Assessing ChatGPT-4's Capabilities in Generating Dermatology Board Examination Content: An Explorational Study

Jonathan SHAPIRO[1], Anna LYAKHOVITSKY[2,3] Tamar FREUD[4], Felix PAVLOTSKY[2,3,] Ziad KHAMAYSI[5,6], Yulia VALDMAN-GRINSHPOUN[7], Roni DODIUK-GAD[5,8,9], Ilan GOLDBERG[3,10], Arieh INGBER[11], Baruch KAPLAN[12] and Emily AVITAN-HERSH[5,6]

[1]Maccabi Healthcare Services, Tel Aviv-Yafo, Israel, [2]Department of Dermatology, Sheba Medical Center, Tel Hashomer, Ramat Gan, Israel, [3]School of Medicine, Faculty of Medical and Health Sciences, Tel Aviv University, Tel Aviv, Israel, [4]Ben-Gurion University of the Negev, Beer-Sheva, Israel, [5]Department of Dermatology, Rambam Health Care Campus, Haifa, Israel, [6]Technion Faculty of Medicine, Haifa, Israel, [7]Soroka Medical University Center, Beer-Sheva, Israel, [8]Department of Dermatology, Emek Medical Center, Afula, Israel, [9]Division of Dermatology, Department of Medicine, University of Toronto, Canada, [10]Division of Dermatology, Tel-Aviv Sourasky Medical Center, Tel-Aviv, Israel, [11]Hadassah Medical Center, Jerusalem, Israel, and [12]Adelson School of Medicine, Ariel University, Ariel, Israel. E-mail: jonmidi@gmail.com

Submitted Jul 20, 2024. Accepted after revision Oct 28, 2024.
Published Jan 3, 2025 DOI: 10.2340/actadv.v105.41208. Acta Derm Venereol 2025; 105: adv41208.

The Chat Generative Pre-Trained Transformer (ChatGPT) series is pivotal in natural language and image processing (1). ChatGPT has shown near-passing results in medical licensing exams, including dermatology (2–4). An assessment of ChatGPT-3.5 for the American Board of Dermatology Applied Exam found 40% of its questions accurate and suitably complex (5). ChatGPT-4 advances further with improved linguistic processing, deeper subject understanding, and a broader knowledge base, potentially improving its question-generation capability.

The Israeli Dermatology Board exam preparation involves a multi-stage process. Based on the textbook "Dermatology, 4th Edition", by Bolognia et al. (6), committee members create 150 multiple-choice questions, each based on a different chapter. The chair reviews these questions for accuracy and structure. The committee then discusses each question and stratifies questions by difficulty. Key rules include having one correct answer, avoiding "all of the above", "none of the above", and double negatives, and ensuring answers are syllabus-based. The exam also features complex clinical cases requiring diagnoses based on descriptions and images, and the questions relate to different clinical or laboratory characteristics of the diagnosis.

This study assesses the effectiveness of ChatGPT-4 in producing accurate and contextually relevant examination content for dermatology board exams.

MATERIALS AND METHODS

Twelve thematic areas were randomly chosen from the textbook "Dermatology, 4th Edition", by Jean L. Bolognia, Julie V. Schaffer, and Lorenzo Cerroni (6). The text of each specific chapter was copied into a Word document and securely imported into the paid version of ChatGPT-4, which was commercially available between 27 December 2023, and 3 January 2024. The "Chat & History Training" parameter in ChatGPT-4's data control settings was disabled to prevent the data from being used for training or stored on its servers. Chats were automatically deleted upon completion, with no option for recovery. Subsequently, the model was tasked to generate multiple-choice questions. The prompt was refined after a systematic process of trial and error and is detailed in Appendix S1. The following final prompt version was consistently used for all the subjects: "Based only on the Word document I uploaded, ask extremely hard complicated, and very diverse questions including regular and clinical questions and a two-step thought process and provide the answer after every question and write at what page in the document I uploaded I can find the answer. If the question requires a two-step thought process where the physician must first deduce the diagnosis from the clinical presentation before answering the specific question, don't mention the diagnosis in the questions and add the diagnosis to the answer in a separate line. The questions should be multiple choice numbered questions.". The prompt and the questions were both in English.

Eight board-certified dermatology experts reviewed the questions. Of those, 5 (FP, YVG, IG, AI, and EAH) are long-term members of the Israeli board exam committee (10, 4, 8, 8, and 7 years, respectively). Two authors chaired the committee (FP, AI) and 1 is the current chair (EAH).

Each questionnaire was assessed by 2 reviewers, of which at least 1 was a long-term committee member. All questions were evaluated as "Suitable", which were further graded by difficulty, or "Not Suitable", which were categorized based on the reason. In cases of disagreement, mutual consultations were aimed at reconciling differences in scoring. Reviewers also recorded the time spent reviewing each exam and estimated how long it would have taken to write the same number of appropriate questions.

Statistical analysis

Statistical analysis was primarily descriptive. Categorical variables were presented as frequency and percentage. Inter-rater reliability was calculated utilizing Cohen's Kappa. All analyses were performed with IBM SPSS statistic software version 29.0 (IBM Corp, Armonk, NY, USA). $P<0.05$ was chosen as the significance level.

RESULTS

ChatGPT-4 generated 402 questions, with 208 (51.7%) deemed acceptable by at least 1 reviewer. However, only 72 questions (18%) were accepted by both reviewers. After consensus discussions, 53 of the 136 initially disputed questions were approved, resulting in a total of 125 questions deemed suitable for the exam. The suitable questions were classified as 51 (40.8%) easy, 45 (36%) medium-difficulty, and 29 (23.2%) hard. The main issues with unsuitable questions included questions that contained errors or improperly structured or with potential for an appeal (118 questions, 27.8%) and excessive simplicity (113 questions, 28.1%).

Table I provides a breakdown of the generated questions by subject area. Biopsy techniques and B-cell lymphoma had the highest rates of suitable questions (63–65%). In

Table I. Overview of question suitability and rejection reasons by subject

Subject	NOQ	Suitable *n* (%)	Dispute *n* (%)	R1 *n* (%)	R2 *n* (%)	Reason for rejection
Biopsy techniques	20	13 (65.0)	11 (55.0)	8 (40.0)	17 (85.0)	To easy
CBCL	30	19 (63.3)	10 (33.3)	21 (70.0)	19 (63.3)	Errored
CTCL	30	16 (53.3)	12 (40.0)	10 (33.3)	16 (53.3)	To easy
HPV	32	13 (40.6)	8 (25.0)	9 (28.1)	15 (46.9)	To hard
Alopecia	40	16 (40.0)	21 (52.5)	7 (17.5)	22 (55.0)	To easy
Systemic disease	30	10 (33.3)	11 (36.7)	8 (26.7)	15 (50.0)	To easy
Mycobacteria	30	8 (26.7)	15 (50.0)	15 (50.0)	8 (26.7)	To easy
Darier disease	20	5 (25.0)	10 (50.0)	4 (20.0)	12 (60.0)	Errored
Ichthyoses	40	6 (15.0)	8 (20.0)	10 (25.0)	6 (15.0)	Errored
Acne	60	9 (15.0)	17 (28.3)	21 (21.0)	10 (16.7)	To easy
Rosacea	20	3 (15.0)	8 (40.0)	9 (45.0)	9 (45.0)	Errored
Vasculitis	50	7 (14.0)	5 (10.0)	10 (20.0)	5 (10.0)	To easy
Total	402	125 (31.1)	136 (33.8)	132 (32.8)	154 (38.3)	

NOQ: number of questions: total questions evaluated per subject. Suitable: questions deemed appropriate for use. Dispute: questions with disagreements between reviewers. R1, R2: Number of suitable questions as determined by Reviewer 1 and Reviewer 2, respectively. Reason for rejection: primary reason for rejecting questions.% Suitable questions: proportion of questions considered suitable in each subject category.

addition, 37 questions were 2-stage complicated questions. Of those 7 were determined as appropriate (18.9%).

In 19 of the 24 assessments, the reviewers acknowledged that using ChatGPT-4 could potentially reduce the time needed by up to 55 min per question (range –110 to –55). **Table II** presents the time invested to review each subject and the estimated duration for designing suitable questions. Most reviewers rated the platform as useful and exhibited their willingness to employ it in the future.

In our cohort, the inter-rater reliability was low, indicating a generally low level of agreement before consensus (**Table III**). The Kappa values were highest in the vasculitis and HPV chapters. Of the 136 disputes, 55 (40.4%) arose from 1 reviewer finding the question too easy, while 48 (35.3%) involved errors or poor structure.

DISCUSSION

ChatGPT has gained significant popularity for its natural language processing and content generation capabilities. Given the complexity of structuring board exams, which demands consistency, proper question structure, and a balanced mix of difficulty levels and clinical scenarios, we explored ChatGPT-4's ability to generate suitable multiple-choice questions for dermatology board exams. This study extends previous work with ChatGPT-3.5 (5, 7–10) by increasing the number of questions and incorporating two-step reasoning tasks, such as diagnostic deductions and follow-up actions (e.g., "What would be your next step?'), to evaluate the model's performance comprehensively.

Table II. Efficiency of AI in Dermatology Board Exam question generation

Subject	NOQ	R1 time (min)	Expected/ question (min)	R2 time (min)	Expected/ question (min)	Saved time (min)±SD
Biopsy techniques	13	2.7	16.3	2.7	7.05	9.0±4.6
CBCL	19	1.4	57.1	2.3	9.47	31.5±24.3
CTCL	16	3	60	3.75	11.25	32.3±24.8
HPV	13	3.3	10	2	12	8.4±1.7
Alopecia	16	5.14	17.1	2.4	13.6	11.6±0.4
Systemic disease	10	3.75	30	2	20	22.1±4.1
Mycobacteria	8	2.4	12	7.5	22.5	12.3±2.7
Darier disease	5	45	30	1.25	3	–6.6±8.4
Ichthyoses	6	7.5	12	45	3	–18.8±23.3
Acne	9	2.8	8.6	1	48	26.4±20.6
Rosacea	3	2.2	20	100	20	–31.1±48.9
Vasculitis	7	6	48	128	18	–34.0±76.0

NOQ: Number of suitable questions: total suitable questions identified for each subject. R1 time (min): average time, in minutes, Reviewer 1 spent evaluating each AI-generated question. Expected time/Question (min): estimated average time, in minutes, that a reviewer would typically take to manually create a suitable question for a specific subject. R2 time (min): average time, in minutes, Reviewer 2 spent evaluating each AI-generated question. Saved time (min)±SD: average time saved per question, in minutes, by using AI-generated questions compared with manually creating them, with standard deviation indicating the variability

In generating multiple-choice questions, initial attempts yielded overly simple questions. Therefore, we revised the prompt to request highly complex questions with two-step reasoning. Despite this, over a third of the questions were still deemed too easy. As not all easy questions are inappropriate, we included 51 such questions, recognizing

Table III. Inter-rater reliability analysis of question suitability for Dermatology Board Exam

		Reviewer 1				
Reviewer 2		Unsuitable question *n* (%)	Suitable question *n* (%)	*n*	Kappa score	*p*-value
CBCL	Unsuitable	5 (16.7)	4 (13.3)	30	0.254	0.160
	Suitable	6 (20.0)	15 (50.0)			
CTCL	Unsuitable	11 (36.7)	9 (30.0)	30	0.217	0.196
	Suitable	3 (10.0)	7 (23.3)			
Biopsy techniques	Unsuitable	2 (10.0)	10 (50.0)	20	0.035	0.798
	Suitable	1 (5.0)	7 (35.0)			
Vasculitis	Unsuitable	40 (80.0)	0 (0.0)	50	0.615	<0.001
	Suitable	5 (10.0)	5 (10.0)			
Acne	Unsuitable	36 (60.0)	3 (5.0)	60	0.292	0.011
	Suitable	14 (23.3)	7 (11.7)			
Systemic disease	Unsuitable	13 (43.3)	9 (30.0)	30	0.267	0.099
	Suitable	2 (6.7)	6 (20.0)			
Ichthyoses	Unsuitable	28 (7.3)	2 (0.5)	40	0.385	0.011
	Suitable	6 (1.6)	4 (1.0)			
Darier disease	Unsuitable	7 (35.0)	9 (45.0)	20	0.107	0.494
	Suitable	1 (5.0)	3 (15.0)			
Alopecia	Unsuitable	15 (37.5)	18 (45.0)	40	0.014	0.900
	Suitable	3 (7.5)	4 (10.0)			
Mycobacteria	Unsuitable	11 (36.7)	4 (13.3)	30	0.000	1.00
	Suitable	11 (36.7)	4 (13.3)			
Rosacea	Unsuitable	10 (50.0)	1 (5.0)	20	0.140	0.413
	Suitable	7 (35.0)	2 (10.0)			
HPV	Unsuitable	16 (50.0)	7 (21.9)	32	0.486	0.003
	Suitable	1 (3.1)	8 (25.0)			
All questions	Unsuitable	**194 (48.3)**	**76 (18.9)**	**402**	**0.256**	**<0.001**
	Suitable	**60 (14.9)**	**72 (17.9)**			

the difficulty in distinguishing "too easy" from "easy but acceptable". Of the accepted questions, 29 were classified as hard, and 7 involved two-stage reasoning. This underscores the challenge of using ChatGPT-4 to produce suitably complex questions that accurately assess clinical scenarios. Additionally, 118 questions were flagged due to ambiguous wording or multiple correct answers. This highlights a key challenge for ChatGPT-4: ensuring clarity and a single correct answer to prevent disputes. Effective examination design requires questions to be not only factually accurate but also clear and precise, a standard that remains challenging for AI platforms to meet.

We assessed ChatGPT-4's ability to produce diverse questions and found that 2.2% of the questions were repeated, indicating limited novelty compared with human-generated questions. Students and residents might also generate questions that will be similar to those appearing on exams, highlighting a potential issue with using the platform. However, the reviewers recognized the educational value of assessing AI-generated questions, noting that this process fosters deeper engagement with the curriculum, which may be beneficial for medical students. This suggests that, with further refinement, AI could be adapted for various levels of medical education, enhancing learning outcomes across different stages (7, 10).

Future AI iterations in question generation should enhance algorithms to better assess question complexity and reduce ambiguities that may lead to appeals. This requires integrating expert feedback into the AI training process to align outputs with board-certified professionals' expectations. Collaboration between AI developers and educational experts is essential for advancing AI capabilities, ensuring outputs meet educational standards and learning objectives, and potentially improving both question-generation efficiency and educational support in medical training.

In conclusion, this analysis highlights both the potential and limitations of ChatGPT-4 in generating questions of varying difficulty and complex clinical scenarios. Key constraints include a significant proportion of overly simplistic questions and inaccurate distractor options. With improved training and contextual understanding, AI tools could better leverage their potential, addressing current limitations and generating diverse questions across subjects. Thus, ChatGPT-4 currently emerges as a supplementary tool for dermatology board exam preparation and may become more effective with forthcoming modifications.

The authors have no conflicts of interest to declare.

REFERENCES

1. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel) 2023; 11: 887. https://doi.org/10.3390/healthcare11060887
2. Joly-Chevrier M, Nguyen AXL, Lesko-Krleza M, Lefrançois P. Performance of ChatGPT on a practice Dermatology Board certification examination. J Cutan Med Surg 2023; 27: 407-409. https://doi.org/10.1177/12034754231188437
3. Eysenbach G. The role of ChatGPT, generative language models, and artificial intelligence in medical education: a conversation with ChatGPT and a call for papers. JMIR Med Educ 2023; 9: e46885. https://doi.org/10.2196/46885
4. Lee H. The rise of ChatGPT: exploring its potential in medical education. Anat Sci Educ 2024; 17: 926–931. https://doi.org/10.1002/ase.2270
5. Ayub I, Hamann D, Hamann CR, Davis MJ. Exploring the potential and limitations of Chat Generative Pre-trained Transformer (ChatGPT) in generating board-style dermatology questions: a qualitative analysis. Cureus 2023; 15: e43717. https://doi.org/10.7759/cureus.43717
6. Bolognia JL, Schaffer JV, Cerroni L. Dermatology, 4th edition. London?Hoboken, NJ: Elsevier, 2017.
7. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023; 2: e0000198. https://doi.org/10.1371/journal.pdig.0000198
8. Barbour AB, Barbour TA. A radiation oncology board exam of ChatGPT. Cureus 2023; 15: e44541. https://doi.org/10.7759/cureus.44541
9. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023; 9: e45312. https://doi.org/10.2196/45312
10. Passby L, Jenko N, Wernham A. Performance of ChatGPT on Specialty Certificate Examination in Dermatology multiple-choice questions. Clin Exp Dermatol 2024; 49: 722–727. https://doi.org/10.1093/ced/llad197