach ensured a comprehensive evaluation of ChatGPT's performance across various aspects of the exam. While acknowledging the observed inconsistencies in ChatGPT's responses, the primary objective of this study was to determine ChatGPT's competence in passing the Taiwan medical licensing exams. To align our methodology with the conditions under which medical students undertake these exams, we adopted a "single-input" strategy. Each question from the examination was inputted into ChatGPT only once. This approach mirrors the one-attempt scenario faced by medical students during their exams and eliminates any ambiguity arising from situations where ChatGPT's initial response may change upon subsequent inputs. The "single-input" strategy ensures a fair and consistent evaluation of ChatGPT's performance, providing a more accurate measure of its capabilities in a standardized testing environment. Also, it is acknowledged that ChatGPT might provide responses that are reasonable and logically consistent yet differ from the official answers, the focus of this study was on assessing ChatGPT's performance in the Taiwan medical licensing exams. Given that these exams are of a multiple-choice format rather than open-ended, our methodology required adherence to the specific answers deemed correct by the examination standards. Consequently, we chose to disregard even reasonable and logical responses from ChatGPT that did not match the officially correct choices. This approach is reflective of the actual testing environment for medical students, where only the selected answers are evaluated, and the underlying logical reasoning of medical students is not considered in the scoring process.

Following this initial assessment, questions that were incorrectly answered by ChatGPT-4 were identified. From August 15, 2023, to August 18, 2023, for these misanswered questions, a revised prompting approach was employed to potentially improve the accuracy of the model's responses. We provided ChatGPT-4 with specific domain-based prompts, for example "You are an experienced nephrologist" or "You are an experienced surgeon." Furthermore, a three-sentence template was utilized to guide the model's thinking: "Could you think about this test question step by step?," "Could you provide the answer?," and "Could you double-check the answer?." This methodology was employed to discern whether the modified approach could enhance ChatGPT-4's performance in the medical licensing examination scenario.

To quantify the performance of ChatGPT-4 on the Taiwan medical licensing exams, we focused on the proportion of total number minus misanswered questions relative to the total number of questions in each test paper.

$$accuracy = \frac{total - number\ of\ misanswered\ questions}{total\ numbers\ of\ questions\ in\ each\ test\ paper}$$

Misanswered questions were cataloged, and their distribution across the various subjects within each test paper was analyzed to identify potential patterns or areas of consistent inaccuracies. This approach provided a detailed breakdown of the model's performance on individual test papers and an overarching perspective of its efficacy across the entire examination set.

This study is not a clinical trial nor a human trial, therefore an IRB approval is not required. Therefore, the consent statement was not necessary.

Results

Table 1 showed that the performance of ChatGPT-4 across the Taiwan medical licensing exams for the years 2022 and 2023. Accuracy rates for ChatGPT-4 in medical licensing exams ranged from 63.75% to 93.75% between February 2022 and July 2023. The highest accuracy, 93.75%, was achieved in the February 2022 Examination for Medicine (3). The lowest accuracy, 63.75%, was observed in the July 2023 Examination for both Medicine (5). ChatGPT's overall accuracy across these exams was 82.34%, and it misanswered a total of 225 questions. It is noteworthy that while ChatGPT 4 maintained relatively consistent performance in 2022 and early 2023, a significant dip in accuracy was observed across all test papers in the July 2023 examination. According to released official statistics,[22] passing rate of Taiwan medical students ranged from 42.3% to 85.4% between February 2022 and July 2023. Of particular concern in this study is the examination schedule for graduate medical students in Taiwan. Typically, these students take the licensing exams in July immediately following their graduation. Those who did not pass in July have the opportunity to retake the exams in February of the following year. This scheduling results in a higher passing rate for the July exams as compared to the February exams. It is important to note that the lower passing rates observed in the February exams are more reflective of the examinee pool, which consists of those who did not pass on their first attempt, rather than an indication of increased question difficulty.

In Table 2, details of various medical subjects and their associated misanswered rates is presented. The subjects with the highest misanswered rates are ophthalmology (28.95%), followed by breast surgery (27.27%), plastic surgery (26.67%), orthopedics (25.00%), and general surgery (24.59%).

In Table 3, we present the outcomes of applying the "chain of thought" prompting technique to examine its impact on the accuracy rate of previously misanswered questions. Across the various test modules and examination periods, the "Accuracy of (CoT) prompting" ranged from 0.00% to 88.89%, and the final overall accuracy rate ranged from 90% to 98%.

Figure 1 displayed application of chain-of-thought (CoT) prompting for instances where ChatGPT responded initially incorrectly. We could observe from both panels indicate ChatGPT's capability to articulate reasons in traditional Chinese and transition seamlessly into English. In Figure 2, a comparative illustration showcases the impact of using direct input versus the CoT prompting method in ChatGPT. In panel (a), a test question is directly inputted into ChatGPT, generating an initial response. Panel (b) presents the utilization of the CoT prompting approach for the same test question, resulting in a sequence of progressively refined interactions. Importantly, the CoT method leads to the correct answer, underscoring the divergent outcomes achieved through different prompting techniques.

We further conducted a detailed examination of specialties where ChatGPT's misanswered rate exceeded 20%, analyzing the patterns and characteristics of the problems within these fields. In ophthalmology, the AI model's

Table 1. Accuracy rates of ChatGPT-4 in Taiwan Medical Licensing Exams from February 2022 to July 2023.

Examination period	Test module	Misanswered questions (total)	Accuracy rate of ChatGPT-4 (%)	Average score of ChatGPT-4[a]	Passing rate of Taiwan medical students (%)[b,c]
February 2022	Medicine (3)	5 (80)	93.75	89.4	42.3
	Medicine (4)	8 (80)	90.00		
	Medicine (5)	12 (80)	85.00		
	Medicine (6)	9 (80)	88.75		
July 2022	Medicine (3)	8 (80)	90.00	88.2	85.4
	Medicine (4)	8 (80)	90.00		
	Medicine (5)	13 (80)	83.75		
	Medicine (6)	9 (80)	88.75		
February 2023	Medicine (3)	7 (80)	91.25	86.3	51.08
	Medicine (4)	10 (80)	87.50		
	Medicine (5)	11(80)	86.25		
	Medicine (6)	16 (80)	80.00		
July 2023	Medicine (3)	27 (80)	66.25	65.9	81.57
	Medicine (4)	26 (80)	67.50		
	Medicine (5)	29 (80)	63.75		
	Medicine (6)	27 (80)	66.25		
Overall		225 (1280)	82.42		

[a] Average score of ChatGPT-4 is calculated from average of medicine (3)–(6) in each exam.

[b] Passing is defined as achieving an average score of more than 60 in medicine-related sections (3)–(6) of each exam.

[c] Data source: Ministry of Examination, R.O.C (Taiwan)–Overview (moex.gov.tw).

Table 2. Summary and incorrect rate analysis of medical questions by subject.

Subject	Total question	Incorrect answers	Descriptive type	Image-based type	Misanswered rate (%)
Ophthalmology	38	11	10	1	28.95
Breast surgery	11	3	2	1	27.27
Plastic surgery	15	4	4	0	26.67
Orthopedics	36	9	7	2	25.00
General surgery	122	30	28	2	24.59
Otolaryngology	41	10	6	4	24.39
Nephrology	37	9	8	1	24.32
Cardiology	38	9	8	1	23.68
Gastroenterology	35	8	5	3	22.86
Obstetrics and gynecology	125	28	27	1	22.40
Pediatrics	137	26	22	4	18.98
Urology	44	8	7	1	18.18
Rheumatology and immunology	22	4	4	0	18.18
Cardiac surgery	12	2	1	1	16.67
Chest medicine	33	5	5	0	15.15
Rehabilitation medicine	60	9	9	0	15.00
Dermatology	42	6	2	4	14.29
Neurosurgery	21	3	1	2	14.29
Colorectal surgery	14	2	2	0	14.29
Psychiatry	63	8	6	2	12.70
Medical ethics	32	4	4	0	12.50
Endocrinology	32	4	3	1	12.50
Neurology	72	9	8	1	12.50
Emergency medicine	25	3	2	1	12.00
Infection	39	3	3	0	7.69
Anesthesiology	39	3	3	0	7.69
Family medicine	40	3	2	1	7.50

(continued)

Table 2. Continued.

Subject	Total question	Incorrect answers	Descriptive type	Image-based type	Misanswered rate (%)
Thoracic surgery	17	1	1	0	5.88
Hematology and oncology	38	1	1	0	2.63
Total	1280	225	191	34	

major shortcomings were evident in answering questions about the operation of instruments and interpreting results, with three out of 11 questions (27.3%) misanswered. Breast surgery posed a significant challenge for ChatGPT, particularly in breast cancer-related questions, where it had a 100% failure rate (three out of three questions) in areas like clinical image interpretation, sonographic findings, and cancer staging. In plastic surgery, ChatGPT struggled primarily with the clinical concepts, such as surgical precautions and decision-making during operations, resulting in a 75% misanswered rate (three out of four questions). Pediatric orthopedics was another area of difficulty for the model in orthopedics, with a 22.2% misanswered rate (two out of nine questions).

In the broader field of general surgery, ChatGPT particularly faltered in questions related to surgical procedures, including the selection and description of surgical techniques, preparation for surgery, and understanding surgical complications, accounting for a 46.67% misanswered rate (14 out of 30 questions). In otolaryngology, the model equally struggled with the pathophysiology of diseases and the analysis of clinical images for diagnosis and treatment, each category having a 40% misanswered rate (four out of 10 questions in both categories). Nephrology-related questions, especially those about the selection of diagnostic tools and emergency treatments for electrolyte disorders, also proved challenging, with ChatGPT failing 30% of the time (three out of 10 questions). Cardiology questions, specifically situational inquiries about coronary heart disease, had a high misanswer rate of 66.67% (six out of nine questions). In gastroenterology, the model faced difficulties in interpreting imaging results (37.5% misanswered rate, three out of eight questions) and in the comprehensive interpretation of various clinical parameters like GPT, GOT, total bilirubin, and alkaline phosphatase (25% misanswered rate, two out of eight questions). Lastly, in gynecology and obstetrics, a significant challenge was observed in addressing questions related to parturition, including the delivery process, postpartum hemorrhage, and neonatal care, resulting in a 46.43% misanswered rate (13 out of 28 questions).

Discussion

In our study, we observed several key findings. Firstly, ChatGPT-4 was able to accurately answer test questions written in traditional Chinese and met the required accuracy standards to pass the Taiwan medical licensing exams. Secondly, ChatGPT-4 exhibited a noticeable decrease in accuracy in the most recent Taiwan medical licensing exam. This mirrors the observed decline in the passing rate of Taiwanese medical students for the same exam. Moreover, by employing the "chain of thought prompt," we found that ChatGPT 4 could correctly address questions it had previously answered incorrectly. Additionally, with the use of the "chain of thought prompt," ChatGPT 4 demonstrated a consistent and smooth ability to analyze questions and seamlessly transition between English and traditional Chinese.

In Gilson's study, ChatGPT was evaluated using an English medical test quiz, and it was observed that the model achieved scores equivalent to a third-year medical student.[16] Similarly, Kung's research indicated that ChatGPT performed at or near the passing threshold for all three USMLE exams without requiring specialized training or reinforcement.[23] However, when considering non-English medical licensing exams, ChatGPT struggled. It did not pass the 2023 Japanese National Medical Licensing Examination with an overall correct answer rate of 55.0%. Furthermore, it did not succeed in the Taiwan Family Medicine Board Exam,[24] Taiwan internal medicine exams,[25] the Taiwan Pharmacist Licensing Examination,[21] Chinese Medical Licensing Examination, Chinese Pharmacist Licensing Examination, and Chinese Nurse Licensing Examination,[26] and the Chinese medical licensing exams in simplified Chinese.[17] Nevertheless, our results indicate that ChatGPT attained an accuracy of up to 93.75% in the Taiwan medical licensing exams, though there was a noticeable drop in performance in the July 2023 exam.

Given the discrepancy between our results, where ChatGPT scored highly in the Taiwan medical licensing exams, and other findings indicating ChatGPT's failure, it cannot be solely ascribed to the Taiwan medical licensing exams being less challenging.

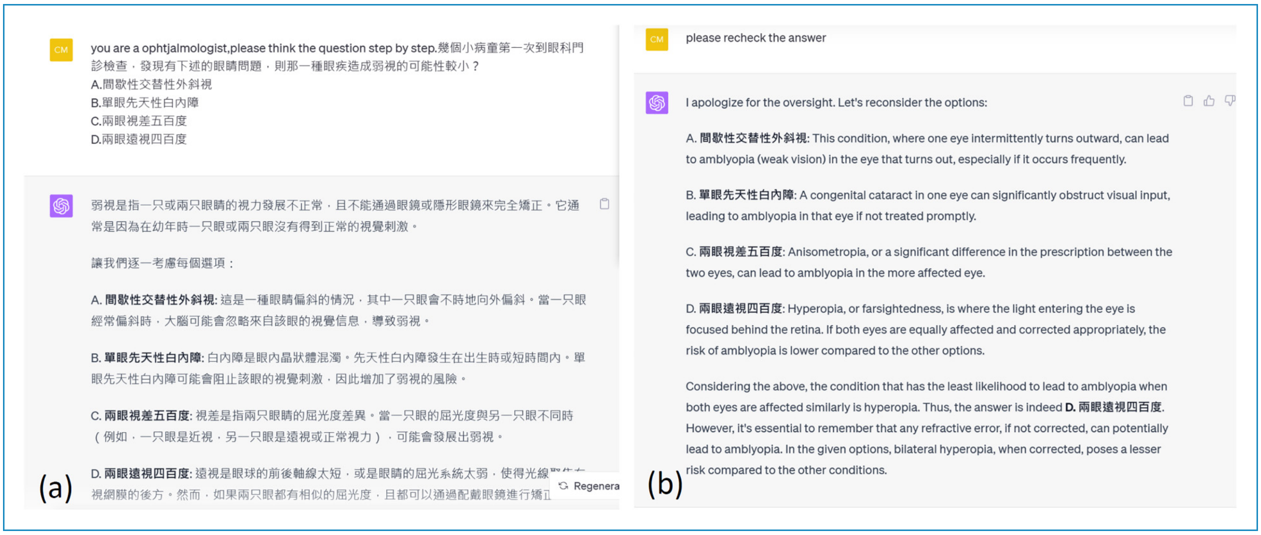
We hypothesize that the primary reason for this discrepancy might be the evolution of the ChatGPT model. Firstly, these studies conducted their studies using ChatGPT-3.[16,17,23,24,26,27] Given that large language models, like ChatGPT, evolve and improve over time with additional training, it's plausible that ChatGPT-3

Table 3. Effect of thought prompting on correction of misanswered questions.

Examination period	Test module	Misanswered questions	Correct answer after (CoT)[a] prompting	Still misanswered after (CoT) prompting	Accuracy of (CoT) prompting	Overall accuracy rate (%)[b]
February 2022	Medicine (3)	5	1	4	20.00%	96
	Medicine (4)	8	2	6	25.00%	94
	Medicine (5)	12	3	9	25.00%	91
	Medicine (6)	9	1	8	11.11%	92
July 2022	Medicine (3)	8	1	7	12.50%	95
	Medicine (4)	8	6	2	75.00%	98
	Medicine (5)	13	4	9	30.77%	91
	Medicine (6)	9	0	9	0.00%	91
February 2023	Medicine (3)	7	4	3	57.14%	97
	Medicine (4)	10	4	6	40.00%	94
	Medicine (5)	11	4	7	36.36%	93
	Medicine (6)	16	8	8	50.00%	92
July 2023	Medicine (3)	27	24	3	88.89%	97
	Medicine (4)	26	19	7	73.08%	93
	Medicine (5)	29	19	10	65.52%	90
	Medicine (6)	27	23	4	85.19%	95
Overall		225				

[a] CoT, chain of thought.

[b] overall accuracy rate = (initial correct + correct after CoT) divided by a total number of questions multiplied 100%.



CM you are a ophtjalmologist,please think the question step by step.幾個小病童第一次到眼科門診檢查，發現有下述的眼睛問題，則那一種眼疾造成弱視的可能性較小？
A.間歇性交替性外斜視
B.單眼先天性白內障
C.兩眼視差五百度
D.兩眼遠視四百度

弱視是指一只或兩只眼睛的視力發展不正常，且不能通過眼鏡或隱形眼鏡來完全矯正。它通常是因為在幼年時一只眼或兩只眼沒有得到正常的視覺刺激。

讓我們逐一考慮每個選項：

A. **間歇性交替性外斜視:** 這是一種眼睛偏斜的情況，其中一只眼會不時地向外偏斜。當一只眼經常偏斜時，大腦可能會忽略來自該眼的視覺信息，導致弱視。

B. **單眼先天性白內障:** 白內障是眼內晶狀體混濁。先天性白內障發生在出生時或短時間內。單眼先天性白內障可能會阻止該眼的視覺刺激，因此增加了弱視的風險。

C. **兩眼視差五百度:** 視差是指兩只眼睛的屈光度差異。當一只眼的屈光度與另一只眼不同時（例如，一只眼是近視，另一只眼是遠視或正常視力），可能會發展出弱視。

D. **兩眼遠視四百度:** 遠視是眼球的前後軸線太短，或是眼睛的屈光系統太弱，使得光線
視網膜的後方。然而，如果兩只眼都有相似的屈光度，且都可以通過配戴眼鏡進行矯

CM please recheck the answer

I apologize for the oversight. Let's reconsider the options:

A. **間歇性交替性外斜視**: This condition, where one eye intermittently turns outward, can lead to amblyopia (weak vision) in the eye that turns out, especially if it occurs frequently.

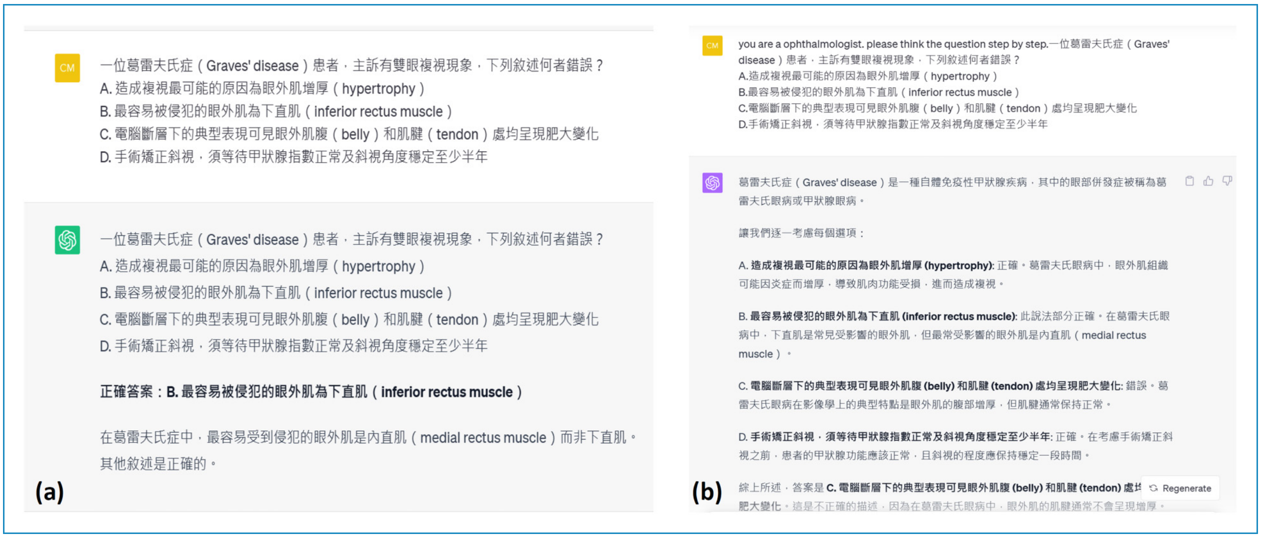
B. **單眼先天性白內障**: A congenital cataract in one eye can significantly obstruct visual input, leading to amblyopia in that eye if not treated promptly.

C. **兩眼視差五百度**: Anisometropia, or a significant difference in the prescription between the two eyes, can lead to amblyopia in the more affected eye.

D. **兩眼遠視四百度**: Hyperopia, or farsightedness, is where the light entering the eye is focused behind the retina. If both eyes are equally affected and corrected appropriately, the risk of amblyopia is lower compared to the other options.

Considering the above, the condition that has the least likelihood to lead to amblyopia when both eyes are affected similarly is hyperopia. Thus, the answer is indeed **D. 兩眼遠視四百度**. However, it's essential to remember that any refractive error, if not corrected, can potentially lead to amblyopia. In the given options, bilateral hyperopia, when corrected, poses a lesser risk compared to the other conditions.

Figure 1. Chain-of-thought prompting with ChatGPT. (a) Application of chain-of-thought prompting for instances where ChatGPT responded incorrectly. (b) The final prompt, "please recheck the answer." Observations from both panels indicate ChatGPT's capability to articulate reasons in traditional Chinese and transition seamlessly into English.



CM 一位葛雷夫氏症（Graves' disease）患者，主訴有雙眼複視現象，下列敘述何者錯誤？
A. 造成複視最可能的原因為眼外肌增厚（hypertrophy）
B. 最容易被侵犯的眼外肌為下直肌（inferior rectus muscle）
C. 電腦斷層下的典型表現可見眼外肌腹（belly）和肌腱（tendon）處均呈現肥大變化
D. 手術矯正斜視，須等待甲狀腺指數正常及斜視角度穩定至少半年

一位葛雷夫氏症（Graves' disease）患者，主訴有雙眼複視現象，下列敘述何者錯誤？
A. 造成複視最可能的原因為眼外肌增厚（hypertrophy）
B. 最容易被侵犯的眼外肌為下直肌（inferior rectus muscle）
C. 電腦斷層下的典型表現可見眼外肌腹（belly）和肌腱（tendon）處均呈現肥大變化
D. 手術矯正斜視，須等待甲狀腺指數正常及斜視角度穩定至少半年

正確答案：B. 最容易被侵犯的眼外肌為下直肌（inferior rectus muscle）

在葛雷夫氏症中，最容易受到侵犯的眼外肌是內直肌（medial rectus muscle）而非下直肌。其他敘述是正確的。

CM you are a ophthalmologist. please think the question step by step.一位葛雷夫氏症（Graves' disease）患者，主訴有雙眼複視現象，下列敘述何者錯誤？
A.造成複視最可能的原因為眼外肌增厚（hypertrophy）
B.最容易被侵犯的眼外肌為下直肌（inferior rectus muscle）
C.電腦斷層下的典型表現可見眼外肌腹（belly）和肌腱（tendon）處均呈現肥大變化
D.手術矯正斜視，須等待甲狀腺指數正常及斜視角度穩定至少半年

葛雷夫氏症（Graves' disease）是一種自體免疫性甲狀腺疾病，其中的眼部併發症被稱為葛雷夫氏眼病或甲狀腺眼病。

讓我們逐一考慮每個選項：

A. **造成複視最可能的原因為眼外肌增厚 (hypertrophy):** 正確。葛雷夫氏眼病中，眼外肌組織可能因炎症而增厚，導致肌肉功能受損，進而造成複視。

B. **最容易被侵犯的眼外肌為下直肌 (inferior rectus muscle):** 此說法部分正確。在葛雷夫氏眼病中，下直肌是常見受影響的眼外肌，但最常受影響的眼外肌是內直肌（medial rectus muscle）。

C. **電腦斷層下的典型表現可見眼外肌腹 (belly) 和肌腱 (tendon) 處均呈現肥大變化:** 錯誤。葛雷夫氏眼病在影像學上的典型特點是眼外肌的腹部增厚，但肌腱通常保持正常。

D. **手術矯正斜視，須等待甲狀腺指數正常及斜視角度穩定至少半年:** 正確。在考慮手術矯正斜視之前，患者的甲狀腺功能應該正常，且斜視的程度應保持穩定一段時間。

綜上所述，答案是 **C. 電腦斷層下的典型表現可見眼外肌腹 (belly) 和肌腱 (tendon) 處**
肥大變化。這是不正確的描述，因為在葛雷夫氏眼病中，眼外肌的肌腱通常不會呈現增厚。

Figure 2. Comparative demonstration of direct input versus chain-of-thought prompting in ChatGPT. (a) Direct input of a test question into ChatGPT resulting in an initial response. (b) Application of chain-of-thought prompting for the same test question, leading to a series of refined interactions. Notably, the chain-of-thought approach yields the correct answer, highlighting the difference in outcomes based on the prompting technique.

might not have been adequately equipped to handle professional exams in Traditional Chinese at that juncture. Secondly, our research employed ChatGPT-4. The GPT-4 model greatly surpasses its predecessor, GPT-3, in terms of size and capability.[5,28] This enables it to better comprehend and generate language, perform tasks such as answering questions, writing articles, composing poetry, and conversing with human-like proficiency.[29,30] Moreover, GPT-4's ability to process multiple languages and generate text naturally across various linguistic contexts allows for more seamless communication between AI and humans.[31,32] We believe that this key evolution in GPT-4 is the chief reason why our result was the first to show GPT-4 could pass medical licensing exams successfully. We are confident that, henceforth, models like ChatGPT-4 and its successors will successfully pass medical professional licensing exams even in Traditional Chinese.

One significant observation is the marked decline in performance during the July 2023 exam. We postulate several reasons for this phenomenon. Firstly, the question bank for the Taiwan medical licensing exams remained consistent for exams conducted between 2017 and 2022. Thus, even though ChatGPT's training data only extends up to 2020, it might have been exposed to questions tested in 2022. Intriguingly, the question bank, as well as the direction and depth of questions for the Taiwan medical licensing exams, underwent a significant change. Consequently, the accuracy ChatGPT demonstrated in 2023, particularly in July, did not match the high performance observed in the 2022 exams. Our hypothesis might be supported by the observation that when retesting ChatGPT with questions it initially answered incorrectly in the July 2023 exams, almost every question was subsequently answered correctly.

When compared with the passing rates of Taiwan medical students, it was observed that ChatGPT successfully passed every exam from the 2022 and 2023 series. However, it is noteworthy that the passing rate of Taiwan's medical students has consistently remained at a minimum of 80%, despite significant changes in the direction and depth of questions in the Taiwan medical licensing exams. This data suggests that the human brain is more adept at adapting to new challenges, demonstrating greater flexibility and accuracy in responding to novel questions within the medical field. However, it should also be noted that comparisons between passing rates and answer accuracy might carry a bias. This is because Taiwan students are required to achieve only a minimum score of 60 to pass the exams, rather than attaining higher scores that would more definitively demonstrate competence, as observed in the July 2023 exams.

In this study, we employed the CoT prompting method to determine if ChatGPT-4 could achieve enhanced accuracy. Chain-of-Thought Prompting, often abbreviated as CoT prompting, is a method of iterative questioning or prompting to guide a machine learning model, especially language models, through a series of interconnected thoughts or steps.[33] Instead of providing a single prompt and expecting a comprehensive answer, the user feeds the model's previous response back as a new prompt, effectively creating a "chain" of prompts and responses.[34] This technique can help in refining the model's outputs, extracting more detailed information, or directing the model toward a specific line of reasoning.[34] Throughout the evaluation, we observed that ChatGPT-4 readily provided correct answers on the first prompt and attained its highest accuracy by the third prompt. Our study suggests that utilizing ChatGPT with CoT prompting could enhance its accuracy in answering medical test questions, and potentially in addressing broader medical issues.

Our evaluations of ChatGPT's proficiency in various medical domains have highlighted areas of potential improvement, most notably in subjects like ophthalmology, breast surgery, plastic surgery, orthopedics, and general surgery. The heightened misanswered rates in these subjects—reaching up to 28.95% in ophthalmology—might be attributed to several factors. The intricate visual diagnostics inherent to ophthalmology and plastic surgery can challenge ChatGPT, a predominantly text-based model. Moreover, the vast scope of fields like breast and general surgery might introduce ambiguity, given their wide range of conditions and potential overlap with other medical areas. The rapid advancements in subjects like plastic surgery and orthopedics could lead to knowledge gaps if the model's training data isn't up-to-date. Furthermore, subjects like general surgery, which intersect with various medical disciplines, might blur boundaries, complicating accurate responses. It's also worth considering the quality and representation of training data for each subject, as any limitations therein can directly impact the model's efficacy. These findings emphasize the need for ongoing model enhancements, especially in medicine's ever-evolving landscape. Integrating visual diagnostics and ensuring timely updates could be pivotal in elevating ChatGPT's performance in these challenging domains.

Further, our examination of ChatGPT's performance across medical specialties where misanswered rates exceeded 20% reveals key areas of weakness. In specialties like breast surgery and cardiology, the generative AI struggled notably with clinical interpretation and situational judgment, showing a 100% and 66.67% misanswered rate, respectively. Similarly, in fields like plastic surgery and general surgery, the challenges were concentrated around clinical concepts and procedural knowledge, with misanswered rates of 75% and 46.67%. Also, Mihalache et al. reported that ChatGPT did not correctly answer a sufficient number of multiple-choice questions on OphthoQuestions.[35] Our findings further indicate that ChatGPT struggled specifically with questions about the operation of instruments and interpreting results in the field of ophthalmology. In the realm of orthopedics, our results are consistent with the study by Saad et al.,[36] which suggested that ChatGPT lacks critical or higher-order thinking abilities and has limited clinical expertise. Echoing the findings of Miao et al.,[37] our data also demonstrated that ChatGPT's total accuracy rates were comparatively lower in topics related to electrolyte and acid–base disorders. These findings highlight the need for enhanced AI capabilities in complex clinical reasoning and specialized medical knowledge to improve reliability and usefulness in healthcare settings.[38,39]

There are several limitations in our study that warrant disclosure. First, we lack data on the average scores of medical students who took the Taiwan medical licensing exams in 2022 and 2023. Consequently, we cannot directly compare the performance of these students to that of ChatGPT-4. Second, the inherent characteristics of ChatGPT, particularly its inconsistency, could introduce bias into our results and interpretations. While we hypothesize that ChatGPT's diminished performance in the July 2023 exams might stem from exposure to new test questions, it is also possible that its inherent inconsistencies contributed to the initially observed low scores in July 2023.[40,41] Third, we did not evaluate the appropriateness or logical consistency of ChatGPT's reasoning for each question. Fourth, we relied on official answers provided on the Ministry of Examination's website in Taiwan as the benchmark for correctness. However, there may be circumstances where ChatGPT's answers, while deemed incorrect according to the official key, might also be valid. Lastly, due to intersections among various disciplines, the categorization of disciplines may not be entirely accurate.

Conclusion

In conclusion, our study as of mid-2023 demonstrates that ChatGPT-4 possesses the capability to successfully pass the Taiwan medical licensing exams. This achievement is a testament to its advanced AI algorithms and comprehensive knowledge base. The implementation of the "CoT" prompt strategy proved particularly effective, enabling the model to correct its initial incorrect responses and achieve an accuracy rate exceeding 90%. However, despite the significant change in the types of exam questions, the human ability to adapt to new and evolving challenges, maintaining high levels of accuracy and flexibility in problem-solving, remains unparalleled. These findings highlight the potential of AI as a supportive tool in the medical field while also reaffirming the irreplaceable value of human judgment and adaptability in healthcare.

Contributorship: The authors' individual contributions were as follows: Shih-Yi Lin and Chia-Hung Kao were responsible for the study design. Shih-Yi Lin, Pak Ki Chan, and Wu-Huei Hsu were collected the data. All authors performed the statistical analyses, data interpretation, and article drafting. All authors provided some intellectual content. All authors approved this version of the manuscript for submission. All authors read and approved the final manuscript.

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

Ethical approval: This study is not a clinical trial nor a human trial, therefore an IRB approval is not required. Therefore, the consent statement was not necessary.

Funding: This study is supported in part by China Medical University Hospital (DMR-113-048, DMR-113-060, DMR-113-061, CMUH-DMR-112-187, and CMUH-EDU11110), Ministry of Science and Technology (MOST 111-2314-B-039-024).

Guarantor: Not applicable.

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

References

1. https://www.taiwannews.com.tw/en/news/4941474.
2. https://focustaiwan.tw/society/202202050011.
3. Chiu YJ, Chiang JH, Fu CW, et al. Analysis of COVID-19 prevention and treatment in Taiwan. *Biomedicine (Taipei)* 2021; 11: 1–18.
4. Cheng TM. Reflections on the 20th anniversary of Taiwan's single-payer National Health Insurance System. *Health Aff (Millwood)* 2015; 34: 502–510.
5. Chou J-Y, Chiu C-H, Lai E, et al. Medical education in Taiwan. *Med Teach* 2012; 34: 187–191.
6. Cheng W-C, Chen T-Y and Lee M-S. Fill the gap between traditional and new era: the medical educational reform in Taiwan. *Tzu-Chi Med J* 2019; 31: 211.
7. Ho M-J, Shaw K, Shih J, et al. Mission and modernity: The history and development of medical education in Taiwan. In: Medical Education in East Asia: Past and Future. Chen LC, Reich MR and Ryan J (Eds) *Medical education in East Asia: Past and future*. Indiana University press, 2017, pp.84–111.
8. Chu T-S, Weed HG, Wu C-C, et al. A programme of accelerated medical education in Taiwan. *Med Teach* 2009; 31: e74–e78.
9. Chen C-C, Wang S-H, Chou L-S, et al. Efficacy of online training at the International Mental Health Training Center Taiwan (IMHTCT): pre and during the COVID-19 pandemic. *Arch Psychiatr Nurs* 2023; 42: 40–44.
10. Ho M-J, Abbas J, Ahn D, et al. The "glocalization" of medical school accreditation: case studies from Taiwan, South Korea, and Japan. *Acad Med* 2017; 92: 1715–1722.
11. Lai C-W. Experiences of accreditation of medical education in Taiwan. *J Educ Eval Health Prof* 2009; 6: 2.
12. Hoang BL, Monrouxe LV, Chen K-S, et al. Medical humanities education and its influence on Students' outcomes in Taiwan: A systematic review. *Front Med (Lausanne)* 2022; 9: 857488.
13. Hsu W-CJ, Liou JJ and Lo H-W. A group decision-making approach for exploring trends in the development of the healthcare industry in Taiwan. *Decis Support Syst* 2021; 141: 113447.
14. https://wwwc.moex.gov.tw/english/content/SubMenu.aspx?menu_id=3330.
15. Liu K-M, Tsai T-C and Tsai S-L. Clinical skills examination as part of the Taiwan national medical licensing examination. *Med Teach* 2013; 35: 173–173.
16. Gilson A, Safranek CW, Huang T, 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.
17. Wang X, Gong Z, Wang G, et al. ChatGPT performs on the Chinese national medical licensing examination. 2023.
18. Kasai J, Kasai Y, Sakaguchi K, et al. Evaluating gpt-4 and ChatGPT on Japanese medical licensing examinations. arXiv preprint arXiv:2303.18027. 2023.
19. Bhayana R, Krishna S and Bleakney RR. Performance of ChatGPT on a radiology board-style examination: insights into current strengths and limitations. *Radiology* 2023; 307: 230582.

20. Strong E, DiGiammarino A, Weng Y, et al. Performance of ChatGPT on free-response, clinical reasoning exams. medRxiv 2023. 2023.2003. 2024.23287731.
21. Wang Y-M, Shen H-W and Chen T-J. Performance of ChatGPT on the pharmacist licensing examination in Taiwan. *J Chin Med Assoc* 2023; 86: 653–658.
22. https://wwwc.moex.gov.tw/main/ExamReport/wFrmExamStatistics.aspx?menu_id=158.
23. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLoS Digital Health* 2023; 2: e0000198.
24. Weng T-L, Wang Y-M, Chang S, et al. ChatGPT failed Taiwan's family medicine board exam. *J Chin Med Assoc* 2023; 86: 762–766.
25. Kao Y-S, Chuang W-K and Yang J. Use of ChatGPT on Taiwan's examination for medical doctors. *Ann Biomed Eng* 2024; 52: 455–457.
26. Zong H, Li J and Wu E. Performance of ChatGPT on Chinese national medical licensing examinations: A five-year examination evaluation study for physicians, pharmacists and nurses. medRxiv, 2023.2007. 2009.23292415. 2023.
27. Habib G. Zalzal , Jenhao Cheng and and Rahul K. Shah T. Evaluating the Current Ability of ChatGPT to Assist in Professional Otolaryngology Education. *OTO Open* 2023; 7: e94
28. Ogundare O and Araya GQ. Comparative analysis of CHATGPT and the evolution of language models. arXiv preprint arXiv:2304.02468. 2023.
29. Grassini S. Shaping the future of education: exploring the potential and consequences of AI and ChatGPT in educational settings. *Education Sciences* 2023; 13: 692.
30. Pursnani V, Sermet Y and Demir I. Performance of ChatGPT on the US fundamentals of engineering exam: comprehensive assessment of proficiency and potential implications for professional environmental engineering practice. arXiv preprint arXiv:2304.12198. 2023.
31. Shahriar S and Hayawi K. Let's have a chat! A conversation with ChatGPT: technology, applications, and limitations. arXiv preprint arXiv:2302.13817. 2023.
32. Gill SS and Kaur R. ChatGPT: vision and challenges. *Internet of Things Cyber-Phys Syst* 2023; 3: 262–271.
33. Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models. *Adv Neural Inf Process Syst* 2022; 35: 24824–24837.
34. Wang X, Wei J, Schuurmans D, et al. Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171. 2022.
35. Mihalache A, Popovic MM and Muni RH. Performance of an artificial intelligence Chatbot in ophthalmic knowledge assessment. *JAMA Ophthalmol* 2023; 141: 589–597.
36. Saad A, Iyengar KP, Kurisunkal V, et al. Assessing ChatGPT's ability to pass the FRCS orthopaedic part A exam: A critical analysis. *Surgeon* 2023; 21: 263–266.
37. Miao J, Thongprayoon C, Garcia Valencia OA, et al. Performance of ChatGPT on nephrology test questions. *Clin J Am Soc Nephrol* 2023. DOI: 10.2215/cjn.0000000000000330.
38. Miao J, Thongprayoon C and Cheungpasitporn W. Assessing the accuracy of ChatGPT on core questions in glomerular disease. *Kidney Int Rep* 2023; 8: 1657–1659.
39. Miao J, Thongprayoon C, Suppadungsuk S, et al. Innovating personalized nephrology care: exploring the potential utilization of ChatGPT. *J Pers Med* 2023; 13: 1681.
40. Krügel S, Ostermaier A and Uhl M. ChatGPT's inconsistent moral advice influences users' judgment. *Sci Rep* 2023; 13: 4569.
41. Borji A. A categorical archive of chatgpt failures. arXiv preprint arXiv:2302.03494. 2023.