Artificial intelligence in dermatology: GPT-3.5-Turbo demonstrates strengths and limitations in residency exams

Dylan Haynes MD MCR, William Lewis MD, Neha N Jariwala MD

Affiliations: Department of Dermatology, Hospital of the University of Pennsylvania, Philadelphia, Pennsylvania, USA

Corresponding Author: Dylan Haynes MD MCR, 2 Maloney Building, 3600 Spruce Street, Philadelphia, PA 19104, Tel: 215-662-7883, Email: Dylan.Haynes@pennmedicine.upenn.edu

Keywords: artificial intelligence, basic, ChatGPT, exam, test

To the Editor:

OpenAI's GPT-3.5-Turbo (GPT), an artificial intelligence (AI) language model which powers ChatGPT, exhibits impressive advancements in complex problem solving and language comprehension. We sought to explore its applications in dermatologic education and practice.

Utilizing OpenAI's API within python, we prompted GPT to answer all questions from the Basic Exam database from a popular online question bank, DermQBank (www.dermqbank.com). On a standard desktop computer (Intel Core™ i9-10900K CPU with 32 GB of RAM, solid-state storage, and an NVIDIA RTX 3070 GPU), the model processed and answered all questions in approximately 755 seconds, costing $0.30 for 150126 tokens—a standardized unit of text reflecting the cumulative prompt-answer size.

The results of GPT's performance across each domain are as follows: dermatopathology 37% (46/124), general dermatology 54% (128/236), pediatric dermatology 53% (16/30), science and research 58% (43/74), surgical dermatology 49% (54/109), and visual recognition (general dermatology) 38% (51/131). Overall, GPT scored 48.0% (338/704), in contrast to the average score of postgraduate year two test takers at 64.4%.

These findings point to relative strengths of GPT in dermatology, notably in general dermatology and science and research. However, the model's inferior performance in dermatopathology and visual recognition emphasizes a significant drawback: the inability to process image input, a critical component of dermatological diagnosis and education. Despite the current limitations of GPT, AI as a whole has shown robustness in image analysis within dermatology, suggesting an avenue for natural progression [1].

GPT's performance herein aligns with a growing trend in AI's role in standardized tests. Studies have shown AI scoring at, or near the passing threshold for the Uniform Bar Exam and United States Medical Licensing Exam [2]. It has even outperformed human candidates in a Virtual Objective Structured Clinical Examination in obstetrics and gynecology [3]. Yet, GPT's capabilities extend beyond merely taking tests, as evidenced by a recent single-blinded observer study which demonstrated GPT's ability to generate AI-derived dermatology case reports indistinguishable from those written by humans [4]. These findings underline the potential of AI in medical education and practice, as well as its versatility in creating high-quality clinical documentation.

However, it is important to note that even with AI's incorporation, it doesn't necessarily improve patient outcomes in point-of-care settings. Studies have highlighted the impact of Electronic Medical Record alerts and "alert fatigue" on emergency physician workflow and medical management, signifying potential limitations [5]. AI's potential role in dermatology education and practice is undoubtedly exciting, especially with the advent of GPT-4. However, the results of this study underscore that AI,

in its current form, should be viewed as a tool to supplement, not replace, human expertise in dermatology. Further research is needed to enhance AI's understanding of visual inputs and further refine its potential applications in the medical field.

As we continue to leverage AI in dermatology, we must balance enthusiasm with a healthy skepticism. An AI's true value lies in its ability to enhance human expertise and patient care, not replace it. This study serves as a stepping stone in understanding and improving AI's role in dermatology.

Potential conflicts of interest

The authors declare no conflicts of interest.

References

1. El-Khatib H, Popescu D, Ichim L. Deep Learning-Based Methods for Automatic Diagnosis of Skin Lesions. *Sensors (Basel)*. 2020;20:1753. [PMID: 32245258].
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:e0000198. [PMID: 36812645].
3. Li SW, Kemp MW, Logan SJS, et al. ChatGPT Outscored Human Candidates in a Virtual Objective Structured Clinical Examination (OSCE) in Obstetrics and Gynecology. *Am J Obstet Gynecol.* 2023; S0002-9378(23)00251-X. [PMID: 37088277].
4. Dunn C, Hunter J, Steffes W, et al. Artificial intelligence-derived dermatology case reports are indistinguishable from those written by humans: A single-blinded observer study. *J Am Acad Dermatol*. 2023;S0190-9622(23)00587-X. [PMID: 37054810].
5. Todd B, Shinthia N, Nierenberg L, et al. Impact of Electronic Medical Record Alerts on Emergency Physician Workflow and Medical Management. *J Emerg Med*. 2021;60:390-395. [PMID 33298357].