Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Evaluating four consumer artificial intelligence chatbots on smartphone dermatology images: diagnostic accuracy, the contribution of clinical context, and skin-tone equity

Kevin Tabatabaei

tabatabaei.kevin@mayo.edu

Mayo Clinic in Florida

Fatmah AlZahrani

University of British Columbia

Matthew R. Hall

Mayo Clinic in Florida

Research Article

Keywords: Artificial intelligence, Machine learning, Smartphone, Dermatology, Skin diseases

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10683647/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Patients increasingly upload smartphone photographs to consumer artificial intelligence (AI) chatbots seeking dermatological diagnoses, yet accuracy on such images is unexplored. We compared the diagnostic accuracy and equity of four leading AI chatbots across varying input conditions. GPT-5.4-Pro, Claude-Opus-4.6, Gemini-3.1-Pro and Grok-4.20-Reasoning were evaluated on 3,005 cases from the crowdsourced Skin Condition Image Network (SCIN) dataset with physician-validated ground truth. Each case was assessed under four conditions: single image (n = 1,073), multiple images (n = 1,932), images plus clinical context (n = 2,567) and text-only (n = 2,567), using top-1 and top-3 hierarchical matching. In the primary analysis (images plus context), models clustered into two tiers: Grok (53.2%, 95% CI 51.3–55.1) and Gemini (52.6%, 95% CI 50.7–54.5) significantly outperformed GPT (45.5%, 95% CI 43.5–47.5) and Claude (44.6%, 95% CI 42.7–46.5) for top-1 accuracy (Cochran's Q, $p < 0.001$). Gemini led top-3 accuracy at 79.3% (95% CI 77.7–80.8). Adding clinical context improved accuracy for Grok (+ 5.5%, $p < 0.001$) and GPT (+ 3.8%, $p = 0.003$) only, whereas removing images caused losses of 11 to 18 percentage points across all models. Inter-model agreement was low across image-based analyses (Fleiss' kappa 0.34–0.41). For top-1 accuracy, Fitzpatrick skin tone disparity was significant only for Grok, the best-performing model (56.0% light versus 45.1% dark skin, $p = 0.007$), whereas top-3 disparities were significant for Claude, Gemini and Grok. Consumer AI chatbots show promise, but skin tone disparities, low inter-model agreement and content moderation limitations indicate these tools are not ready for standalone clinical use.

Introduction

The use of computing to analyze natural language (i.e., Natural Language Processing - "NLP") has been documented since as early as the 1970s.[4] It was only in the 2010s where the domain of NLP began leveraging deep learning techniques to make exponential advances, culminating in consumer-facing conversational large-language-models (LLMs) by late-2022 with the public launch of OpenAI's ChatGPT 3.5.[4, 10, 12]

Not too long after, OpenAI enabled "multi-modal" capabilities for ChatGPT, announcing that it "can now see, hear, and speak" in September 2023. This meant that, in addition to text, ChatGPT could interact with images and many other file formats.[12, 14] Other rivals soon followed, with top competitors such as Google, Anthropic, and xAI soon enabling multi-modal capabilities in their own models: Gemini, Claude, and Grok, respectively.[1, 5, 8]

With the ability of consumer artificial intelligence (AI) chatbots to "see," it is anticipated that patients may turn to these chatbots to get a first opinion on dermatological conditions that they may be experiencing.[13, 17]

Prior dermatology literature has largely focused on the multi-modal capabilities of ChatGPT in taking medical licensure assessments (e.g., Speciality Certificate Examination in Dermatology or the United States Medical Licensing Examination [USMLE]) or looking exclusively at dermoscopic images of

conditions (specifically melanomas and other cancerous lesions).[3, 13, 16–17] This leaves three notable gaps: First, current literature focuses on a narrow range of dermatological conditions, with the overwhelming focus on melanomas and cancerous lesions. Second, it almost entirely ignores ChatGPT's main rivals (Google's Gemini, Anthropic's Claude, and xAI's Grok). Last, it focuses on professional photography or dermoscopic images, neglecting the imperfections inherent in the smartphone photos that users actually upload.

Furthermore, existing studies have not examined how diagnostic accuracy changes when models are given varying amounts of visual and clinical information, nor have they assessed whether these AI models perform equitably across different skin tones. The role of clinical context (e.g., patient age, symptoms, anatomical location) in augmenting or substituting for visual input also remains unexplored.

Consequently, this article aims to address these gaps by exploring the diagnostic accuracy of the leading consumer AI models, specifically in the setting of representative real-world "smartphone" images. The four leading flagship consumer models, OpenAI's GPT-5.4-Pro, Anthropic's Claude-Opus-4.6, Google's Gemini-3.1-Pro, and xAI's Grok-4.20-Reasoning, will be compared in assessing a novel smartphone-style dermatology dataset, the Skin Condition Image Network (SCIN) dataset by Google Research and Stanford Medicine.[7] We further evaluate the effect of image quantity, the addition of clinical context, text-only (no image) performance, and diagnostic equity across Fitzpatrick skin types.

Methods

Dataset

The Skin Condition Image Network (SCIN) dataset was used. This dataset offers over 10,000 user-contributed, smartphone style images of dermatological conditions. It is a highly diverse dataset, featuring a strong representation of skin types and shades (over 50% of the dataset is at dermatologist estimated Fitzpatrick Skin Type (FST) 3 or above). It emphasizes eruptions, cutaneous infections, hair disorders, and vascular conditions, as opposed to the mainly malignant vs. pre-malignant focus seen in the majority of dermatology datasets in the literature (eg, HAM10000, ISIC 2020, PAD-UFES, etc.).[2, 7, 11, 15] Images are taken from different angles, with different devices, in different lighting conditions, etc., thus providing the "chaos" desired to mimic real-world scenarios.

The dataset is organized into "cases" which represent a patient and the associated skin concern. Each case may have one or more images and a dermatologist-provided weighted differential diagnosis of 1–5 conditions, with confidence values summing to one.

Image Filtering

Inclusion Criteria

- All cases with a differential diagnosis rendered by a dermatologist (n = 3,061)

Exclusion Criteria

- Cases where there was no differential diagnosis listed
- Cases flagged by the dermatologist as having insufficient image quality for grading (dermatologist_gradable_for_skin_condition = "NO_IMAGE_QUALITY_INSUFFICIENT")

Of the 5,033 cases in the SCIN dataset, 3,061 had a non-empty dermatologist-provided differential diagnosis. An additional 56 cases were excluded due to insufficient image quality as assessed by the dermatologist, yielding 3,005 eligible cases. Of these, 1,073 had a single image, 504 had two images, and 1,428 had three images.

Image Processing

Images were converted from Portable Network Graphics format (.PNG) to Joint Photographic Experts Group (.JPEG) at 95% quality to reduce the size of the images uploaded to the AI models. The conversion resulted in an average 85.29% reduction in file size at the same image dimensions. Images did not undergo any further modification (eg, dimensions, upscaling) before being uploaded to the AI models via their respective Application Programming Interfaces (APIs).

AI Model Evaluation

Four flagship consumer AI models were used in this evaluation:

- **GPT-5.4-Pro** (OpenAI)
- **Claude-Opus-4.6** (Anthropic)
- **Gemini-3.1-Pro** (Google)
- **Grok-4.20-Reasoning** (xAI)

Each model was provided the same identical images and was prompted to provide a differential diagnosis, both a top-1 diagnosis (i.e., the model's best guess) and a top-3 diagnosis (i.e., the model's top 3 guesses at the condition represented in the photos). The dermatologist provided predictions were treated as the "ground truth."

When assessing top-1 accuracy, a "hit" (success scenario) was considered when the AI model's prediction was found within the dermatologist-provided prediction. When assessing top-3 accuracy, a "hit" was considered when at least one of the AI model's top 3 guesses was found within the dermatologist-provided predictions.

Diagnostic matching used a hierarchical term-matching system with over 30 curated synonym groups to account for terminological variation between AI outputs and ground-truth labels (e.g., "urticaria" matched "hives," "tinea corporis" matched "ringworm").

Analysis Design

To isolate the contributions of visual and contextual information, each case was evaluated under four distinct conditions:

- **Analysis 1 (A1): Single Image.** Only the first image per case was provided, with no clinical metadata (n = 1,073).
- **Analysis 2 (A2): All Available Images.** All images for each case were provided, with no clinical metadata (n = 1,932 cases with 2 + images).
- **Analysis 3 (A3): Images Plus Context.** All images plus clinical metadata (age, sex, anatomical location, duration, and symptoms) were provided for the 2,567 cases with available metadata. This represents the most complete input and serves as the primary analysis.
- **Analysis 4 (A4): Text Only.** Only clinical metadata was provided, with no images (n = 2,567 cases with available metadata).

All models received the same standardized prompt requesting a structured response with a top-1 diagnosis and top-3 differential. For image-based analyses (A1, A2, A3), a prefix instructed the model to analyze the provided dermatological image(s). For text-only analysis (A4), the prompt presented clinical metadata without images. All four models received identical prompts for each analysis condition (**Online Resource 1**). Some consumer AI models enforce content moderation policies that may cause them to refuse to analyze certain medical images. Any such refusals were documented and the affected cases were excluded from accuracy calculations for the refusing model.

Statistical Analysis

Between-model comparisons used Cochran's Q test for omnibus differences among the four models, followed by pairwise McNemar's tests with Bonferroni correction. Within-model cross-analysis comparisons (e.g., effect of adding context or removing images) used paired McNemar's tests on overlapping case populations. Subgroup analyses by Fitzpatrick skin type, Monk skin tone, sex, and age used chi-square tests. Inter-model agreement was assessed with Fleiss' kappa. Significance was set at $p < 0.05$. All analyses were conducted in Python 3.11.

Results

Accuracy results are tabulated in Table 1. In the primary analysis (A3: images plus context, n = 2,567), Grok-4.20-Reasoning achieved the highest top-1 accuracy at 53.2% (95% confidence interval [CI] 51.3–55.1), followed by Gemini-3.1-Pro at 52.6% (95% CI 50.7–54.5). GPT-5.4-Pro achieved 45.5% (95% CI

43.5–47.5) and Claude-Opus-4.6 achieved 44.6% (95% CI 42.7–46.5). For top-3 accuracy, Gemini led at 79.3% (95% CI 77.7–80.8), followed by Grok (75.0%, 95% CI 73.3–76.6), GPT (73.7%, 95% CI 71.9–75.4), and Claude (72.5%, 95% CI 70.8–74.2). GPT gave predictions on 2,454 of 2,567 cases due to content filtering failures (4.4% failure rate).

Table 1
Diagnostic accuracy (%) across four analyses

Model	A1	A2	A3	A4
Top-1 Accuracy				
n	1,050*	1,846*	2,454*	2,561*
GPT-5.4-Pro	40.3 (37.4–43.3)	43.7 (41.5–46.0)	45.5 (43.5–47.5)	28.8 (27.1–30.6)
Claude-Opus-4.6	41.8 (38.8–44.7)	42.3 (40.1–44.5)	44.6 (42.7–46.5)	29.8 (28.0–31.6)
Gemini-3.1-Pro	44.2 (41.2–47.2)	52.7 (50.5–55.0)	52.6 (50.7–54.5)	34.4 (32.6–36.3)
Grok 4.20	45.9 (43.0–48.9)	49.3 (47.1–51.6)	53.2 (51.3–55.1)	42.0 (40.1–43.9)
Top-3 Accuracy				
n	1,050*	1,846*	2,454*	2,561*
GPT-5.4-Pro	67.6 (64.7–70.4)	70.0 (67.9–72.1)	73.7 (71.9–75.4)	59.7 (57.8–61.6)
Claude-Opus-4.6	67.8 (64.9–70.5)	70.9 (68.8–72.8)	72.5 (70.8–74.2)	55.7 (53.7–57.6)
Gemini-3.1-Pro	76.9 (74.3–79.3)	80.9 (79.1–82.6)	79.3 (77.7–80.8)	59.8 (57.9–61.7)
Grok 4.20	70.3 (67.5–72.9)	73.4 (71.4–75.3)	75.0 (73.3–76.6)	62.6 (60.7–64.5)

*A1 = single image; A2 = multiple images (2–3); A3 = images + context; A4 = text only. Values are % (95% Wilson CI). *GPT n is lower due to content policy refusals; other models analysed all cases in each analysis (A1: 1,073; A2: 1,932; A3: 2,567; A4: 2,567)*

Cochran's Q test confirmed significant between-model differences in both top-1 and top-3 accuracy across all four analyses (all $p < 0.002$). Pairwise McNemar's tests with Bonferroni correction revealed a consistent pattern: GPT and Claude did not differ significantly from each other in any analysis (all corrected $p > 0.05$), nor did Gemini and Grok (all corrected $p > 0.05$). However, both Gemini and Grok significantly outperformed both GPT and Claude in analyses A2, A3, and A4 (all corrected $p < 0.001$). In text-only analysis (A4), Grok further outperformed all other models, achieving 42.0% top-1 accuracy versus Gemini's 34.4%, Claude's 29.8%, and GPT's 28.8% (all pairwise corrected $p < 0.001$).

Effect of Image Count

Accuracy stratified by the number of images per case (Table 2) revealed that providing two images numerically improved top-1 accuracy over a single image for all models, though only Gemini's gain was statistically significant (+ 9.8 percentage points, from 44.2% to 54.0%, $p < 0.001$). A third image provided

no significant additional benefit for any model. Top-3 accuracy showed a similar pattern, with Gemini (76.9% to 82.3%, p = 0.02) and Grok (70.3% to 75.2%, p = 0.049) showing nominal gains from one to two images.

Table 2
Accuracy (%) by number of images per case (analyses 1 and 2)

Model	1 Image	2 Images	3 Images	P 1→2	P 2→3	P 1→3
GPT	40.3 / 67.6	44.9 / 70.5	43.3 / 69.9	0.10 / 0.29	0.58 / 0.85	0.15 / 0.25
Claude	41.8 / 67.8	42.7 / 69.0	42.2 / 71.5	0.78 / 0.65	0.89 / 0.33	0.87 / 0.048*
Gemini	44.2 / 76.9	54.0 / 82.3	52.3 / 80.4	< 0.001* / 0.02*	0.56 / 0.37	< 0.001* / 0.04*
Grok	45.9 / 70.3	49.8 / 75.2	49.2 / 72.8	0.17 / 0.049*	0.84 / 0.31	0.12 / 0.19

*1-image data from A1 (single image, no context); 2- and 3-image data from A2 (multiple images, no context). Top-1 / top-3 hierarchical accuracy shown. P values from pairwise chi-square tests. *Statistically significant (p < 0.05)*

Effect of Clinical Context

Adding clinical metadata (age, sex, anatomical location, duration, symptoms; availability detailed in **Online Resource 2, Table S1**) to images produced variable effects on top-1 accuracy. Grok showed the largest significant improvement (+ 5.5 percentage points, p < 0.001 by McNemar's test), followed by GPT (+ 3.8%, p = 0.003). Claude (+ 1.7%, p = 0.15) and Gemini (-0.3%, p = 0.85) showed no significant change. Top-3 accuracy showed similar variability: GPT (+ 4.3%), Grok (+ 1.9%), Claude (+ 1.9%), and Gemini (-2.0%). These results suggest that the benefit of clinical context is model-dependent rather than universally negligible.

Image vs. Text-Only Performance

Removing images while retaining clinical metadata (A3 to A4, same overlapping cases) resulted in significant accuracy drops for all models. Top-1 losses ranged from 11.3 (Grok) to 18.2 (Gemini) percentage points (all p < 0.001). Top-3 losses were similarly pronounced: Gemini lost 19.5 percentage points, Claude lost 16.9, GPT lost 13.6, and Grok lost 12.4 (all p < 0.001). Grok demonstrated the smallest decrement in both metrics and the highest text-only top-1 accuracy (42.0%), suggesting stronger utilization of clinical metadata.

Equity Analysis

Accuracy in Analysis 3 stratified by dermatologist-assessed Fitzpatrick skin type (binned as light [I-II], medium [III-IV], and dark [V-VI]) revealed significant disparities for some models (Table 3). For top-1 accuracy, only Grok showed a significant disparity between light and dark skin (56.0% vs. 45.1%, a 10.9-

point difference, p = 0.007). Claude (6.7-point gap, p = 0.14), GPT (4.3-point gap, p = 0.37), and Gemini (-1.1-point gap, p = 0.55) did not reach significance. For top-3 accuracy, skin tone disparities were significant for Grok (8.3-point gap, p = 0.006), Claude (7.3-point gap, p = 0.004), and Gemini (4.2-point gap, p = 0.03), but not GPT (1.9-point gap, p = 0.53). Claude showed a significant sex-based disparity in top-1 accuracy (42.8% female vs. 49.8% male, p = 0.01).

Table 3
Accuracy (%) by Fitzpatrick skin type (binned), analysis 3

Model	Light (I-II)	Medium (III-IV)	Dark (V-VI)	P
Top-1 Accuracy				
GPT	45.1	46.3	40.8	0.37
Claude	46.4	43.6	39.7	0.14
Gemini	53.4	51.3	54.4	0.55
Grok	56.0	51.9	45.1	0.007*
Top-3 Accuracy				
GPT	74.8	72.7	72.8	0.53
Claude	75.9	70.2	68.6	0.004*
Gemini	81.7	77.3	77.5	0.03*
Grok	77.9	73.2	69.6	0.006*

*Top-1 / top-3 hierarchical accuracy shown. P values from chi-square test. *Statistically significant (p < 0.05). GPT n is slightly lower per group due to content policy refusals*

Inter-Model Agreement

Fleiss' kappa for top-1 agreement among all four models ranged from 0.34 (A1 and A2) to 0.52 (A4), indicating fair to moderate agreement. In Analysis 3, kappa was 0.41. At least one model produced a correct top-1 diagnosis in 74.9% of cases, while all four agreed on the correct diagnosis in only 21.7% of cases. All four were incorrect in 25.1% of cases. The low agreement across image-based analyses suggests complementary error patterns among models.

Content Policy Refusals

GPT was the only model to exhibit query failures, at rates of 2.1% in A1 (23/1,073), 4.5% in A2 (86/1,932), 4.4% in A3 (113/2,567), and 0.2% in A4 (6/2,567), attributed to content policy filters for image-based queries. These cases were excluded from accuracy calculations.

Condition-Level Performance

Among the 20 most common ground-truth conditions in Analysis 3 (**Online Resource 2, Table S2**), notable variation was observed. Urticaria was among the most accurately diagnosed conditions (Claude 84%, Gemini 84%, Grok 71%, GPT 68%). Models performed poorly on pigmented purpuric eruption (Claude 24%, Gemini 19%, GPT 9%, Grok 9%) and impetigo (Grok 44%, Gemini 39%, GPT 35%, Claude 24%). Eczema, the most common condition, showed moderate accuracy with Grok leading at 70% versus GPT and Claude at approximately 49%.

Discussion

Our analysis demonstrated that while current state-of-the-art consumer AI models have promising performance on diagnosing patient-uploaded smartphone-style dermatological images, there is still much room for improvement. Patients should not completely trust their AI-provided diagnoses, especially due to the poor top-1 performance exhibited by all models. Moreover, while an interesting metric, top-3 predictions may not be of much utility to patients, as they might not have the ability to discern between the three provided diagnoses without clinical support from a trained dermatologist. The smartphone-style photos used in this study are more representative of real-world scenarios and go beyond the professional dermoscopic images used in prior research. Furthermore, the strong representation of eruptions, cutaneous infections, hair disorders, and vascular conditions in the SCIN dataset aligns well with the "bread-and-butter" of what patients deal with on a daily basis,[9] and would thus be more likely to upload to an AI chatbot.

A striking finding is the consistent performance clustering into two tiers: GPT and Claude performed statistically equivalently to one another, as did Gemini and Grok, but the latter pair significantly outperformed the former across all image-based analyses. This consistent two-tier pattern has practical relevance for clinicians advising patients on AI tool selection, as not all commercially available chatbots perform equivalently in dermatological assessment. Notably, only Gemini significantly benefited from additional images, suggesting that most models extract the majority of diagnostic information from a single representative photograph, a reassuring finding for patients uploading limited images.

The effect of clinical context was model-dependent. Grok and GPT showed significant accuracy gains when metadata was added to images, while Gemini and Claude did not. This variability has practical implications: clinicians cannot assume that providing more history to an AI tool will uniformly improve its output. In contrast, the removal of images caused accuracy declines of 11 to 18 percentage points across all models, underscoring the primacy of visual information in dermatologic diagnosis.

The equity analysis raises concerns for clinical deployment. Grok, despite being the top performer overall, was the only model with a significant top-1 skin tone disparity (10.9-percentage-point gap between light and dark Fitzpatrick skin types, $p = 0.007$). Gemini demonstrated the most equitable performance across skin tones. For top-3 accuracy, disparities were more widespread, with Grok, Claude, and Gemini all showing significant gaps. These disparities highlight the need to evaluate AI fairness before deployment in diverse patient populations. The low inter-model agreement (Fleiss' kappa 0.34 to

0.41 for image-based analyses) means that different chatbots frequently disagree on the same case, further cautioning against reliance on any single AI tool.

These findings have important implications for healthcare professionals as patients increasingly build trust in AI chat tools for their healthcare concerns. Clinicians should be aware that patients could be using tools like ChatGPT to seek preliminary diagnoses, and should counsel patients about the high propensity of AI to get diagnoses incorrect, especially with haphazardous smartphone images.

Limitations

This study has several limitations. It is unclear why GPT refused to analyze 2.1 to 4.5% of image-based cases, as there were no readily apparent patterns in the types of cases refused. Additionally, the lack of transparency regarding training sources makes it impossible to verify if any of these models were already trained on SCIN data. The SCIN dataset relies on weighted multi-label ground truth from dermatologist assessments rather than biopsy-confirmed diagnoses, which may introduce labeling variability, and is limited to U.S.-based adults. A single standardized prompt was used; prompt optimization for individual models could yield different results. Hierarchical term matching may still under- or over-count accuracy for ambiguous diagnoses. API model versions may change over time, and results may not be reproducible with future updates.

Conclusion

Four leading consumer AI chatbots achieved top-1 diagnostic accuracies of 44.6% to 53.2% on smartphone-style dermatological images, clustering into two consistent performance tiers. Visual input contributed substantially more diagnostic signal than clinical metadata, and the benefit of added context was model-dependent. Persistent skin tone disparities in the highest-performing model, low inter-model agreement, and content moderation failures indicate that these tools are not ready for standalone diagnostic use. Equity-stratified evaluation should be a prerequisite for any patient-facing deployment of general-purpose multimodal models in dermatology.

Declarations

The authors describe no conflicts of interest. This research received no funding.

Competing interests

The authors declare that they have no competing interests.

Ethics approval

This study used only publicly available, de-identified data from the Skin Condition Image Network (SCIN) dataset, which was originally collected with informed consent under an institutional review board-

approved study at Stanford Medicine. As the present analysis used only the public dataset, no additional institutional review board review was required.

Consent to participate

Not applicable. This study used previously collected, publicly available, de-identified data; no new participants were enrolled.

Consent to publish

Not applicable. No identifiable patient information or images are reproduced in this manuscript.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author Contribution

Kevin Tabatabaei: conceptualization, data curation, formal analysis, investigation, methodology, software, visualization, writing - original draft, writing - review and editing. Fatmah AlZahrani: supervision, writing - review and editing. Matthew R. Hall: supervision, writing - review and editing.

Data Availability

The Skin Condition Image Network (SCIN) dataset analysed in this study is publicly available at https://github.com/google-research-datasets/scin. Analysis code is available from the corresponding author on reasonable request.

References

1. Carolan K, Fennelly L, Smeaton AF (2024) A review of multi-modal large language and vision models. arXiv. https://doi.org/10.48550/arXiv.2404.01322
2. Cassidy B, Kendrick C, Brodzicki A et al (2022) Analysis of the ISIC image datasets: usage, benchmarks and recommendations. Med Image Anal 75:102305. https://doi.org/10.1016/j.media.2021.102305
3. Fan KS, Fan KH (2024) Dermatological knowledge and image analysis performance of large language models based on the Specialty Certificate Examination in Dermatology. Dermato 4:124–135. https://doi.org/10.3390/dermato4040013

4. Fanni SC, Febi M, Aghakhanyan G, Neri E (2023) Natural language processing. In: Klontzas ME, Fanni SC, Neri E (eds) Introduction to artificial intelligence. Springer, Cham, pp 87–99
5. Gemini Team, Anil R, Borgeaud S et al (2023) Gemini: a family of highly capable multimodal models. https://doi.org/10.48550/arXiv.2312.11805. arXiv
6. Google R (2024) Skin Condition Image Network (SCIN) dataset. GitHub. https://github.com/google-research-datasets/scin
7. Jeong Y, Schaekermann M, Lin S (2024) Crowdsourcing dermatology images with Google search ads: creating a real-world skin condition dataset for AI development. Ann Fam Med 22(Suppl 1):6186. https://doi.org/10.1370/afm.22.s1.6186
8. Kaczmarczyk R, Wilhelm TI, Martin R, Roos J (2024) Evaluating multimodal AI in medical diagnostics. npj Digit Med 7:205. https://doi.org/10.1038/s41746-024-01208-3
9. Li Q, Patrick MT, Sreeskandarajan S et al (2024) Large-scale epidemiological analysis of common skin diseases to identify shared and unique comorbidities and demographic factors. Front Immunol 14:1309549. https://doi.org/10.3389/fimmu.2023.1309549
10. Otter DW, Medina JR, Kalita JK (2021) A survey of the usages of deep learning for natural language processing. IEEE Trans Neural Netw Learn Syst 32:604–624. https://doi.org/10.1109/TNNLS.2020.2979670
11. Pacheco AGC, Lima GR, Salomao AS et al (2020) Data Brief 32:106221. https://doi.org/10.1016/j.dib.2020.106221. PAD-UFES-20: a skin lesion dataset composed of patient data and clinical images collected from smartphones
12. Ray PP (2023) ChatGPT: a comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet Things Cyber-Physical Syst 3:121–154. https://doi.org/10.1016/j.iotcps.2023.04.003
13. Shifai N, van Doorn R, Malvehy J, Sangers TE (2024) Can ChatGPT vision diagnose melanoma? An exploratory diagnostic accuracy study. J Am Acad Dermatol 90:1057–1059. https://doi.org/10.1016/j.jaad.2023.12.062
14. Temsah R, Altamimi I, Alhasan K et al (2023) Healthcare's new horizon with ChatGPT's voice and vision capabilities: a leap beyond text. Cureus 15:e47469. https://doi.org/10.7759/cureus.47469
15. Tschandl P, Rosendahl C, Kittler H (2018) The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci Data 5:180161. https://doi.org/10.1038/sdata.2018.161
16. Yamamura Y, Fujii K, Nakashima C, Otsuka A (2025) Evaluation of the accuracy of artificial intelligence (AI) models in dermatological diagnosis and comparison with dermatology specialists. Cureus 17:e77067. https://doi.org/10.7759/cureus.77067
17. Zhang Y, Chung J, Chen R et al (2025) Testing the ability of artificial intelligence chatbots and dermatologic diagnosis mobile applications to diagnose melanoma and seborrheic keratosis. J Am Acad Dermatol 92:161–163. https://doi.org/10.1016/j.jaad.2024.08.077

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementaryastandardizedprompts.docx
- supplementarybtablesS1S2.docx