

Assessing AI efficacy in medical knowledge tests: A study using Taiwan's internal medicine exam questions from 2020 to 2023

DIGITAL HEALTH
Volume 10: 1–10
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20552076241291404
journals.sagepub.com/home/dhj



Shih-Yi Lin^{1,2}, Ying-Yu Hsu³ , Shu-Woei Ju², Pei-Chun Yeh⁴,
Wu-Huei Hsu^{1,5} and Chia-Hung Kao^{1,4,6,7} 

Abstract

Background: The aim of this study is to evaluate the ability of generative artificial intelligence (AI) models to handle specialized medical knowledge and problem-solving in a formal examination context.

Methods: This research utilized internal medicine exam questions provided by the Taiwan Internal Medicine Society from 2020 to 2023, testing three AI models: GPT-4o, Claude_3.5 Sonnet, and Gemini Advanced models. Rejected queries for Gemini Advanced were translated into French for resubmission. Performance was assessed using IBM SPSS Statistics 26, with accuracy percentages calculated and statistical analyses such as Pearson correlation and analysis of variance (ANOVA) performed to gauge AI efficacy.

Results: GPT-4o's top annual score was 86.25 in 2022, with an average of 81.97. Claude_3.5 Sonnet reached a peak score of 88.13 in 2021 and 2022, averaging 84.85, while Gemini Advanced lagged with an average score of 69.84. In specific specialties, Claude_3.5 Sonnet scored highest in Psychiatry (100%) and Nephrology (97.26%), with GPT-4o performing similarly well in Hematology & oncology (97.10%) and Nephrology (94.52%). Gemini's best scores were in Psychiatry (86.96%) and Hematology & Oncology (82.76%). Gemini Advanced models struggled with Neurology, scoring below 60%. Additionally, all models performed better on text-based questions than on image-based ones, without significant differences. Claude 3 Opus scored highest on COVID-19-related questions at 89.29%, followed by GPT-4o at 75.00% and Gemini Advanced at 67.86%.

Conclusions: AI models showed varied proficiency across medical specialties and question types. GPT-4o demonstrated higher image-based correction rates. Claude_3.5 Sonnet generally and consistently outperformed others, highlighting significant potential for AI in assisting medical education.

Keywords

Generative AI models, internal medicine exam, ChatGPT-4, Claude 3 Opus, Gemini

Submission date: 22 May 2024; Acceptance date: 27 September 2024

¹Graduate Institute of Biomedical Sciences, College of Medicine, China Medical University, Taichung

²Division of Nephrology and Kidney Institute, China Medical University Hospital, Taichung

³11th Grade Student, National Changhua Senior High School, Changhua

⁴Artificial Intelligence Center, China Medical University Hospital, Taichung

⁵Department of Chest Medicine, China Medical University Hospital, Taichung

⁶Department of Nuclear Medicine and PET Center, China Medical University Hospital, Taichung

⁷Department of Bioinformatics and Medical Engineering, Asia University, Taichung

S-YL, Y-YH, and S-WJ contributed equally and shared the first authorship.

Corresponding author:

Chia-Hung Kao, Graduate Institute of Biomedical Sciences and School of Medicine, College of Medicine, China Medical University, No. 2, Yuh-Der Road, Taichung 404.

Emails: dr.kaochiahung@gmail.com, 010040@tool.caamed.org.tw



Introduction

Since the development of generative artificial intelligence (AI) in 2020,¹ it has played an increasingly significant role in medical education, training, and advanced medical practice.² There is growing confidence that generative AI will revolutionize the field of medicine.³ Previous studies have explored the use of generative AI for interpreting data and medical images.^{4,5} Additionally, some research has employed medical licensing exams from various countries to assess whether generative AI could potentially meet the qualifications required of a doctor.^{6–10} However, few studies have specifically examined AI performance in medical specialty certification exams to assess its ability to independently manage patient care in subspecialty fields and determine if generative AI could potentially pass these examinations, thereby potentially qualifying as a knowledgeable subspecialist.¹¹

Internal medicine encompasses a wide range of subspecialties, each with its distinct scientific considerations, therefore, when discussing a patient's condition, a practitioner must actively apply a diverse set of knowledge.¹² Consequently, using internal medicine exam questions to evaluate generative AI goes beyond merely determining pass/fail outcomes or scores. We could identify the cognitive limitations of generative AI through such questions. By understanding these limitations, it would be more effective and better to integrate generative AI into our clinical practice in the future.

The COVID-19 pandemic emerged in 2019, and with its outbreak, numerous internal medicine challenges arose, including issues related to immunology,¹³ infections, pharmacology, chronic diseases, and cardiovascular care. We also aim to use topical questions from the internal medicine specialty exams to evaluate whether generative AI can handle such emerging knowledge and novel diseases. Additionally, this approach helps us discern which generative AI systems are better equipped to address these contemporary and emergent issues.

In 2024, the landscape of generative AI expanded beyond ChatGPT-4 to include competitors such as Claude 3 Opus,¹⁴ and Gemini.¹⁵ Evaluating these products using medical exams can provide valuable insights into their capabilities regarding data bias and quality, transparency and explainability, ethical considerations, and adaptability to new medical challenges like COVID-19. Given that the Internal Medicine Exam questions in Taiwan from 2020 to 2023 encompassed COVID-19-related content—aligning with the advent of generative AI—we employed these three AI models to attempt the exam. Our goal was to assess their effectiveness in addressing internal medicine-related questions, current medical issues such as COVID-19, and situational problem-solving. This study aims to assess the current capabilities of generative AI models in handling specialized medical knowledge and problem-solving within the context of a formal examination setting.

Methods

This study utilized the internal medicine specialty exam questions and standard answers provided by the Taiwan Internal Medicine Society from 2020 to 2023. The official answers served as the reference standard.¹⁶ The Taiwan Internal Medicine Specialty Examination comprised 200 questions in 2020, while from 2021 to 2023, each annual examination consisted of 160 questions. Researchers conducted the experiment using three AI models: GPT-4o, Claude_3.5 Sonnet, and Gemini Advanced models. The research on three generative AI models—GPT-4o, Claude_3.5 Sonnet, and Gemini Advanced—was conducted during the period from 11 to 15 August 2024 in Taichung, Taiwan.

The written exam format involves multiple-choice questions presented in Chinese, with specialized terms accompanied by English translations. The duration of the exam ranges from two to three hours. The subject areas examined included cardiovascular, respiratory, gastrointestinal, metabolic and endocrine, renal, rheumatological, immunological, and allergic disorders, as well as hematological, oncological, infectious, neurological, psychiatric, and dermatological diseases pertinent to internal medicine. The question distribution in the annual Taiwan Internal Medicine Specialty Exams mainly emphasized Cardiovascular, Chest, Gastroenterology, and Nephrology. This was followed by Endocrinology, Infectious Diseases, Hematology & Oncology, and Rheumatology, with Neurology, Psychiatry, and Dermatology receiving fewer questions. A comprehensive breakdown of question distribution by specialty from 2020 to 2023 is detailed in Supplemental Appendix Table 1. The passing criterion of Taiwan Internal Medicine Specialty written Examination was a score of 60 points or above out of 100.

Due to usage restrictions associated with the GPT-4o and Claude_3.5 Sonnet models, questions were alternately posed to each system. Additionally, the Gemini Advanced model was programmed to decline questions pertaining to medical images. For queries rejected by Gemini Advanced, the questions were translated into French using Google Translate and then successfully resubmitted for responses.

Each model was asked to answer the exam questions, and their responses were compared with the official answers to verify accuracy. The input method for GPT-4o, Claude_3.5 Sonnet, and Gemini Advanced was consistent to maintain fairness in the evaluation.

The method of data input varied depending on the type of question, which was categorized into two main types: text-based and image-based. For text-based questions, all models employed a standardized input system where each batch consisted of 20 questions. In instances where the total number of questions did not divide evenly by 20, as observed in the 2020 dataset which included 185 text-based questions, the final batch was comprised of the remaining questions (e.g. 5 questions) along with 15 previously

answered questions to complete the batch of 20. The responses for these reintegrated questions were based on their original inputs.

In contrast, for image-based questions, each question was processed individually.

The input process was conducted by the same researcher to ensure consistency throughout the study. This study was conducted to test AI models' performance on Taiwan internal medicine exams. As the study involved no human participants or personal data collection, it did not require approval from the Institutional Review Board (IRB).

Statistical analysis

The statistical analysis in this study employed IBM SPSS Statistics 26, leveraging a comprehensive approach to evaluate the performance of three AI models across medical subject questions from 2020 to 2023. The study examined distributions of question types, particularly focusing on "Knowledge" and "Context" categories as well as COVID-19-related questions (Supplemental Appendix Table 2). By calculating the percentage of correctly answered questions for each model, the analysis provided a quantitative measure of each model's accuracy within specific medical subjects. This score, derived by dividing the number of correct responses by the total questions, allows for a direct comparison of model performance, offering insight into their respective capabilities in diverse domains of medical knowledge.

The use of Pearson correlation coefficients to analyze the interrelationships among the three models adds another layer of complexity to the analysis. By assessing the strength and direction of linear relationships between the models' performances across different subjects, the study not only identified where model outputs aligned but also highlighted areas where they diverged. A strong positive correlation would suggest consistency in how the models handle specific medical content, while weaker correlations may indicate variability in their underlying processing

mechanisms. Understanding these correlations is critical for interpreting how different AI models approach the same set of clinical data, which has implications for their potential integration into real-world medical settings.

Analysis of variance (ANOVA) was applied to assess the statistical significance of differences in performance between the models, providing a rigorous method to determine whether observed variations are due to chance or represent meaningful discrepancies. The ANOVA results, coupled with the Chi-square test—used to evaluate the differences between observed and expected frequencies—offer robust evidence regarding the models' differential capabilities. With a significance level set at 0.05, the statistical tools employed in this study were designed not just to confirm variability but to explore the underlying causes of these differences. This multifaceted analysis moves beyond mere description, offering insights into how each AI model handles distinct types of medical data and providing a foundation for understanding how these systems may be optimized for clinical use.

Results

Table 1 presents a comparative analysis of the performance of GPT-4o, Claude_3.5 Sonnet, and Gemini Advanced over the years 2020 to 2023, with distinct trends emerging. GPT-4o achieved its peak performance in 2022 with a score of 86.25, while maintaining a relatively stable average of 81.97 across the 4-year span. Claude_3.5 Sonnet, however, reached its highest scores in 2021 and 2022, with a notable average of 84.85, indicating superior consistency across the years compared to the other models. In contrast, Gemini Advanced consistently underperformed, achieving the lowest scores across all years, with an average score of 69.84, suggesting a potential limitation in its ability to handle the diverse question types and medical specialties examined.

When dissecting performance by specialty, Claude_3.5 Sonnet stood out with the highest overall average score

Table 1. Comparative performance and scoring of GPT-4o, Claude 3 Opus, and Gemini from 2020 to 2023.

Year/model	GPT-4o	Claude_3.5 Sonnet		Gemini Advanced		Passing rate	Total questions	
	Correct	Score	Correct	Score	Correct			Score
2020	162	81.00	167	83.50	140	70.00	90.57%	200
2021	128	80.00	141	88.13	118	73.75	91.03%	160
2022	138	86.25	141	88.13	112	70.00	90.53%	160
2023	129	80.63	128	80.00	105	65.63	90.75%	160
Average	139.25	81.97	144.25	84.85	118.75	69.84	90.72%	170

rate of 84.85%, followed by GPT-4o at 81.91%, while Gemini Advanced lagged at 69.85%. Claude_3.5 Sonnet demonstrated exceptional performance in Psychiatry, achieving a perfect score (100%), and excelled in Nephrology (97.26%), underscoring its strength in specialties requiring nuanced understanding. In comparison, GPT-4o's strongest area was Hematology & Oncology, with a score of 97.1%, showing its competence in this domain. Despite its overall lower scores, Gemini Advanced performed relatively better in Psychiatry (86.96%) and Hematology & Oncology (82.91%), suggesting its potential strength in specific areas. Notably, Claude_3.5 Sonnet's weakest performance was in Gastroenterology, while Gemini Advanced exhibited significant deficiencies in Neurology, with a concerning score of 47.83%, highlighting variability in model specialization and effectiveness across different medical fields (Table 2).

The ANOVA further reinforces these observations, revealing statistically significant differences in performance among the three models across various specialties, with a significance level of $p < 0.01$. The F-values (GPT-4o: 4.575, Claude_3.5 Sonnet: 5.593, Gemini Advanced: 2.572)

suggest that the differences in mean scores between specialties are not due to random variation but rather reflect meaningful distinctions in model performance. Claude_3.5 Sonnet exhibited the greatest variability between specialties, suggesting a higher degree of specialization in certain fields, while Gemini Advanced showed the least variability, though this may reflect its consistently lower performance. Interestingly, despite its higher variability across specialties, Claude_3.5 Sonnet demonstrated the most consistent performance within each specialty, as evidenced by its low mean square value of 0.121, indicating smaller variations in performance across different items within a specific specialty (Supplemental Appendix Table 3).

The data highlights the varying strengths of the AI models when categorized by question type and content. Claude_3.5 Sonnet demonstrated a clear advantage in knowledge-based questions, leading with a score of 86.99, followed by GPT-4o at 81.89, and Gemini Advanced at 71.68. This suggests that Claude_3.5 Sonnet is more adept at assimilating and recalling factual medical information. For scenario-based questions, Claude_3.5 Sonnet and GPT-4o performed similarly,

Table 2. The correction rates and scores of three generative AI models—ChatGPT, Claude 3 Opus, and Gemini—across different medical specialties for the Taiwanese internal medicine exams from 2020 to 2023.

Specialty /correct score	Total	GPT-4o		Claude_3.5 Sonnet		Gemini Advanced	
		Questions	Score (%)	Questions	Score (%)	Questions	Score (%)
Cardiovascular	82	61	74.39	64	78.05	51	62.20
Chest	81	63	77.78	65	80.25	54	66.67
Dermatology	22	68	80.95	71	84.52	54	64.29
Endocrinology	72	57	70.37	69	85.19	53	65.43
Gastroenterology	84	53	73.61	47	65.28	48	66.67
Hematology & oncology	73	67	97.10	67	97.10	57	82.61
Infection	70	64	91.43	65	92.86	54	77.14
Nephrology	81	69	94.52	71	97.26	59	80.82
Neurology	23	15	65.22	17	73.91	11	47.83
Psychiatry	23	20	86.96	23	100.00	20	86.96
Rheumatology	69	20	90.91	18	81.82	14	63.64
Total	680	557	81.91	577	84.85	475	69.85

Note. Data overview: This table presents the performance of three AI models (GPT-4o, Claude opus, and Gemini) across various medical subjects. Performance is measured by the percentage of correctly answered questions out of the total questions presented in each subject.

Scoring method: The score (%) for each model in each subject is derived from the number of correctly answered questions divided by the total questions, multiplied by 100. This provides a percentage score reflecting the accuracy of each model in the specific subject area.

each achieving a score of 81.94, further emphasizing their ability to analyze and interpret clinical scenarios effectively. Conversely, Gemini Advanced underperformed in scenario-based questions with a score of 67.36, reflecting limitations in applying knowledge to real-world clinical contexts. These results suggest that while Claude_3.5 Sonnet and GPT-4o are comparably strong in scenario-based reasoning, Gemini Advanced may struggle with the complexity of clinical applications (Table 3).

The statistical analysis, including p-values, further highlights differences in how these models handle various question types. GPT-4o exhibited p-values well above 0.05 across all tests, indicating no significant differences between its handling of knowledge-based questions and contextual data. This suggests that GPT-4o maintains stable performance without notable peaks or weaknesses across different data types. In contrast, Claude_3.5 Sonnet showed a p-value close to significance in a two-tailed test ($p=0.070$) and reached significance in a one-tailed test ($p=0.045$), indicating potentially stronger performance in a specific direction. Similarly, Gemini Advanced displayed no significant differences in performance across data types, with p-values above 0.05, suggesting a consistent—though generally lower—performance level (Supplemental Appendix Table 4).

When evaluating text-based and image-based questions, the results indicate that all three models performed better on text-based questions than on image-based ones, although the degree of difference varied. Claude_3.5 Sonnet achieved a score of 86.52% on text-based questions and dropped to

66.67% on image-based questions, indicating a notable decline in performance when dealing with visual data. GPT-4o scored 82.34% on text-based questions and 77.19% on image-based questions, showing a more balanced and consistent performance across both question types. Meanwhile, Gemini Advanced scored 72.39% on text-based questions but fell significantly to 42.11% on image-based questions, reflecting a severe limitation in its ability to interpret and process visual information. Interestingly, GPT-4o demonstrated a relatively higher proficiency in handling image-based questions, with 44 out of its 557 correct responses being image-based, compared to Claude_3.5 Sonnet, which answered 38 image-based questions correctly out of 577 correct responses (Table 4). This finding indicates GPT-4o's relative strength in processing visual data compared to its counterparts.

The Chi-square test further supports these observations, indicating significant differences in correction rates for text-based and image-based questions for Claude_3.5 Sonnet and Gemini Advanced ($p<0.001$). This suggests that these two models exhibit substantial differences in performance between these question types. GPT-4o, however, did not show significant differences in its performance across text-based and image-based questions ($p=0.368$), suggesting a more uniform ability to handle both types of data. This consistent performance across different data types implies that GPT-4o may be more versatile, while Claude_3.5 Sonnet and Gemini Advanced are more specialized, particularly in how they differentiate between text-based and image-based questions (Supplemental Appendix Tables 5 and 6).

Table 3. Performance of Claude 3 Opus, ChatGPT-4, and Gemini on knowledge-based and scenario-based questions (2020-2023).

Integrated/correct score	Total	GPT-4o		Claude_3.5 Sonnet		Gemini Advanced	
		Questions	Score (%)	Questions	Score (%)	Questions	Score (%)
Knowledge-based questions	392	321	81.89	341	86.99	281	71.68
Scenario-based questions	288	236	81.94	236	81.94	194	67.36
Total	680	557	81.91	577	84.85	475	69.85

Table 4. Comparison of AI models' performance on text-based and image-based questions.

Category/correct score	Total	GPT-4o		Claude_3.5 Sonnet		Gemini Advanced	
		Questions	Score (%)	Questions	Score (%)	Questions	Score (%)
Image-based questions	57	44	77.19	38	66.67	24	42.11
Text-based questions	623	513	82.34	539	86.52	451	72.39
Total	680	557	81.91	577	84.85	475	69.85

Table 5. Scoring rates of AI models on COVID-19-related versus unrelated questions.

Category/correct score	Total	GPT-4o		Claude_3.5 Sonnet		Gemini Advanced	
		Questions	Score (%)	Questions	Score (%)	Questions	Score (%)
COVID-19 related	27	21	75.00	25	89.29	19	67.86
COVID-19 unrelated	653	536	82.21	552	84.66	456	69.94
Total	680	557	81.91	577	84.85	475	69.85

When focusing on questions related to current events, specifically COVID-19, the analysis reveals key insights into the models' adaptability. Between 2020 and 2023, there were 27 COVID-19-related questions and 653 unrelated questions. Claude_3.5 Sonnet achieved the highest score on COVID-19-related questions with 89.29%, outperforming GPT-4o, which scored 75.00%, and Gemini Advanced, which scored 67.86%. This suggests that Claude_3.5 Sonnet was better equipped to process and incorporate rapidly evolving information related to the pandemic. Despite this, all three models performed better on COVID-19-unrelated questions, indicating that while they can handle pandemic-related content, their performance may still be optimized in areas with more stable, established knowledge bases (Table 5). Notably, no model demonstrated strong statistical significance in their performance differences between COVID-19-related and unrelated questions, indicating that the differences in handling these two categories may not be substantial across all tests (Supplemental Appendix Tables 7 and 8).

Discussion

This study evaluated the performance of three AI models—GPT-4o, Claude_3.5 Sonnet, and Gemini Advanced—from 2020 to 2023 across various question formats and specialties. Claude_3.5 Sonnet consistently outperformed the others, achieving high scores particularly in Psychiatry and Nephrology, with its peak performance in 2021 and 2022. GPT-4o demonstrated variability but excelled in scenario-based, though it underperformed in specialties like Neurology and Endocrinology. Gemini Advanced, on the other hand, generally scored lower, with some exceptions in Psychiatry. Additionally, variations in correction rates between knowledge-based and scenario-based questions were significant for Claude_3.5 Sonnet, highlighting different capabilities in question handling. Claude_3.5 Sonnet also showed superior performance in text-based questions compared to image-based ones. Performance on COVID-19-related questions was better of Claude_3.5 Sonnet models compared to COVID-19-unrelated topics, though the differences were not statistically significant.

Overall, this analysis provides insights into each model's strengths and weaknesses, indicating Claude_3.5 Sonnet as the most consistent and versatile performer.

In this study, Claude 3 Opus emerged as the top-performing model among three generative AI models. Additionally, Claude_3.5 Sonnet demonstrated consistent excellence across various medical specialties. As of August 2024, Claude_3.5 Sonnet stands out as the most advanced model developed by Anthropic (CA, USA). Despite having fewer studies in the medical domain compared to Ched GPT, Claude series has exhibited notable performance in several research endeavors. For example, Mensah et al.¹⁷ demonstrated Claude's superior performance in the application of Emergency Medicine when used with appropriate prompts. Uppalapati et al.¹⁸ reported that Claude AI performs better than ChatGPT and Bard because it gives complete responses. Similarly, Kurokawa et al.¹⁹ found that Claude 3 Opus showcased significantly enhanced diagnostic capabilities by incorporating key images alongside clinical history, facilitating the listing of important differential diagnoses. Additionally, Venerito et al.²⁰ successfully integrated Claude 2 into scoring systems for idiopathic inflammatory myopathies. However, some studies have indicated instances where Claude performed less effectively than GPT-4o in the medical field.²¹ Wu et al.²¹ demonstrated that widely used open-source language models (LLMs) exhibited poor zero-shot reasoning abilities in nephrology compared to GPT-4o and Claude 2, with Claude 2 scoring lower than GPT-4o. Furthermore, Abbas et al.²² found that GPT-4o's overall performance surpassed that of Claude by 15.3% on the National Board of Medical Examiners Sample Questions. A key reason for the differing outcomes in the studies by Wu et al.²¹ and Abbas et al.²² is that they utilized Claude 2, whereas most other studies demonstrating superior performance employed Claude 3 Opus.¹⁹ This variation underscores the distinct performance levels among AI models in the medical field. It emphasizes the importance of careful comparison of study results, especially as these generative AI models are rapidly evolving and revolutionizing the field.

It's important to note that Gemini Advanced models—demonstrated low performance in neurology, achieving an average accuracy of <50%. Despite previous research

indicating the potential of AI in this field, such as studies by Huang et al.²³ on debunking Alzheimer's disease myths among caregivers, Inojosa et al.²⁴ on assessing generative AI's clinical decision making in multiple sclerosis, and Kim et al.²⁵ on employing AI for the differential diagnosis of epilepsy, our latest findings reveal a noticeable discrepancy in the AIs' ability to handle neurology exam questions. These prior studies focused on specific neurological issues, testing the AI's proficiency within narrow topics.^{23–25} Our data suggest that a careful, comprehensive evaluation is required when deploying generative AI across broader neurological applications, ensuring the technology is adapted appropriately to meet diverse clinical needs.

Another point to note is that models like GPT-4o, Claude_3.5 Sonnet, and Gemini Advanced models, which are LLMs, demonstrate a discrepancy in handling text-based versus image-based questions.²⁶ LLMs are primarily trained using natural language processing (NLP) techniques focused on predicting and completing text, and their databases consist predominantly of textual data.^{27,28} This makes them highly adept at understanding and generating contextually appropriate text responses, a capability enhanced by their use of a deep learning architecture known as transformers, ideal for processing sequential data.^{26,29,30} However, LLMs do not inherently process images; they can only handle image-based questions through textual descriptions of the images, not by analyzing the images directly.³¹ The lack of direct image processing capabilities stems from the models' design, which does not incorporate the computational techniques—such as pattern recognition and object identification—used in image analysis.³² This specialization in text rather than images explains why LLMs have superior capabilities in answering text-based questions over image-based ones. The current advancement of multimodal LLMs is set to enhance the capability of generative AI in image interpretation, complementing its existing proficiency in text understanding.^{33,34}

Since the onset of the COVID-19 pandemic in late 2019, all three generative AI models examined in this study have shown the capability to effectively answer questions related to COVID-19. This demonstrates that these AI systems possess the ability to rapidly incorporate and update their knowledge bases with current and emerging information, even before 2021.³⁵ The swift adaptation and knowledge acquisition of these AI models are particularly significant in addressing fast-evolving global health crises like the COVID-19 pandemic. The capacity of AI to quickly integrate new data and respond to novel situations underscores its potential as a critical tool in medical education, especially in the context of advancing medical knowledge. Claude_3.5 Sonnet, in particular, stands out as a model that could play a pivotal role in helping medical professionals and students stay abreast of the latest developments in the field. This feature of AI models—to capture, process,

and disseminate new and important information rapidly—ensures that medical education can be conducted in a timely and relevant manner. The implications extend beyond merely answering pandemic-related questions; they suggest that AI can serve as an essential adjunct in the continuous education of healthcare professionals, particularly in areas where medical knowledge is rapidly evolving. As the landscape of medicine continually shifts with new discoveries and treatments, the ability to incorporate cutting-edge information quickly becomes a vital asset in both clinical practice and education. Future studies should explore how AI models like Claude_3.5 Sonnet can be further optimized and systematically integrated into medical curricula and ongoing professional development to maximize their utility in an ever-changing medical environment.

The study's findings open important discussions about the implications for both clinical practice and medical education. The demonstrated effectiveness of Claude_3.5 Sonnet suggests that integrating advanced AI tools into medical training could significantly enhance learners' diagnostic skills and understanding of complex medical scenarios. In clinical practice, the AI's ability to adeptly handle diverse question types may support physicians in making more informed decisions, ultimately leading to improved patient care. If we consider an average score above 90 as indicative of being qualified to assist in clinical decision making, our data suggests that Claude_3.5 Sonnet could be particularly valuable in addressing clinical problems within the subspecialties of Psychiatry, Hematology & Oncology, Infection, and Nephrology. This highlights the potential for AI to not only support general medical practice but also to provide specialized assistance in areas that require nuanced expertise.

While this study primarily focused on evaluating the performance metrics of AI models in the context of medical performance, it is essential to acknowledge that ethical considerations and potential biases could impact their real-world applicability. AI models like GPT-4o, Claude_3.5 Sonnet, and Gemini Advanced may carry biases rooted in their training data, which could affect their decision making and outcomes across different contexts and demographics. Additionally, it's important to note that the internal medicine exams used in this study contain very few questions related to ethics. Therefore, the AI models' performance on these exams may not translate to an accurate assessment of their ability to handle ethical issues. Addressing ethical considerations in AI is critical, but it exceeds the scope and data available in this study.

Based on the findings of this study, we recommend several specific avenues for future research. First, the integration of AI models into medical training should be approached systematically, with controlled studies designed to measure the impact of AI assistance on learning outcomes. AI models like Claude_3.5 Sonnet, which showed particular strengths in areas such as Psychiatry and Nephrology, could

be piloted in specialty-specific education programs to assess their effectiveness in enhancing diagnostic skills.

Second, AI models should be continuously evaluated for their ability to handle evolving medical knowledge, such as COVID-19-related content, to ensure they remain a valuable educational resource. Further studies could also investigate the potential for AI to support continuous medical education, helping clinicians stay updated on the latest advances in treatment and guidelines.

Third, improvements in AI's handling of visual data are necessary, as evidenced by Gemini Advanced's relatively poor performance on image-based questions. Research should focus on enhancing AI's image recognition capabilities, particularly in specialties like radiology and dermatology, where visual interpretation is crucial.

Finally, addressing ethical concerns and ensuring equitable AI deployment in diverse clinical settings will be critical. Future research should investigate strategies for minimizing algorithmic bias, particularly in subspecialties that involve complex patient populations, such as Oncology and Infectious Disease.

Limitations and potential biases

This study has several limitations, some of which may introduce potential biases that could have impacted the results. First, the definition of medical specialties was determined by the research team, which may differ from the official classifications recognized by the Taiwan Society of Internal Medicine. Although the distribution of questions across subspecialties closely mirrors the annual distribution in the Taiwan internal medicine exams, with discrepancies in the number of questions per subspecialty not exceeding two, this could still introduce a subtle bias. While we attempted to minimize the impact, the subjective nature of defining specialties could affect how well the AI models' performance generalizes to real-world internal medicine classifications.

A further limitation arises from the uneven distribution of questions across specialties. Certain specialties, such as Neurology and Psychiatry, had a very limited number of questions, which may not provide a robust assessment of the AI models' performance in these areas. The small sample size for these specialties could skew the models' scores and make it difficult to draw reliable conclusions about their true capabilities in such domains. This limitation underscores the challenge of adequately evaluating AI performance across a broad range of subspecialties, especially when question representation is uneven.

Another significant source of potential bias lies in the language processing of the three AI models. All the questions were written in Chinese, but when Gemini Advanced could not respond, the questions were translated into French using Google Translate. This additional translation step introduces the possibility of linguistic discrepancies,

misinterpretation of nuanced medical terminology, or even cultural bias, which could have skewed the performance outcomes. Translation inconsistencies may not only affect accuracy but could also alter the context or meaning of clinical scenarios, thereby affecting the AI's ability to generate appropriate responses.

Additionally, the limitations imposed by the AI systems themselves must be considered. Due to constraints on the number of questions generative AI models can process at one time, it was not possible to test all questions in a single session. This may introduce variability in the models' performance, as the different testing environments and potential changes in model behavior over time could contribute to inconsistencies in the results. Moreover, the limited capacity of the internal medicine question bank and the finite number of specialty-specific questions may not have fully captured the complexity and breadth of knowledge required to comprehensively assess the models' abilities across all medical fields. Lastly, the rapid evolution of AI technology introduces another layer of complexity. The models tested in this study are part of a fast-changing field, where new updates and improvements are frequently released. As this study predominantly relies on studies published in 2024 for comparison, the results may quickly become outdated as AI models continue to advance. Furthermore, the rapidly evolving capabilities of generative AI models could mean that the performance observed in this study might not be fully indicative of future versions, particularly as these models improve in processing language, adapting to medical terminology, and mitigating biases. In sum, while this study provides valuable insights into the performance of AI models in internal medicine exams, the inherent biases related to specialty definitions, question distribution, translation issues, and the evolving nature of AI technology must be carefully considered when interpreting the results. Future studies should address these limitations by ensuring larger and more balanced question sets across specialties, improving cross-linguistic accuracy in model evaluation, and continually updating assessments to reflect advancements in AI capabilities.

Conclusions

In conclusion, GPT-4o and Claude_3.5 Sonnet performed better on the Taiwanese internal medicine specialty exam questions compared to Gemini Advanced. Notably, GPT-4o, Claude_3.5 Sonnet, and Gemini Advanced each exhibited significant variations in performance across different subspecialties of internal medicine, indicating inconsistencies in their proficiency in the areas of internal medicine. These disparities may be attributed to the unique characteristics of the exam questions themselves or inherent differences in the generative AI models. It is hoped that the findings of this study will further aid in the application of generative AI in clinical settings.

Acknowledgements: We extend our sincere thanks to the Taiwan Society of Internal Medicine for providing the average passing rates of internal medicine fellows in each annual exam. Their support and data have been invaluable to this study.

Contributorship: The authors' individual contributions were as follows: S-YL and C-HK were responsible for the study design. S-YL, Y-YH, S-WJ, P-CY, W-HH, and CHK collected the data. All authors performed the statistical analyses, data interpretation, and article drafting, provided some intellectual content, approved this version of the manuscript for submission, and read and approved the final manuscript.


Data sharing statement: These considerations pertain to the protection of proprietary information inherent in the AI platforms' responses and the sensitivity of medical data. The study protocol adhered to the guidelines established by the journal. The protocol was also under the premise of safeguarding the confidentiality and integrity of the proprietary AI methodologies and ensuring the privacy of hypothetical patient data embedded in the biopsy reports.

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

Funding: The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study is supported in part by China Medical University Hospital (DMR-113-048, DMR-113-060, DMR-113-061, CMUH-DMR-112-187, CMUH-DMR-113-178, DMR-HHC-111-3 and CMUH-EDU11110), National Science and Technology Council (NSTC 112-2410-H-039-009-MY3).

Guarantor: C-HK

ORCID iDs: Ying-Yu Hsu  <https://orcid.org/0009-0006-9025-8102>

Chia-Hung Kao  <https://orcid.org/0000-0002-6368-3676>

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

References

1. Stokel-Walker C and Van Noorden R. What ChatGPT and generative AI mean for science. *Nature* 2023; 614: 214–216.
2. Reddy S. Generative AI in healthcare: an implementation science informed translational path on application, integration and governance. *Implement Sci* 2024; 19: 27.
3. Mesko B. The ChatGPT (generative artificial intelligence) revolution has made artificial intelligence approachable for medical professionals. *J Med Internet Res* 2023; 25: e48392.
4. Yu P, Xu H, Hu X, et al. Leveraging generative AI and large language models: a comprehensive roadmap for healthcare integration. *Healthcare (Basel)* 2023; 11: 2776.
5. Busch F, Hoffmann L, Rueger C, et al. Systematic review of large language models for patient care: current applications and challenges. *medRxiv* 2024. doi:10.1101/2024.03.04.24303733
6. Wang H, Wu W, Dou Z, et al. Performance and exploration of ChatGPT in medical examination, records and education in Chinese: pave the way for medical AI. *Int J Med Inform* 2023; 177: 105173.
7. Sumbal A, Sumbal R and Amir A. Can ChatGPT-3.5 pass a medical exam? A systematic review of ChatGPT's performance in academic testing. *J Med Educ Curric Dev* 2024; 11: 23821205241238641.
8. Gilson A, Safranek C, Huang T, et al. How does ChatGPT perform on the medical licensing exams? The implications of large language models for medical education and knowledge assessment. *MedRxiv* 2022. doi:10.1101/2022.12.23.22283901
9. Sallam M and Al-Salahat K. Below average ChatGPT performance in medical microbiology exam compared to university students. *Front Educ* 2023; 8: 1333415.
10. Lin S-Y, Chan PK, Hsu W-H, et al. Exploring the proficiency of ChatGPT-4: an evaluation of its performance in the Taiwan advanced medical licensing examination. *Digit Health* 2024; 10: 20552076241237678.
11. Jindal JA, Lungren MP and Shah NH. Ensuring useful adoption of generative artificial intelligence in healthcare. *J Am Med Inform Assoc* 2024; 31(6): 1441–1444.
12. Loscalzo J, Fauci AS, Kasper DL, et al. Harrison's principles of internal medicine. 21 ed. New York, NY: McGraw-Hill Education, 2022. Accessed May 08, 2024.
13. Palmer RD. COVID 19 vaccines and the misinterpretation of perceived side effects clarity on the safety of vaccines. *Biomedicine (Taipei)* 2022; 12: 1–4.
14. Introducing the next generation of Claude. <https://www.anthropic.com/news/claude-3-family>.
15. Gemini. <https://gemini.google.com/>.
16. Taiwan Society of Internal Medicine. <http://www.tsim.org.tw/>.
17. Mensah PB, Quao NS, Dagadu S, et al. Can large language models provide emergency medical help where there is no ambulance? A comparative study on large language model understanding of emergency medical scenarios in resource-constrained settings. *medRxiv* 2024. doi:10.1101/2024.04.17.24305971
18. Uppalapati V and Nag D. A comparative analysis of AI models in complex medical decision-making scenarios: evaluating ChatGPT, Claude AI, Bard, and Perplexity. *Cureus* 2024; 16: e52485.
19. Kurokawa R, Ohizumi Y, Kanzawa J, et al. Diagnostic performance of Claude 3 from patient history and key images in diagnosis please cases. *medRxiv* 2024. doi:10.1101/2024.04.11.24305622
20. Venerito V, Fornaro M, Sabbagh S, et al. Integrating large language models in medicine: a study of Claude 2's performance in MDAAT scoring for idiopathic inflammatory myopathies. *Rheumatology (Oxford)* 2024; 63(10): e292–e293.
21. Wu S, Koo M, Blum L, et al. Benchmarking open-source large language models, GPT-4o and Claude 2 on multiple-choice questions in nephrology. *NEJM AI* 2024; 1(2). doi:10.1056/AIdbp2300092

22. Abbas A, Rehman MS and Rehman SS. Comparing the performance of popular large language models on the national board of medical examiners sample questions. *Cureus* 2024; 16: e55991.
 23. Huang SS, Song Q, Beiting KJ, et al. Fact check: assessing the response of ChatGPT to Alzheimer's disease statements with varying degrees of misinformation. *Medrxiv* 2024. doi:10.1101/2023.09.04.23294917
 24. Inojosa H, Gilbert S, Kather JN, et al. Can ChatGPT explain it? Use of artificial intelligence in multiple sclerosis communication. *Neurol Res Pract* 2023; 5: 48.
 25. Kim H-W, Shin D-H, Kim J, et al. Assessing the performance of ChatGPT's responses to questions related to epilepsy: a cross-sectional study on natural language processing and medical information retrieval. *Seizure* 2024; 114: 1–8.
 26. Jin H, Zhang Y, Meng D, et al. A comprehensive survey on process-oriented automatic text summarization with exploration of LLM-based methods. *arXiv*. 2024; preprint arXiv:2403.02901.
 27. Huang Y, Tang K and Chen M. Leveraging large language models for enhanced NLP task performance through knowledge distillation and optimized training strategies. *arXiv*. 2024; preprint arXiv:2402.09282.
 28. Gao M, Hu X, Ruan J, et al. LLM-based NLG evaluation: current status and challenges. *arXiv*. 2024; preprint arXiv:2402.01383.
 29. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *arXiv*. 2017; arXiv:1706.03762.
 30. OpenAI. <https://www.openai.com/research/>.
 31. Ia G. Deep learning. An MIT press book. Ian Goodfellow and Yoshua Bengio and Aaron Courville, 2016. Cambridge, MA.
 32. Yun Y and Kim J. CIC: a framework for culturally-aware image captioning. *arXiv*. 2024; preprint arXiv:2402.05374.
 33. Lyu C, Wu M, Wang L, et al. Macaw-LLM: multi-modal language modeling with image, audio, video, and text integration. *arXiv*. 2023; preprint arXiv:2306.09093.
 34. Lang O, Yaya-Stupp D, Traynis I, et al. Using generative AI to investigate medical imagery models and datasets. *EBioMedicine* 2024; 102: 105075.
 35. Ciotti M, Ciccozzi M, Terrinoni A, et al. The COVID-19 pandemic. *Crit Rev Clin Lab Sci* 2020; 57: 365–388.
-