

ChatGPT-4's Level of Dermatological Knowledge Based on Board Exam Review Questions and Bloom's Taxonomy

Hansen Tai, Carrie Kovarik

Submitted to: JMIR Dermatology
on: March 17, 2025

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript.....	4
---------------------------------	----------

Preprint
JMIR Publications

ChatGPT-4's Level of Dermatological Knowledge Based on Board Exam Review Questions and Bloom's Taxonomy

Hansen Tai¹ BA; Carrie Kovarik² MD

¹ SUNY Upstate Medical University Syracuse US

² Professor University of Pennsylvania Philadelphia US

Corresponding Author:

Carrie Kovarik MD

Professor

University of Pennsylvania

3600 Spruce Street, 2 Maloney Building

Philadelphia

US

Abstract

Our study demonstrated the ability of ChatGPT-4 to answer 77.5% of all sampled text-based board review type questions correctly. Questions requiring recall of factual information were answered correctly most often, with slight decreases in correctness as higher-order thinking requirements increased. Improvements to ChatGPT's visual diagnostics capabilities will be required before it can be used reliably for clinical decision-making and visual diagnostics

(JMIR Preprints 17/03/2025:74085)

DOI: <https://doi.org/10.2196/preprints.74085>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to all users.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in a JMIR journal, I will be able to make the full manuscript available to all users.

No. Please do not make my accepted manuscript PDF available to anyone.

Original Manuscript

ChatGPT-4's Level of Dermatological Knowledge Based on Board Exam Review Questions and Bloom's Taxonomy

Manuscript word count: 750

Tables: 2

Names of authors: Hansen Tai, BA¹ and Carrie Kovarik, MD²

Names of the institutions: SUNY Upstate Medical University¹ and Department of Dermatology at the Hospital of the University of Pennsylvania²

Name and email address of the corresponding author: Carrie Kovarik, MD, carrie.kovarik@pennmedicine.upenn.edu

A statement of all funding sources that supported the work: None

Any conflict of interest disclosures: No conflicts of interest

Details of ethics approval and informed consent, if applicable: None

Introduction

ChatGPT, a multimodal language model capable of answering multiple choice questions, incorporates visual inputs in its latest version, Generative Pre-Trained Transformer 4 (GPT4). Lewandowski et al recently assessed ChatGPT-3.5 and ChatGPT-4's performance on dermatology exams, finding ChatGPT-4 significantly outperformed its predecessor, achieving over a 60% pass rate overall and >84% accuracy on photo-based questions.¹ Building on this, our study classified ChatGPT-4's correctly answered question types using Bloom's taxonomy for cognitive complexity.²

Methods

We evaluated ChatGPT4's capabilities on the Basic, Core, and Applied exam questions from Dermatology-In-Review, an online dermatology board review preparation course. The Basic exam is a required exam for first-year US dermatology residents and tests dermatology fundamental knowledge. The Core and Applied examinations are taken later in residency and after residency, respectively, and examine more advanced clinical knowledge and focus more on higher-order thinking. In total, 167 Basic, 210 Core, and 166 Applied multiple-choice questions without photos were formatted and fed into GPT-4 using an algorithm in Python's Pandas. GPT-4's in-depth responses to each query were captured, reviewed, and coded as correct or incorrect (Table 1).

Additionally, we categorized text-based questions according to Bloom's taxonomy using a Python function. Bloom's categories included Remember (includes lower-level thinking, such as knowledge and comprehension), Apply, Analyze, Evaluate, and Synthesize. The results are compiled in Table 2. All statistics were performed using R statistical software, including Pearson's Chi-squared test (Table 1) and Fisher's Exact Test (Table 2).

Photo-based questions were entered directly into Chat-GPT4, along with structured messages and answer choices, and responses were recorded. Fifty-three photo cases from all board categories were used.

Results

Overall, ChatGPT answered 77.5% of all sampled text-based questions correctly. Varying levels of accuracy were demonstrated in answering board questions within different Bloom categories. In the "Remember" category, the model correctly answered 158/193 (81.9%). "Remember" is considered the most basic level of educational understanding, with the ability to recall or comprehend information without applying the concept.² Chat-GPT4 performed the best in this category; however, it did significantly ($p=0.0382$) better on the "Remember" questions from the Basic exam compared to those on the Core and Applied sections combined (Table 1). As the Bloom categories progress from Apply to Analyze, Evaluate, and Synthesize, a solid foundation of knowledge and higher-order thinking is necessary. Table 2 demonstrates a decreasing trend ($p=0.059$) in the percent correctness for the Chat-GPT4 answers moving from "Remember" to the classes of higher-order thinking.

Eighteen out of 53 (34%) of the questions with photos were answered correctly, with none of the "What is the histologic diagnosis?" question stems answered correctly. Excluding these, 18/38 (47.3%) had the correct answer. Notably, photo questions with leading information were more likely to be given the correct response.

Discussion

ChatGPT4 correctly answered between 77.5% of all text questions correctly, which is similar to the study by Lewandowski et al., in which ChatGPT4 answered 80.7-84% of the questions correctly on English-based Dermatology assessments. Our outcomes differed in that they were able to obtain a much higher number of correct responses on photo-based questions compared to our study, where ChatGPT4 was only able to answer approximately one-third of the photo cases correctly. A recent investigation by Hirosawa et al. assessed the impact of adding image data to clinical textual data on

ChatGPT-4's diagnostic accuracy.³ They found that integrating image data into ChatGPT-4 did not significantly enhance diagnostic accuracy, and it predominantly relies on textual data, limiting its ability to use the diagnostic potential of visual information fully.³ This corroborates our findings of poor analysis of photo cases and improved correctness when leading question stems were given. In addition, another recent study explored ChatGPT4's ability to answer psychosomatic clinical exam questions and group the responses into Bloom categories.⁴ This study noted that ChatGPT had more problems with the "Remember" category; however, these questions were long, detailed clinical scenarios, which are quite different from the question types presented in the dermatology exams.

Overall, our study demonstrates the ability of ChatGPT4 to answer text-based questions from Dermatology-In-Review at a high level. Questions requiring recall of factual information were answered correctly most often, with slight decreases in correctness as higher-order thinking requirements increased. Improvements to ChatGPT's visual diagnostics capabilities will be required before it can be used reliably for visual interpretation and clinical decision-making.

Our study has several limitations. Bloom's taxonomy is a continuum, and question classification can be complex. We used board review questions, and this may not be generalizable to true board questions. The current edition of ChatGPT has been trained with data only up to December 2023.⁵

References

1. Lewandowski M, Łukowicz P, Świetlik D, Barańska-Rybak W. ChatGPT-3.5 and ChatGPT-4 dermatological knowledge level based on the Specialty Certificate Examination in Dermatology. *Clinical and Experimental Dermatology*. Published online August 4,

2023:llad255. doi:10.1093/ced/llad255

2. Zaidi, N. Modified Bloom's Taxonomy for Evaluating Multiple Choice Questions. 15 June 2015, www.bcm.edu/sites/default/files/2019/04/principles-and-guidelines-for-assessments-6.15.15.pdf. Accessed May 21, 2024.
3. Hirosawa T, Harada Y, Tokumasu K, Ito T, Suzuki T, Shimizu T. Evaluating ChatGPT-4's Diagnostic Accuracy: Impact of Visual Data Integration. JMIR Med Inform. 2024 Apr 9;12:e55627
4. Herrmann-Werner A, Festl-Wietek T, Nat R; Holderried F, Herschbach L, Griewatz J, Masters K, Zipfel S, Mahling M. Assessing ChatGPT's Mastery of Bloom's Taxonomy Using Psychosomatic Medicine Exam Questions: Mixed-Methods Study. J Med Internet Res 2024;26:e52113.
5. OpenAI Platform. Accessed April 27, 2024. <https://platform.openai.com>

Table 1: Chat-GPT 4 Cases Correct by Testing Category*

	Correct	Incorrect	% Correct	Remember Type Questions: Correct %, Total %
Basic	139	28	83.20%	71/82 (86.6%), 82/167 (49.1%)
Core	158	52	75.20%	52/66 (78.8%), 66/210 (31.4%)
Applied	123	43	74.10%	35/46 (76.1%), 46/166 (27.7%)

*p-value = 0.0382, Pearson's Chi-squared test for Basic vs Core+Applied

Table 2: Chat-GPT 4 Cases Correct by Bloom Category (All cases)*

	Correct	Incorrect	Total	% Correct
Remember	158	35	193	81.9%
Apply	168	51	219	76.7%
Analyze	56	19	75	74.7%
Evaluate	37	14	52	72.5%
Synthesize	1	3	4	25.0%
Total	420	122	542	77.5%

* p-value = 0.059, Fisher's Exact Test