Commentary

Evaluating ChatGPT-4o's accuracy in answering American Board of Dermatology practice questions: an analysis of AI in dermatology residency education

Lauren McGrath, BA[1a], Nathan Schedler, BS, BA[2], Sarah Martin, BS[2], Matthew Hrin, MD[3], Maria Mariencheck, MD, PhD[3], Steven Feldman, MD, PhD[3,4,5], Zeynep Akkurt, MD[3]

[1] Center for Dermatology Research, Department of Dermatology, Wake Forest School of Medicine, Winston-Salem, North Carolina, United States, [2] Wake Forest School of Medicine, Winston-Salem, North Carolina, United States, [3] Department of Dermatology, Wake Forest University School of Medicine, Winston-Salem, North Carolina, United States, [4] Department of Pathology, Wake Forest School of Medicine, Winston-Salem, North Carolina, United States, [5] Department of Social Sciences & Health Policy, Wake Forest School of Medicine, Winston-Salem, North Carolina, United States

Keywords: artificial intelligence, board examinations, ChatGPT-4o, education

Dermatology Online Journal

Vol. 31, Issue 4, 2025

Abstract

Artificial intelligence may enhance medical education. This study evaluates ChatGPT-4o's accuracy in answering sample questions from the American Board of Dermatology BASIC, CORE, and APPLIED examinations. Fifty publicly available questions, with and without images, were analyzed for accuracy and performance across difficulty levels and categories. Its performance varied significantly between text-only and image-based questions, with lower accuracy on image-based questions (47%). Improvements in artificial intelligence for the use in dermatology residency education are necessary, as limitations in visual diagnostic skills were evident.

Introduction

Artificial intelligence and large language processing models could transform medicine by enhancing diagnostics, treatment care plans, language translation, and overall decision support.[1-3] This study assesses ChatGPT-4o's performance on American Board of Dermatology (ABD) practice questions, utilizing sample questions from the ABD's BASIC, CORE, and APPLIED examinations. The evaluation aims to provide insights into the accuracy and reliability of artificial intelligence, specifically ChatGPT-4o, in dermatology education and examination preparation and its potential role in supporting knowledge-based decision-making in patient care.

Fifty publicly available sample questions from the ABD's BASIC, CORE, and APPLIED examinations were entered into ChatGPT-4o without the corresponding answer key.[4] ChatGPT-4o's responses were then compared to the ABD's answer key. The selected questions were analyzed to see if they contained images of histology, pathology, or physical examination findings. A dermatology resident and an attending dermatologist rated the difficulty of each question on a scale of 1 to 3, with 1 indicating *easy* and 3 indicating *difficult*. If ratings were not equivalent, the more difficult rating was selected.

Discussion

ChatGPT-4o scored 53%, 65%, and 67% on the BASIC, CORE, and APPLIED examinations (Table 1). The score for questions containing images was 47%. ChatGPT-4o exhibited lower performance in questions related to Diagnostic Skills and Disease Recognition, Therapeutics and Management, Surgical and Procedural Knowledge, and Histopathology and Immunology, with scores of 62%, 64%, 50%, and 40% (Table 1). ChatGPT-4o scored 38% on easy questions, 70% on medium questions, and 56% on difficult questions (Table 2).

Chi-square tests of independence were performed. There was no statistically significant difference in the distribution of correct and incorrect responses across the BASIC, CORE, and APPLIED examination types ($\chi^2(2)=0.69$, $P=0.707$), the question topic categories ($\chi^2(5)=5.39$, $P=0.371$), or the question difficulty rating ($\chi^2(2)=3.03$, $P=0.220$). There was a statistically significant difference in ChatGPT-4o's performance between questions with and without images. Questions without images yielded more correct responses ($\chi^2(1)=6.94$, $P=0.008$).

There is no predetermined score for passing the BASIC, CORE, or APPLIED examinations. Psychometricians

a Corresponding Author: Lauren McGrath BA, Center for Dermatology Research, Department of Dermatology, Wake Forest University School of Medicine, 4618 Country Club Road, Winston-Salem, NC 27104, Email: laurmcg21@gmail.com

Table 1. *ChatGPT-4o Performance Metrics by ABD Examination, Question Type, and Category.*

		BASIC	CORE	APPLIED	Total
Total Questions	Correct, n	8	13	10	31
	Incorrect, n	7	7	5	19
	Total, n	15	20	15	50
	Percent Score, %	53%	65%	67%	62%
Questions with Images	Correct, n	4	3	8	15
	Incorrect, n	7	5	5	17
	Total, n	11	8	13	32
	Percent Score, %	36%	38%	62%	47%
Questions without Images	Correct, n	4	10	2	16
	Incorrect, n	0	2	0	2
	Total, n	4	12	2	18
	Percent Score, %	100%	83%	100%	89%
Question topic category: Diagnostic Skills and Disease Recognition	Correct, n	8			
	Incorrect, n	5			
	Percent Score, %	62%			
Question topic category: Pathophysiology and Mechanism of Disease	Correct, n	4			
	Incorrect, n	0			
	Percent Score, %	100%			
Question topic category: Therapeutics and Management	Correct, n	9			
	Incorrect, n	5			
	Percent Score, %	64%			
Question topic category: Surgical and Procedural Knowledge	Correct, n	3			
	Incorrect, n	3			
	Percent Score, %	50%			
Question topic category: Research and Study Design Fundamentals	Correct, n	1			
	Incorrect, n	0			
	Percent Score, %	100%			
Question topic category: Histopathology and Immunology	Correct, n	4			
	Incorrect, n	6			
	Percent Score, %	40%			

analyze test performance each year to determine the passing rate, and mean examination scores for the CORE and APPLIED examinations are not published.[4] The mean total percent correct for the 2024 BASIC examination was 79% with a standard deviation of 7%.[4] ChatGPT-4o's score of 53% on the BASIC examination is greater than three standard deviations below the mean.

Conclusion

ChatGPT-4o was more effective at processing text-based information rather than visual content. This is a key limitation in its ability to handle visual analysis crucial in dermatology diagnosis, treatment, and management. Al-

Table 2. *ChatGPT-4o Performance Metrics by ABD Examination and Rated Question Difficulty.*

Examination	"Easy" Questions		"Medium" Questions		"Difficult" Questions	
	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect
BASIC	1	5	5	2	2	0
CORE	2	0	10	5	1	2
APPLIED	0	0	8	3	2	2
Total	3	5	23	10	5	4

though artificial intelligence may have the capabilities to assist with text-based learning, its integration into dermatology education and patient care requires further advancements in visual analysis capabilities. Improving image analysis in future generations of ChatGPT could potentially help support dermatology residents in educational settings such as board examination preparation.

Potential conflicts of interest

Steven Feldman has received research, speaking and/or consulting support from a variety of companies including Galderma, GSK/Stiefel, Almirall, Leo Pharma, Boehringer Ingelheim, Mylan, Celgene, Pfizer, Valeant, Abbvie, Samsung, Janssen, Lilly, Menlo, Merck, Novartis, Regeneron, Sanofi, Novan, Qurient, National Biological Corporation, Caremark, Advance Medical, Sun Pharma, Suncare Research, Informa, UpToDate, and National Psoriasis Foundation. He is founder and majority owner of www.DrScore.com and founder and part owner of Causa Research, a company dedicated to enhancing patients' adherence to treatment.

References

1. Haug CJ, Drazen JM. Artificial intelligence and machine learning in clinical medicine, 2023. *The New England Journal of Medicine*. 2023;388(13):1201-1208. doi:10.1056/NEJMra2302038. PMID:36988595

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 Digital Health*. 2023;2(2):e0000198. doi:10.1371/journal.pdig.0000198. PMID:36812645

3. Passby L, Jenko N, Wernham A. Performance of ChatGPT-4o on specialty certificate examination in Dermatology multiple-choice questions. *Clinical and Experimental Dermatology*. 2024;49(7):722-727. doi:10.1093/ced/llad197. PMID:37264670

4. ABD Certification Pathway: Sample Items. American Board of Dermatology. 2024. Accessed November 17, 2024. https://www.abderm.org/residents-and-fellows/abd-certification-pathway/abd-certification-pathway-sample-items