doi:10.36849/JDD.9100e

A Qualitative Assessment of Chatbots for Common Dermatologic Conditions

Robert Adler BA, Aaron Lavi BA BBA, Rachel Berglas BA, Netanel Yomtov BA, Isaac Inoyatov BA, Daniel Yusupov BA, David Musheyev BA, Jessica L. Feig MD PhD, Justin W. Marson MD

SUNY Downstate Department of Dermatology, Brooklyn, NY

INTRODUCTION

Artificial Intelligence (AI) chatbots have become increasingly popular tools for patients seeking health information. The ability of AI chatbots to offer accurate, understandable, and reliable information for common conditions can bridge gaps in healthcare information access.[1-3] This study explores how well 4 chatbots, OpenAI ChatGPT 4o mini, Microsoft Copilot, Google Gemini (Flash 1.5), and Perplexity AI (Version 2.33.3), perform in providing guidance on common dermatologic conditions and comparing them to American Academy of Dermatology (AAD) patient resources.[4]

A prompt, "tell me about (condition)," was entered into 4 chatbots for 25 dermatologic conditions (Table 1), which were assessed for readability, accuracy, and quality. In addition, 280 articles from the AAD website that discussed these 25 conditions were also assessed. The DISCERN tool evaluated quality while the Patient Education Materials Assessment Tool (PEMAT) judged understandability and actionability. The Canadian Press's 1-5 scale was employed to rate misinformation. Intraclass Correlation (DISCERN, Misinformation) and Cohen's Kappa (PEMAT) assessed inter-rater reliability, with ANOVA comparing significance.

There was moderate, good, and high level of agreement between raters for Misinformation, DISCERN, and PEMAT, respectively. The quality of responses was moderate for all, with Copilot and Perplexity scoring highest and Gemini lowest. Chatbots generally lacked comparative details regarding potential treatments, leading to lower scores. Understandability and actionability were relatively high, with ChatGPT performing best. All chatbot responses were accurate. ChatGPT offered the most thorough information, whereas Gemini's responses required confirmatory information, warranting a lower score (Table 2). Copilot and Perplexity provided consistent sources within responses, while Gemini was variable in sourcing, and ChatGPT lacked sources.

Chatbot responses were at a 10th-grade reading level; however, as 54% of adults in the US read below a 6th-grade level, chatbot responses may be poorly understood.[5] ChatGPT, in particular, was more verbose than the other chatbots. While this can explain its comprehensiveness and accuracy, it might also overwhelm or distract users from crucial information. While chatbots could customize responses by reading level, it is unlikely patients would specifically prompt this.[2] Compared to the four chatbots, the AAD website's materials were more readable and specifically tailored, though chatbot responses were more concise and consistent in word count.

TABLE 1.

Twenty-Five Conditions Assessed for Chatbot Performance
Condition
Acne
Actinic keratosis
Basal cell carcinoma
Birthmarks
Cellulitis
Cold sores
Contact dermatitis
Eczema
Folliculitis
Hair loss
Hives
Hyperhidrosis
Keloid scars
Lyme disease
Melanoma
Moles
Nail fungus
Psoriasis
Rosacea
Seborrheic Dermatitis
Seborrheic keratoses
Skin tags
Squamous cell carcinoma
Vitiligo
Warts

doi:10.36849/JDD.9100e

TABLE 2.

Scores Comparing Chatbot Performance

Chatbot	Microsoft Copilot (±SD)	OpenAI ChatGPT (±SD)	Google Gemini (±SD)	Perplexity AI (±SD)	AAD Articles (±SD)	Inter-Rater Reliability (95% CI)	Chatbot Significance *P*=
DISCERN (Score Totals/90)	64.4 (±2.94)	50.86 (±1.47)	45.9 (±4.66)	62.24 (±1.61)	--	0.7708 (0.2815-0.8747)	<.001
PEMAT (%)	89.17 (±2.85)	92.32 (±1.43)	89.95 (±3.79)	91.23 (±2.31)	--	0.8778 (0.8440 to 0.9117)	0.0032
Grade Level	10.124 (±1.46)	10.088 (±1.34)	10.364 (±1.83)	10.396 (±1.91)	7.546 (±1.47)	—	0.8738
Word Count	349 (±73.51)	604 (±87.37)	277 (±48.40)	339 (±64.44)	779.56 (±1234.58)	—	<.001
Misinformation (1-5)	4.84 (±.345)	4.9 (±.289)	4.58 (±.449)	4.78 (±.384)	--	0.6904 (0.5643-0.7829)	0.018
Top 5 Sources (frequency)	Mayo Clinic (33) Wikipedia (18) Cleveland Clinic (16) WebMD (13) MSN (13)	--	--	Mayo Clinic (20) AAD (18) WebMD (12) NIH (11) Cleveland Clinic (10)	--	--	--

DISCERN scores are on a scale of 0-90. PEMAT scores are ranked as a percentage of "yes" answers. Misinformation scored on a 1-5 scale.

Although chatbots show promise in their ability to deliver quality, accurate, and readable information to patients, patients should still be encouraged to consult healthcare professionals or expert material for comprehensive, accurate guidance. One key limitation of chatbots is their poor reproducibility and the cross-sectional nature of the data; responses to the same query are often inconsistent between users, and chatbots are rapidly evolving. Additionally, as research and treatments continue to advance, chatbots may lag slightly behind regarding access to information concerning ongoing trials and studies. Furthermore, chatbot interactions are often a succession of prompts and responses, which does not assess their ability to synthesize previous responses to provide more tailored information. As chatbots are integrated into healthcare systems and practitioners become more attuned to their various uses and limitations, future versions may be able to overcome crucial gaps in deliverance of quality, accurate healthcare information.

DISCLOSURES

The authors have no conflicts of interest of disclose.

REFERENCES

1. Reynolds K, Tejasvi T. Potential use of ChatGPT in responding to patient questions and creating patient resources. *JMIR Dermatol.* 2024;7:e48451. doi: 10.2196/48451. PMID: 38446541; PMCID: PMC10955382.
2. Lambert R, Choo ZY, Gradwohl K, et al. Assessing the application of large language models in generating dermatologic patient education materials according to reading level: qualitative study. *JMIR Dermatol.* 2024;7:e55898. doi: 10.2196/55898. PMID: 38754096; PMCID: PMC11140271.
3. Lakdawala N, Channa L, Gronbeck C, et al. Assessing the accuracy and comprehensiveness of ChatGPT in offering clinical guidance for atopic dermatitis and acne vulgaris. *JMIR Dermatol.* 2023;6:e50409. doi: 10.2196/50409. PMID: 37962920; PMCID: PMC10685272.
4. For the public. Aad.org. https://www.aad.org/public. Accessed January 25th, 2025.
5. National Literacy Institute. Literacy Statistics 2024- 2025 (Where we are now). National Literacy. https://www.thenationalliteracyinstitute.com/post/literacy-statistics-2024-2025-where-we-are-now. Accessed January 25th, 2025.

AUTHOR CORRESPONDENCE

Robert Adler BA

E-mail:..... robertadler119@gmail.com