

## Introduction

Conversational artificial intelligence (AI) language models like ChatGPT have emerged as promising tools for patients seeking medical information and guidance.<sup>1</sup> Previous studies in dermatological machine-learning have highlighted that the underrepresentation of diverse skin types in research could lead to bias and reduced performance in evaluating skin lesions in darker skin tones.<sup>2</sup> This study aims to assess GPT-4's accuracy in generating differential diagnoses and correct diagnoses for common skin lesions, while also examining differences in diagnostic accuracy between darker and lighter skin tones.

## Methods

Fifty images were randomly selected from the Fitzpatrick 17k dataset.<sup>3</sup> Half of the images selected represented darker skin tones, Fitzpatrick IV-VI, and the other half represented lighter skin tones, Fitzpatrick I-II. For each selected dermatological condition, GPT-4 was presented with pairs of images - one from a lighter skin tone and another from a darker skin tone. GPT-4 was then asked to provide its top three differential diagnoses and a final diagnosis for each pair. The responses generated by GPT-4 were transcribed and compared against the labels provided in the dataset to evaluate accuracy. Subsequently, a univariate linear regression analysis was conducted to investigate the relationship between Fitzpatrick skin type and diagnostic accuracy of GPT-4.



## Results

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

## Discussion

GPT-4's exhibited significantly lower overall accuracy compared to previous studies reporting accuracies as high as 90%.<sup>4</sup> This discrepancy highlights GPT-4's potential limitations in providing accurate information without sufficient clinical context. It is important to note that this study is limited by its relatively small sample size. If GPT-4 is to be considered for use by patients in a clinical setting, it is important to ensure that it demonstrates high accuracy and remains unbiased across all patient demographics and skin types.

## References

1. Shah YB, Ghosh A, Hochberg AR, et al. Comparison of ChatGPT and traditional patient education materials for men's health. *Urology Practice*. 2024;11(1):87-94.
2. Guo LN, Lee MS, Kassamali B, Mita C, Nambudiri VE. Bias in, bias out: Underreporting and underrepresentation of diverse skin types in machine learning research for skin cancer detection—a scoping review. *J Am Acad Dermatol*. 2022;87(1):157-159.
3. Groh M, Harris C, Soenksen L, et al. Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. *J Am Acad Dermatol*. 2021;85(4):1820-1828.
4. Passby L, Jenko N, Wernham A. Performance of ChatGPT on specialty certificate examination in dermatology multiple-choice questions. *Clin Exp Dermatol*. 2023;49(1):197.