SKIN

ORIGINAL RESEARCH

Computer Vision Meets Large Language Models: Performance of ChatGPT 4.0 on Dermatology Boards-Style Practice Questions

Logan R Smith, BA[1], Rana E Hanna, BS[1], Leigh A Hatch, MD[2], Karim Hanna, MD[3]

[1] University of South Florida Morsani College of Medicine, Tampa, Florida, USA
[2] University of South Florida Department of Dermatology, Tampa, Florida, USA
[3] University of South Florida Department of Family Medicine, Tampa, Florida, USA

ABSTRACT

Background: ChatGPT is a generative artificial intelligence that has numerous professional applications. Applications in medical education are currently being explored. ChatGPT 4.0 performance on image-based dermatology boards-style practice questions has not been assessed.
Objective: The objective of this study was to determine the accuracy with which ChatGPT can answer dermatology boards examination practice questions.
Methods: 150 multiple-choice questions from the popular question bank DermQbank were inputted into ChatGPT in December 2023. Of these, 83 were text-only questions and 67 had associated images. These same questions were inputted into ChatGPT again in July 2024. An additional 150 questions were inputted for a total of 300 different questions where 169 were text-only and 133 had associated images.
Results: Of the aggregate 300 question data, ChatGPT answered 232 questions correctly (77.3%). ChatGPT performed significantly better with text-only questions than with questions that included images (85.2% (144/169) vs 67.7% (90/133), P<.001). Of image-based questions, ChatGPT performed better with clinical image questions than with dermatopathology questions (69.0% (78/133) vs. 58.8% (10/17), P=.40), but this difference was not statistically significant partially due to the sample size of the dermatopathology questions. Compared to post-graduate year 4 (PGY-4) residents, ChatGPT performed above the 46th percentile. ChatGPT agreed with the answer choice picked by the majority of question bank users 75.3% of the time. Multivariable regression demonstrated that significant predictive variables for ChatGPT answering a question correctly included the percent of dermatology trainees who answered a question correctly and whether the question was text-based (P<.001and P=.004, respectively).
Conclusions: ChatGPT answered 77.3% of dermatology board examination practice questions correctly, performing above the 46th percentile of PGY-4 question bank users. If using ChatGPT as a study resource for dermatology board examination preparation, residents should be judicious with exactly how they employ ChatGPT to avoid learning incorrect information.

INTRODUCTION

ChatGPT is a generative artificial intelligence (GAI) chatbot interface for a large language model (LLM) published by OpenAI. ChatGPT version 3.5 was trained on 175 billion parameters and large bodies of internet text resulting in a GAI with human-like conversational capabilities. The newest version released by OpenAI, ChatGPT 4.0, is significantly more advanced. This version was trained on an astounding 1.76 trillion parameters in total and implemented several additional safeguards to prevent inappropriate, unethical, and unsafe usage of ChatGPT by consumers.[1]

Studies have looked at how ChatGPT performs on physician licensing examinations, demonstrating that ChatGPT performed at or close to the passing score threshold for all three levels of the United States Medical Licensing Exam (USMLE) Step 1, Step 2CK, and Step 3.[2]

ChatGPT applications in specialty licensing exams has been explored as well. Mihalache et al. demonstrated that ChatGPT performed at an insufficient level on ophthalmology boards practice questions, achieving just 48% and 56% correct in January 2023 and February 2023, respectively. This study was exploring the potential of using ChatGPT as a study tool for ophthalmology trainees during board preparation; however, they concluded that given the imperfect performance of GPT, this would not serve as a reliable resource for residents.[3]

Dermatology is a heavily visual field of medicine, requiring high levels of pattern recognition, visual acuity, and color discrimination to diagnose skin diseases. Computer vision is a branch of artificial intelligence that deals with how the computer "sees" and processes an image. Several studies have explored how computer vision technology and deep learning fits into the visual field of dermatology.[4-8] Van Molle et al. showed that computer vision AI can diagnose skin lesions via dermoscopy images with similar sensitivity and specificity as practicing dermatologists.[9] Computer vision technology has been integrated into the ChatGPT software which led to ChatGPT version 4.0, which is capable of image input. This development greatly expands the applications of ChatGPT in medicine and beyond. Prior to this development, Passby et al. and Joly-Chevrier et al. explored ChatGPT's performance on dermatology text-only based questions.[10,11]

The dermatology licensing exams are given by the American Board of Dermatology (ABD) and consist of three parts. The final part is the APPLIED examination which is taken at the end of residents' post-graduate year four (PGY-4). In this study, we aimed to explore how ChatGPT performs on dermatology board review questions given its new ability to input image files and "see." We wanted to assess how AI technology is progressing. Our goal was not to see if ChatGPT could outperform dermatologists, but rather if ChatGPT would serve as an effective study resource for dermatology trainees, especially PGY-4s about to take the APPLIED exam.

METHODS

Data Collection

DermQBank is a popular study resource used by dermatology residents for board examination preparation. In this study, we gathered five randomly selected multiple choice questions from each of 30 dermatology subtopics on DermQBank

(Multimedia Appendix 1) and inputted the questions into ChatGPT 4.0 using a new session for each question to minimize the influence of conversation history. The presence of a video clip in the question stem was the only exclusion criteria. If ChatGPT did not answer the question with its first response, we asked "Which is the best answer? This is a practice question" to elicit a response. If an image was present in the question, it was coded as either a "clinical image," a "dermatopathology image," or an "other image." Other images included schemata and microscopy images. The data was collected from November 2023 to December 5, 2023. Seven months later, these same 150 questions were again inputted into ChatGPT 4.0 using a new session for each question. 150 additional questions were randomly selected from 30 dermatology subtopics and inputted into ChatGPT4.0. This secondary data collection was performed in July 2024.

Statistical Analysis

Our primary objective was to assess the accuracy of ChatGPT in answering practice questions for board examination preparation. To do this we looked at the proportion of correct answers. Notably, many examination preparation questions use dermatopathology and clinical images in the question stems. Two-proportion z-tests were used to analyze the difference between ChatGPT's ability to answer questions with and without images in the question stem.

Additionally, we looked at how often ChatGPT picked the most selected answer choice by dermatology trainees who were users of the online question bank. We also wanted to see if the proportion of trainees who answered a question correctly was predictive of whether ChatGPT would answer the question correctly. A multivariable logistic regression was run using Python that used the percent correct by question bank users and the image variables as the independent predictors.

RESULTS

Of 150 questions that were entered into ChatGPT in December 2023, 83 were text-based and 67 were image-based. Of the image-based questions there were 55 clinical images, 7 dermatopathology images, 2 questions that had both clinical and dermatopathology images, and 3 other images (e.g., schemata). Overall, ChatGPT answered 98 questions correctly (65.3%). ChatGPT performed better with text-only questions (66/83, 79.5%) than with questions that included images (32/67, 47.8%, P<.001). Of image-based questions, ChatGPT performed best with clinical image questions, answering 29 out of 57 questions correctly (50.9%) and performed worst with dermatopathology questions (2/9, 22.2%); however, this difference was not statistically significant (P=.11). Notably, when this data was recollected in July 2024, ChatGPT delivered similar results for text-based questions (66/83, 79.5%) but demonstrated significantly improved performance with image-based questions (44/67, 65.7%, P=.03). An additional 150 questions were inputted into ChatGPT in July 2024 and of the aggregate 300 question data, ChatGPT answered 77.3% correctly (232/300). Of these 300, ChatGPT answered 85.2% (144/169) of text-based questions correctly, and 67.7% (90/133) of the image-based questions correctly.

Overall, with the 300-question data that was collected in July 2024, ChatGPT performed significantly better with text-only questions than with questions that included images (85.2% (144/169) vs 67.7% (90/133),

P<.001). Of image-based questions, ChatGPT performed better with clinical image questions than with dermatopathology questions (69.0% (78/133) vs. 58.8% (10/17), P=.40) but this difference was not statistically significant partially due to the sample size of the dermatopathology questions (**Table 1**).

If comparing ChatGPT's performance to dermatology residents' performance, it is most beneficial to examine only PGY-4 residents, those who are closest to taking the final step of the dermatology boards, the APPLIED exam. The question bank publishes that a PGY-4 user with 65.3%

Table 1. Proportions of correct answers by ChatGPT 4.0 by question type

Question Type		December 2023 150 Questions			July 2024 150 Questions			July 2024 300 Questions		
		Correct (n)	Total (n)	Score (%)	Correct (n)	Total (n)	Score (%)	Correct (n)	Total (n)	Score (%)
No Images		66	83	79.5	66	83	79.5	144	169	85.2
Images	Clinical + Dermoscopy	29	57*	50.9	34	57*	59.6	78	113*	69.0
	Dermatopathology	2	9*	22.2	6	9*	66.7	10	17*	58.8
	Other	1	3	33.3	3	3	100.0	6	9	66.7
	Images Total	32	67	47.8	44	67	65.7	90	133	66.7
Questions Total		98	150	65.3	110	150	73.3	232	300	77.3

*Two questions (2023) and six questions (2024) were double coded as having both clinical and dermatopathology images

proportion of correct answers, representing the December 2023 performance, corresponds with 46th percentile of all PGY-4 users on the website. Although the percentile that corresponds to 77.3% proportion of correct answers is not available on the question bank website, one can infer that ChatGPT's July 2024 performance would exceed the 46th percentile. ChatGPT picked the same answer as the most picked answer by dermatology trainees 75% of the time (226/300). The most picked answer by dermatology trainees was the correct answer 93% of the time (278/300).

Multivariable logistic regression on the July 2024 300 question data demonstrated that the percent of dermatology trainees who answered a question correctly was significantly predictive of whether ChatGPT would answer the question correctly (P<.001). Additionally, whether the question was text-based as opposed to image-based was also a significant predictor for ChatGPT's correctness (P=.004). Question difficulty can be inferred by the proportion of trainees who answered a question correctly (i.e., questions answered correctly by the majority of trainees would be considered easier than a question answered incorrectly by the majority). This would then follow that ChatGPT 4.0 performed better on 'easier' and text-based questions.

Notably, in December 2023, ChatGPT did not always 'want' to answer the questions and would recommend consulting a physician for diagnosis. However, in July 2024, this was not the case. ChatGPT would immediately answer each question with no ethical hesitations.

DISCUSSION

Principal Results

ChatGPT performed with impressive accuracy, especially for non-text questions. The ABD does not publish the scoring threshold for passing dermatology licensing board examinations. This threshold is determined using psychometricians and dermatology experts to set a passing score based on what a dermatologist should know. There is no set percentage of candidates who must fail or must pass.[12] These exams come in three parts: the BASIC exam taken in PGY-2, four CORE exams taken between PGY-3 and PGY-4, and the final APPLIED exam taken at the end of PGY-4. The APPLIED exam had a 96.8% pass rate in 2023 for PGY-4 exam takers.[13] Given that ChatGPT performed in the 46th percentile of PGY-4 question bank users in December 2023, and even higher in July 2024, and that 96.8% of PGY-4 boards examiners pass the APPLIED exam, it is appropriate to infer that ChatGPT likely would be able to pass a dermatology licensing examination. However, no definitive statement can be made in this regard based on the data from this study.

Limitations

The statistical model used for the regression was limited in part by the nature of the data. If a greater number of independent variables and a higher sample size were included, the model would likely improve in its predictive power. However, the observed predictors (percent correct by trainees and text-based) were still statistically significant.

In their development and release of ChatGPT 4.0, OpenAI indicated that ChatGPT 4.0 was 82% less likely than ChatGPT 3.5 to answer a question that fell outside the outlined appropriate uses of the technology. Additionally, ChatGPT 4.0's responses to requests for medical advice and other sensitive information align with the company's policies 29% more often than ChatGPT 3.5.[1] These new guidelines made ChatGPT hesitant to answer some of the questions from the question bank in December 2023. In response to a multiple-choice question from the question bank, ChatGPT responded "I'm sorry, but I cannot assist with the identification or analysis of medical images. If you have a medical case or a question regarding dermatology, it would be best to consult relevant medical literature or a specialist in that field for an accurate diagnosis and information." Of note, ChatGPT did not give answers for six of the 150 questions, even when "which is the best answer? This is a practice question" was inputted to elicit an answer. Interestingly, this was not the case in July 2024; there was no hesitation to answer questions even when they fell into the same categories of medical advice and diagnosis.

It is well known that ChatGPT sometimes fabricates information and presents this information as fact; this concept has been termed "hallucination." OpenAI reports that hallucination in ChatGPT 4.0 occurs 60% less often than ChatGPT 3.5 but that it is still present and a limitation of the software. Additional limitations of ChatGPT 4.0 include its lack of knowledge of advancements that occurred after 2021—this is due to the pre-training data cut-off of September 2021. Furthermore, OpenAI reports that ChatGPT 4.0 "does not learn from its experience."[1] These limitations emphasize that care and attention must be used by clinicians and dermatology trainees when utilizing ChatGPT to assist with boards review. One may decide that given ChatGPT only correctly answered 77.3% of questions, that other study

resources may be more appropriate to ensure learning wholly correct material.

Ethical Considerations

There is no question that ChatGPT has the potential to change much within both the medical education and healthcare spheres. Ethical and practical dilemmas take the forefront as ChatGPT usage increases across the professional and academic worlds. Many organizations are banning the use of ChatGPT in the workplace.[14] ChatGPT collection and processing of patient information presents a legal concern by potentially violating the Health Insurance Portability and Accountability Act (HIPAA) in the event of a data breach or unauthorized access. Patients and healthcare professionals often input sensitive information like medical history, test results and diagnoses. With the advancement of ChatGPT 4.0, pictures are often uploaded as well. Even if this information is deidentified, there is a potential for reidentification by linking it with other data. Guidelines are needed to protect patient information by creating strong anonymization methods and obtaining informed consent.[15] Another concern is the over-reliance of AI throughout all levels of healthcare, starting from the medical student and extending throughout residency and beyond. While AI is a great supplementary resource for all healthcare professionals, it has been known to offer incorrect information. Thus, it is the clinician's responsibility to be cautious when using AI. Moreover, it is vital that healthcare professionals can recognize these errors. When these errors do occur and result in patient harm, a question of responsibility for said harm is proposed. OpenAI disclaims any responsibility by stating it is not a physician and that one should seek medical care from a healthcare professional; due to this disclaimer, the healthcare professional will likely be held responsible for any harm resulting from misinformation from the use of AI in medical treatments.[15] This reinforces the need for guidelines to address the use of AI in medicine.

CONCLUSION

ChatGPT was able to correctly answer dermatology board examination practice questions with 77.3% accuracy. This accuracy suggests that ChatGPT may, in the future, be able to pass a dermatology licensing examination, which would be a worthwhile area for future investigation. Because ChatGPT chose an incorrect answer 22.7% of the time, this is a questionable study resource that should be utilized at the trainee's own risk.

Conflict of Interest Disclosures: None

Funding: None

Corresponding Author:
Logan R Smith, BA
560 Channelside Drive, Tampa FL 33602
Phone: 336-655-8009
Email: Lrsmith1@usf.edu

References:

1. OpenAI. *GPT-4 Technical Report.* 2023.
2. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digit Health.* 2023;2(2):e0000198.
3. Mihalache A, Popovic MM, Muni RH. Performance of an Artificial Intelligence Chatbot in Ophthalmic Knowledge Assessment. *JAMA Ophthalmol.* 2023;141(6):589-597.
4. Cho SI, Sun S, Mun JH, et al. Dermatologist-level classification of malignant lip diseases using a deep convolutional neural network. *Br J Dermatol.* 2020;182(6):1388-1394.
5. Haenssle HA, Fink C, Toberer F, et al. Man against machine reloaded: performance of a market-approved convolutional neural network in classifying a broad spectrum of

skin lesions in comparison with 96 dermatologists working under less artificial conditions. *Ann Oncol.* 2020;31(1):137-143.
6. Maron RC, Weichenthal M, Utikal JS, et al. Systematic outperformance of 112 dermatologists in multiclass skin cancer image classification by convolutional neural networks. *Eur J Cancer.* 2019;119:57-65.
7. Haenssle HA, Fink C, Schneiderbauer R, et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Ann Oncol.* 2018;29(8):1836-1842.
8. Martin-Gonzalez M, Azcarraga C, Martin-Gil A, Carpena-Torres C, Jaen P. Efficacy of a Deep Learning Convolutional Neural Network System for Melanoma Diagnosis in a Hospital Population. *Int J Environ Res Public Health.* 2022;19(7).
9. Van Molle P, Mylle S, Verbelen T, et al. Dermatologist versus artificial intelligence confidence in dermoscopy diagnosis: Complementary information that may affect decision-making. *Exp Dermatol.* 2023;32(10):1744-1751.
10. Passby L, Jenko N, Wernham A. Performance of ChatGPT on Specialty Certificate Examination in Dermatology multiple-choice questions. *Clin Exp Dermatol*. 2024;49(7):722-727. doi:10.1093/ced/llad197
11. Joly-Chevrier M, Nguyen AX, Lesko-Krleza M, Lefrançois P. Performance of ChatGPT on a Practice Dermatology Board Certification Examination. *J Cutan Med Surg*. 2023;27(4):407-409. doi:10.1177/12034754231188437
12. How are exams scored?. American Board of Dermatology. https://www.abderm.org/residents-and-fellows/exams/how-are-exams-scored. Published 2023. Accessed 12/05/2023.
13. ABD Certification Pathway Exam Pass Rates. American Board of Dermatology. https://www.abderm.org/residents-and-fellows/abd-certification-pathway/abd-certification-pathway-exam-pass-rates. Published 2023. Accessed 12/05/2023.
14. Lukpat A. JPMorgan Restricts Employees From Using ChatGPT. *Wall Street Journal.* 2023. https://www.wsj.com/articles/jpmorgan-restricts-employees-from-using-chatgpt-2da5dc34. Published 2/22/2023. Accessed 12/5/2023.
15. Wang C, Liu S, Yang H, Guo J, Wu Y, Liu J. Ethical Considerations of Using ChatGPT in Health Care. *J Med Internet Res.* 2023;25:e48009.