# The comparative sufficiency of ChatGPT, Google Bard, and Bing AI in answering diagnosis, treatment, and prognosis questions about common dermatological diagnoses

Courtney Andrea Chau, Hao Feng, Gabriela Cobos, Joyce Park

Submitted to: JMIR Dermatology
on: May 22, 2024

# *Table of Contents*

**Original Manuscript** ..... **4**
**Supplementary Files** ..... **11**
Multimedia Appendixes ..... **12**
Multimedia Appendix 1 ..... **12**

# The comparative sufficiency of ChatGPT, Google Bard, and Bing AI in answering diagnosis, treatment, and prognosis questions about common dermatological diagnoses

Courtney Andrea Chau[1] BS; Hao Feng[2] MD; Gabriela Cobos[3] MD; Joyce Park[4] MD

[1]Icahn School of Medicine at Mount Sinai New York US
[2]Department of Dermatology University of Connecticut Health Center Farmington US
[3]Department of Dermatology Tufts Medical Center Boston US
[4]Skin Refinery PLLC Spokane US

**Corresponding Author:**
Gabriela Cobos MD
Department of Dermatology
Tufts Medical Center
260 Tremont St
Fl 13
Boston
US

## *Abstract*

Our team explored the utility of unpaid versions of three artificial intelligence chatbots in offering patient-facing responses to questions about five common dermatological diagnosis, and highlights the strengths and limitations of different AI chatbots, while demonstrating how chatbots present the most potential in tandem with a dermatologist's diagnosis.

(JMIR Preprints 22/05/2024:60827)
DOI: https://doi.org/10.2196/preprints.60827

### 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 v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

# Original Manuscript

**Article type:** Research Letter

**Title:** The comparative sufficiency of ChatGPT, Google Bard, and Bing AI in answering diagnosis, treatment, and prognosis questions about common dermatological diagnoses

**Author list:** Courtney A. Chau, BS[1], Hao Feng MD,[2] Gabriela Cobos MD,[3*] and Joyce Park MD[4*]

**Author affiliations:**

[1]Icahn School of Medicine at Mount Sinai, New York, NY

[2] Department of Dermatology, University of Connecticut Health Center, Farmington, CT

[3] Department of Dermatology, Tufts Medical Center, Boston, MA

[4] Skin Refinery PLLC

*Drs Cobos and Park are cosenior authors.

**Corresponding author:**

Gabriela Cobos, MD

260 Tremont St, Fl 13 Boston, MA 02116

gabriela.cobos@tuftsmedicine.org

**Manuscript word count:** 583 words

**Abstract:** Our team explored the utility of unpaid versions of three artificial intelligence chatbots in offering patient-facing responses to questions about five common dermatological diagnosis, and highlights the strengths and limitations of different AI chatbots, while demonstrating how chatbots present the most potential in tandem with a dermatologist's diagnosis.

**Keywords:** artificial intelligence, ChatGPT, atopic dermatitis, acne vulgaris, cyst, actinic keratosis, rosacea

## Introduction

Artificial intelligence (AI) chatbots, such as ChatGPT, offer a platform for patients to ask medical questions, particularly when access to care is limited.[1] Studies have assessed the utility of ChatGPT in dermatology; however, fewer studies have compared performance between chatbots.[2] This study compares the clinical utility of ChatGPT 3.5, Google Bard, and Bing AI in generating patient-facing responses to questions about five common dermatological diagnoses (atopic dermatitis, acne vulgaris, actinic keratosis, cyst, and rosacea).[3] Only unpaid versions of chatbots were used, as these are most accessible to patients.

## Methods

For each condition, two diagnosis, two treatment, and one prognosis question were devised. Diagnosis questions requested a diagnosis and presented a patient history including age, sex, symptoms (duration/location), treatments tried and outcomes, and medical history. 19 questions were modeled from questions on Reddit forums (“r/AskDocs” and “r/dermatology”). For topics with insufficient Reddit questions, the co-authors devised prompts reflecting common questions in their experience (6 questions).

Questions were inputted into each chatbot (Supplementary Table 1). Three board-certified dermatologists scored the responses on appropriateness for a patient-facing platform (Yes/No), sufficiency for clinical practice (Yes/No: not specific/No: not concise/No: inaccurate information), accuracy from 1 (completely inaccurate) to 6 (completely accurate), and overall from 1 (worst possible answer) to 10 (best possible answer).[4] The Wilcoxon signed-rank test was used for pairwise comparisons (Table 1). P-values were adjusted using the Bonferroni correction.

**Table 1.** Descriptive statistics of scores between ChatGPT 3.5, Google Bard, and Bing AI.

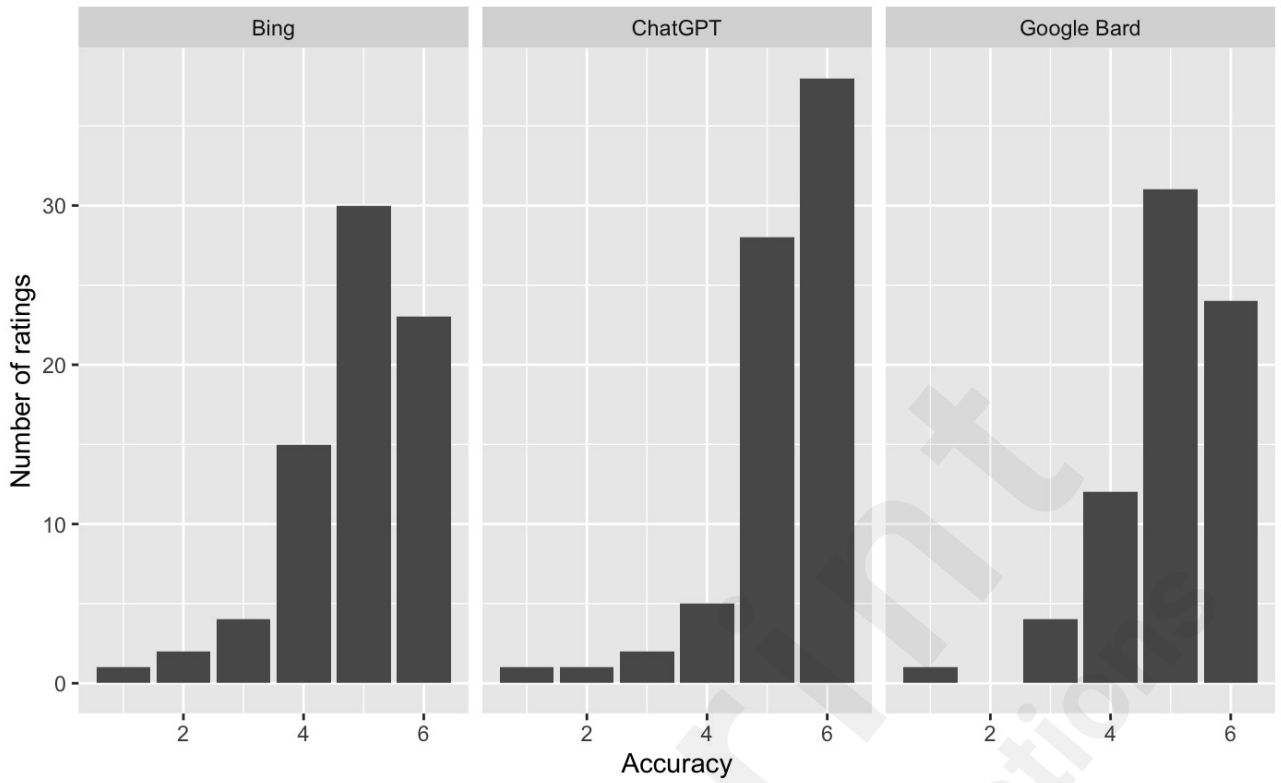
| | **ChatGPT 3.5 (n=75)** | **Google Bard (n=72)** | **Bing AI (n=75)** |
|---|---|---|---|

| | | | |
|---|---|---|---|
| Mean Flesch Reading Ease Score (SD) | 33.90 (8.12) | 49.72 (15.42) | 46.53 (9.67) |
| Mean accuracy (SD) | 5.29 (0.97) | 5.00 (0.98) | 4.87 (1.08) |
| Mean overall rating (SD) | 8.37 (1.81) | 7.94 (1.85) | 7.41 (2.13) |
| Number of responses appropriate for a patient-facing platform (%) | 71 (94.67) | 65 (90.28) | 65 (86.67) |
| Sufficiency for clinical practice | | | |
| Yes (%) | 41 (54.67) | 35 (48.61) | 35 (46.67) |
| No: not specific enough (%) | 14 (18.67) | 15 (20.83) | 23 (30.67) |
| No: inaccurate information (%) | 20 (26.67) | 20 (27.78) | 17 (22.67) |
| No: not concise (%) | 0 | 2 (2.78) | 0 |

**Results**

One response was omitted because Google Bard declined to answer a question. ChatGPT responses had significantly lower Flesch Reading Ease Scores than Google Bard (P<.001) and Bing AI (P<.001), indicating lower comprehensibility. Responses from ChatGPT received significantly higher accuracy (P=.01, Figure 1) and overall (P=.003) ratings than Bing AI. In terms of patient-facing platform appropriateness and clinical practice sufficiency, ChatGPT received the most appropriate (94.67%) and sufficient (54.67%) ratings. Bing AI received the fewest (86.67% and 54.67%, respectively). 45.33%, 48.61%, and 53.33% of ChatGPT, Google Bard, and Bing AI responses, respectively, had inaccurate information or were not specific. For diagnosis prompts, 9/10 of ChatGPT and Bing AI and 7/10 of Google Bard responses included the intended diagnosis. Of the 25 responses from each chatbot, 25 of Bing AI's, 24 of ChatGPT's, and 19 of Google Bard's responses emphasized the importance of consulting a healthcare professional before acting.

**Figure 1.** Distribution of accuracy ratings for each chatbot.




## Discussion

ChatGPT outputs were most accurate and appropriate for patient questions. However, ChatGPT responses had college-level readability, limiting its utility for the public.[5] Only approximately half of the responses were sufficient for clinical practice, primarily due to inaccuracies and lack of specificity. ChatGPT and Bing AI performed best at diagnosis and favorably emphasized the importance of seeking input from a healthcare professional. Google Bard did not perform well in these domains, suggesting that it is less useful for offering patient advice. Despite acceptable diagnostic performance of ChatGPT and Bing AI, an unranked list of conditions with differing treatments is not actionable for patients. Chatbots present more potential in offering advice once a diagnosis has been established.

ChatGPT 3.5 displays the most promise of the chatbots analyzed in this study, consistent with Mu et al. (2024), who compared the chatbots' responses to melanoma questions. However, before AI can be harnessed to address patient concerns, it must be improved by enhancing readability, removing inaccuracies, and improving information specificity. Its utility is most promising in tandem

with a dermatologist’s diagnosis, rather than as an independent entity. As access to AI grows, dermatologists must be aware of the quality of information patients may receive from AI and how it may differ from a dermatologist’s advice.

## References


1. Baker MN, Burruss CP, Wilson CL. ChatGPT: A Supplemental Tool for Efficiency and Improved Communication in Rural Dermatology. Cureus. 2023;15(8):e43812. Published 2023 Aug 20. doi:10.7759/cureus.43812
2. Mu X, Lim B, Seth I, et al. Comparison of large language models in management advice for melanoma: Google's AI BARD, BingAI and ChatGPT. Skin Health Dis. 2023;4(1):e313. Published 2023 Nov 28. doi:10.1002/ski2.313
3. Landis ET, Davis SA, Taheri A, Feldman SR. Top dermatologic diagnoses by age. Dermatol Online J. 2014;20(4):22368. Published 2014 Apr 16.
4. Young JN, Ross O'Hagan, Poplausky D, et al. The utility of ChatGPT in generating patient-facing and clinical responses for melanoma. J Am Acad Dermatol. 2023;89(3):602-604. doi:10.1016/j.jaad.2023.05.024
5. Hutchinson N, Baird GL, Garg M. Examining the Reading Level of Internet Medical Information for Common Internal Medicine Diagnoses. Am J Med. 2016;129(6):637-639. doi:10.1016/j.amjmed.2016.01.008

# Supplementary Files

# Multimedia Appendixes

Supplementary Table 1. Diagnosis, treatment, and prognosis prompts inputted into chatbots.
URL: http://asset.jmir.pub/assets/6587c219af94a72bf334413d2563ac58.docx