

1 **Article Type:** Research Letter

2 **Title:** Assessing GPT-4's Diagnostic Accuracy with Darker Skin Tones: Underperformance and
3 Implications

4 **Authors:** Edgar Akuffo-Addo (MHA)^{1*}, Luna Samman (BA)^{2*}, Leena Munawar (MD)³, Maya
5 Akbik (BS)⁴, Nelly Kokikian (BS)⁵, Raquel Wescott (BS)⁶, Jashin J. Wu (MD)⁷

6 *Edgar Akuffo-Addo and Luna Samman have contributed equally to this study and should be
7 considered as co-first authors.

8

9 **Author Affiliation:**

10 ¹ Division of Dermatology, Department of Medicine, University of Toronto, Toronto, Ontario,
11 Canada

12 ² Rowan School of Osteopathic Medicine, Stratford, New Jersey, USA

13 ³ University of Texas Medical Branch, Galveston, Texas, USA

14 ⁴ Medical College of Georgia, AU/UGA Medical Partnership, Athens, GA, USA

15 ⁵ Department of Medicine, Division of Dermatology, David Geffen School of Medicine,
16 University of California, Los Angeles, California, USA

17 ⁶ University of Nevada, Reno School of Medicine, Reno, Nevada

18 ⁷ Department of Dermatology, University of Miami, Miller School of Medicine, Miami, Florida,
19 USA

20 **Corresponding Author:** Edgar Akuffo-Addo
21 1 King's College Circle
22 Toronto, ON M5S 1A8
23 Email: edgar.addo@mail.utoronto.ca

24 **Manuscript word count:** 489

25 **Supplemental Files:** 0

26 **Keywords:** GPT-4, Images, Dermatology, Darker skin, Underperformance

27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49

Competing Interests: Dr. Wu declares being an investigator, consultant, or speaker for AbbVie, Almirall, Amgen, Arcutis, Aristea Therapeutics, Bausch Health, Boehringer Ingelheim, Bristol Myers Squibb, Dermavant, DermTech, Dr Reddy's Laboratories, Eli Lilly, EPI Health, Galderma, Janssen, LEO Pharma, Mindera, Novartis, Pfizer, Regeneron, Samsung Bioepis, Sanofi Genzyme, Solius, Sun Pharmaceutical Industries, UCB, and Zerigo Health.

Acknowledgements: None

Data Sharing Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

50 **Abstract**

51 **Introduction:** Conversational artificial intelligence (AI) language models like ChatGPT have
52 emerged as promising tools for patients seeking medical information and guidance. However,
53 their use raises ethical concerns due to the potential for inaccurate medical advice that could
54 harm patients. Previous studies in dermatological machine-learning have highlighted that the
55 underrepresentation of diverse skin types in research could lead to bias and reduced performance
56 in evaluating skin lesions in darker skin tones. This study aims to assess the accuracy of GPT-4
57 in generating appropriate differential diagnoses and arriving at the correct diagnoses for common
58 skin lesions. Additionally, we investigate any differences in its diagnostic accuracy between
59 darker and lighter skin tones.

60

61 **Method:** Fifty images were randomly selected from the Fitzpatrick 17k dataset, a publicly
62 available online collection of clinical images labelled with the appropriate diagnoses and skin
63 types based on the Fitzpatrick scoring system. Half of the images selected represented darker
64 skin tones, Fitzpatrick IV-VI, and the other half represented lighter skin tones, Fitzpatrick I-II.
65 For each selected dermatological condition, GPT-4 was presented with pairs of images - one
66 from a lighter skin tone and another from a darker skin tone. GPT-4 was then asked to provide its
67 top three differential diagnoses and a final diagnosis for each pair. The responses generated by
68 GPT-4 were transcribed and compared against the labels provided in the dataset to evaluate
69 accuracy. Subsequently, a univariate linear regression analysis was conducted to investigate the
70 relationship between Fitzpatrick skin type and diagnostic accuracy of GPT-4.

71

72 **Results:** Out of the 50 selected images, the distribution of Fitzpatrick skin types was as follows:
73 40% were Fitzpatrick type I, 10% were type II, 4% were type IV, 26% were type V, and 20%
74 were type VI. Overall, GPT-4 correctly diagnosed the condition in 28% of the images
75 (n=14/50), while the correct diagnosis was included in its list of top differentials for 48% of the
76 images (n=24/50). GPT-4 exhibited better performance in providing the correct diagnosis for
77 lighter skin tones (44%, n=11/25) compared to darker skin tones (12%, n=3/25), and this was
78 statistically significant (p-value < 0.05). Furthermore, with each unit increase in the Fitzpatrick
79 scale, GPT-4s performance decreased by 11.4% in accurately providing a differential diagnosis
80 and by 7.1% in accurately providing the correct diagnosis.

81

82 **Conclusion:** GPT-4 exhibited significantly lower overall accuracy compared to previous studies
83 reporting accuracies as high as 90%. This discrepancy highlights GPT-4s potential limitations in
84 providing accurate information without sufficient clinical context. While GPT-4 could serve as a
85 valuable learning tool for medical students and dermatology residents, it may not be suitable for
86 patients seeking clinical input to self-diagnose lesions at home. It is important to note that this
87 study is limited by its relatively small sample size, which could impact the generalizability of the
88 findings. If GPT-4 is to be considered for use by patients in a clinical setting, it is important to
89 ensure that it demonstrates high accuracy and remains unbiased across all patient demographics
90 and skin types.

91

92

93 **Research letter**

94 Conversational artificial intelligence (AI) language models like ChatGPT have emerged as
95 promising tools for patients seeking medical information and guidance.¹ However, their use
96 raises ethical concerns due to the potential for inaccurate medical advice that could harm
97 patients.² Previous studies in dermatological machine-learning have highlighted that the
98 underrepresentation of diverse skin types in research could lead to bias and reduced performance
99 in evaluating skin lesions in darker skin tones.³ This study aims to assess the accuracy of GPT-4
100 in generating appropriate differential diagnoses and arriving at the correct diagnoses for common
101 skin lesions. Additionally, we investigate any differences in its diagnostic accuracy between
102 darker and lighter skin tones.

103

104 Fifty images were randomly selected from the *Fitzpatrick 17k* dataset, a publicly available online
105 collection of clinical images labelled with the appropriate diagnoses and skin types based on the
106 Fitzpatrick scoring system.⁴ Half of the images selected represented darker skin tones,
107 Fitzpatrick IV-VI, and the other half represented lighter skin tones, Fitzpatrick I-II. For each
108 selected dermatological condition, GPT-4 was presented with pairs of images - one from a
109 lighter skin tone and another from a darker skin tone. GPT-4 was then asked to provide its top
110 three differential diagnoses and a final diagnosis for each pair. The responses generated by GPT-
111 4 were transcribed and compared against the labels provided in the dataset to evaluate accuracy.
112 Subsequently, a univariate linear regression analysis was conducted to investigate the
113 relationship between Fitzpatrick skin type and diagnostic accuracy of GPT-4.

114

115 Out of the 50 selected images, the distribution of Fitzpatrick skin types was as follows: 40%
116 were Fitzpatrick type I, 10% were type II, 4% were type IV, 26% were type V, and 20% were
117 type VI. Overall, GPT-4 correctly diagnosed the condition in 28% of the images (n=14/50),
118 while the correct diagnosis was included in its list of top differentials for 48% of the images
119 (n=24/50). GPT-4 exhibited better performance in providing the correct diagnosis for lighter skin
120 tones (44%, n=11/25) compared to darker skin tones (12%, n=3/25), and this was statistically
121 significant (p-value < 0.05). Furthermore, with each unit increase in the Fitzpatrick scale, GPT-
122 4's performance decreased by 11.4% in accurately providing a differential diagnosis and by
123 7.1% in accurately providing the correct diagnosis.

124
125 GPT-4's exhibited significantly lower overall accuracy compared to previous studies reporting
126 accuracies as high as 90%.⁵ This discrepancy highlights GPT-4's potential limitations in
127 providing accurate information without sufficient clinical context. While GPT-4 could serve as a
128 valuable learning tool for medical students and dermatology residents, it may not be suitable for
129 patients seeking clinical input to self-diagnose lesions at home. It is important to note that this
130 study is limited by its relatively small sample size, which could impact the generalizability of the
131 findings. If GPT-4 is to be considered for use by patients in a clinical setting, it is important to
132 ensure that it demonstrates high accuracy and remains unbiased across all patient demographics
133 and skin types.

134
135
136
137

138 References

- 139 1. Shah YB, Ghosh A, Hochberg AR, et al. Comparison of ChatGPT and traditional patient
140 education materials for men’s health. *Urology Practice*. 2024;11(1):87-94.
- 141 2. Walker HL, Ghani S, Kuemmerli C, et al. Reliability of medical information provided by
142 ChatGPT: Assessment against clinical guidelines and patient information quality instrument.
143 *Journal of Medical Internet Research*. 2023;25:e47479.
- 144 3. Guo LN, Lee MS, Kassamali B, Mita C, Nambudiri VE. Bias in, bias out: Underreporting and
145 underrepresentation of diverse skin types in machine learning research for skin cancer
146 detection—a scoping review. *J Am Acad Dermatol*. 2022;87(1):157-159.
- 147 4. Groh M, Harris C, Soenksen L, et al. Evaluating deep neural networks trained on clinical
148 images in dermatology with the fitzpatrick 17k dataset. . 2021:1820-1828.
- 149 5. Passby L, Jenko N, Wernham A. Performance of ChatGPT on specialty certificate
150 examination in dermatology multiple-choice questions. *Clin Exp Dermatol*. 2023:llad197.
- 151