

Performance of Consumer Multimodal Large Language Model Interfaces on a Selected Publication-derived Challenge Set of Initially Misdiagnosed Dermatologic Emergencies

✉ Bengisu Meral Ketenci¹, ✉ Demet Acar²

¹Konya City Hospital, Clinic of Dermatology, Konya, Türkiye

²Konya City Hospital, Clinic of Emergency Medicine, Konya, Türkiye

Abstract

Aim: To evaluate the diagnostic and safety-related response performance of consumer multimodal large language model interfaces on a selected publication-derived challenge set of initially misdiagnosed dermatologic emergencies.

Materials and Methods: This exploratory cross-sectional study evaluated 50 cases of initially misdiagnosed dermatologic emergencies, selected from published reports identified between May 1 and June 1, 2026. ChatGPT, Gemini, and Claude were tested on June 2, 2026, under text-only and text-plus-image conditions using a single standardized prompt. Outcomes were Top-1 and cumulative Top-4 accuracies, clinically appropriate emergency management, risk-triage failure, and harmful recommendations.

Results: Top-1 accuracy ranged from 32.0% to 50.0% with text-only input and from 36.0% to 58.0% with text-plus-image input; corresponding Top-4 ranges were 50.0%-66.0% and 56.0%-74.0%. Appropriate management ranged from 38.0% to 54.0% and from 44.0% to 64.0%. No within-platform input-mode comparisons remained significant after Holm correction. Between-platform differences remained significant for Top-1 accuracy under both input conditions and for Top-4 accuracy and management appropriateness under text-plus-image input. Post-hoc differences primarily reflected lower observed performance for Claude than for Gemini in this dataset. Appropriate management occurred in 100% of Top-1-correct responses compared with 9.9% of Top-1-incorrect responses, whereas 18.0% of Top-4-correct responses had inappropriate management.

Conclusion: Within this selected challenge set and a single standardized prompting condition, the platforms demonstrated moderate diagnostic performance and variable performance on safety-related responses. Correct Top-1 diagnosis closely corresponded to appropriate management, whereas Top-4 recognition did not ensure safe management. The findings represent a single-run technical benchmark and do not establish output reproducibility, clinical usefulness, real-world safety, or stable platform superiority.

Keywords: Dermatologic emergencies, large language models, artificial intelligence, diagnostic accuracy, patient safety, emergency management

Introduction

Dermatologic emergencies represent a diagnostically challenging subset of skin diseases in which delayed recognition or inappropriate initial management may lead to severe morbidity, permanent sequelae, or mortality (1). Although many dermatologic complaints are managed in outpatient settings, certain presentations require urgent evaluation and rapid intervention (2). Severe cutaneous adverse reactions, necrotizing

soft tissue infections, disseminated viral eruptions, purpura fulminans, erythroderma, acute autoimmune blistering diseases, and airway-threatening angioedema are examples of conditions in which early clinical suspicion directly affects patient outcomes (1,2).

The diagnosis of dermatologic emergencies is often complicated by overlapping clinical morphology (1). Life-threatening diseases may initially resemble common benign conditions such as viral



Corresponding Author: Bengisu Meral Ketenci MD, Konya City Hospital, Clinic of Dermatology, Konya, Türkiye

E-mail: merabengisu@gmail.com **ORCID ID:** orcid.org/0009-0007-5339-1681

Cite this article as: Ketenci BM, Acar D. Performance of consumer multimodal large language model interfaces on a selected publication-derived challenge set of initially misdiagnosed dermatologic emergencies. Eurasian J Emerg Med. 2026;25: 456-66

Received: 15.07.2026

Accepted: 22.09.2026

Published: 09.10.2026



©Copyright 2026 The Author(s). The Emergency Physicians Association of Turkey / Eurasian Journal of Emergency Medicine published by Galenos Publishing House. Licensed by Creative Commons Attribution-NonCommercial-NoDerivatives (CC BY-NC-ND) 4.0 International License

exanthems, urticaria, cellulitis, eczema, or uncomplicated drug eruptions (3). Conversely, some benign dermatoses may appear alarming at first presentation (2). This diagnostic overlap creates a risk of both delayed treatment and inappropriate triage (4). Therefore, in emergency dermatology, accurate diagnosis alone is not sufficient; clinicians must also recognize disease severity, identify red flags, recommend appropriate investigations, and determine whether urgent referral, hospitalization, or multidisciplinary consultation is required (1,5).

Large language models (LLMs) have been investigated as potential clinical decision-support technologies (6). With the development of multimodal systems, these models can process both textual clinical information and visual data, including clinical photographs (7). This capability is relevant to dermatology, where diagnostic reasoning depends on integrating lesion morphology, distribution, symptom history, medication exposure, systemic findings, and visual pattern recognition (8). Previous research has largely focused on diagnostic performance using clinical text, images, or combined text-image inputs (9,10). However, performance on retrospective benchmarks should be distinguished from reproducibility, clinical usefulness, and safety in real-world implementation.

Beyond diagnostic classification, dermatologic emergencies require recognition of disease severity, red flags, and the need for urgent escalation. A model may include the correct diagnosis in its differential list yet provide unsafe reassurance, omit hospitalization, or suggest inappropriate treatment (10). Therefore, safety-related response endpoints such as management appropriateness, risk-triage failures, and harmful recommendations can complement diagnostic accuracy when benchmarking model outputs. These response-based endpoints do not, however, constitute evidence of clinical safety in actual implementation.

This study aimed to benchmark the diagnostic and safety-related response performance of three consumer multimodal LLM interfaces using a selected, publication-derived challenge set comprising 50 initially misdiagnosed dermatologic emergency cases. Each case was assessed under paired text-only and text-plus-image conditions using a single standardized prompt. The study directly evaluated performance on this selected case set but did not evaluate output reproducibility, clinical effectiveness as a decision-support tool, or safety in real-world implementation.

Materials and Methods

Study Design

This exploratory cross-sectional comparative study benchmarked the diagnostic and safety-related response performance of three

consumer multimodal LLM interfaces using a selected publication-derived challenge set comprising initially misdiagnosed dermatologic emergencies. The dataset comprised 50 cases identified from published case reports and case series using the literature search and eligibility assessment process described below. All included cases had been initially misdiagnosed, and the final diagnoses had been confirmed by histopathology, microbiology, laboratory testing, imaging, genetic analysis, or accepted clinical criteria.

The literature search and dataset construction were conducted between May 1 and June 1, 2026, and model evaluations were performed on June 2, 2026. Each case was assessed under two separately administered, paired input conditions: text-only and text-plus-image. Outcomes included Top-1 accuracy, cumulative Top-4 accuracy, clinically appropriate emergency management, risk-triage failure, and harmful recommendations.

Case Selection and Source-publication Provenance

A structured literature search was conducted in PubMed/MEDLINE, Embase, and Google Scholar between May 1 and June 1, 2026, without a publication-date restriction and was limited to English-language publications. Search terms included the following: dermatologic emergencies, specific urgent conditions, diagnostic errors, case reports, and case series. The complete search strategy is provided in Supplementary Table S1.

Eligible publications were required to report at least one case with an explicitly documented initial misdiagnosis, a supported final diagnosis, sufficient clinical information, and at least one suitable clinical image. An emergency or urgent presentation was defined as an acute or rapidly progressive cutaneous or mucocutaneous condition in which delayed management could result in death, airway or organ compromise, severe deterioration, irreversible damage, or the need for urgent hospital-based intervention. Publications not meeting these criteria were excluded.

The complete selection process and the numerical reasons for exclusion are presented in Figure 1. The initial search was performed at the publication level and included both case reports and case series as potentially eligible publication types. Full-text eligibility assessment was also performed at the publication level. All 50 publications that ultimately satisfied the eligibility criteria were single-patient case reports; therefore, each included publication contributed exactly one case to the final dataset ($n=50$ publications; $n=50$ cases). Case-level source and provenance information is provided in Supplementary Table S2. Eligibility was independently assessed by a board-certified dermatologist and an emergency medicine specialist (Cohen's $\kappa=0.83$), with disagreements resolved by consensus.

Disease Categories

Cases were classified into six mutually exclusive categories: severe cutaneous adverse reactions (n=12), infectious emergencies (n=11), vascular or vasculitic emergencies (n=9), autoimmune blistering emergencies (n=8), urticaria/angioedema disorders (n=6), and other dermatologic emergencies (n=4).

Artificial Intelligence Models

The default free consumer web versions of ChatGPT, Gemini, and Claude were evaluated on June 2, 2026, from Türkiye via their official web interfaces. Testing was conducted in private (incognito) browser sessions without selecting a named model or using paid subscription features. Because the consumer

interfaces did not provide verifiable backend metadata, the exact backend model versions were unavailable and the systems were identified only by their platform names.

API access, plugins, fine-tuning, retrieval-augmented generation, automated web search, externally connected tools, and custom system instructions were not used. No explicit rate-limit or fallback-model notification was encountered during testing.

A total of 300 model-case-input-condition evaluations were performed. Each evaluation was conducted on a separate, newly opened page and in an independent conversation to prevent information carryover between cases or input conditions. Text-only and text-plus-image evaluations of the same case were performed separately; each condition was queried once.

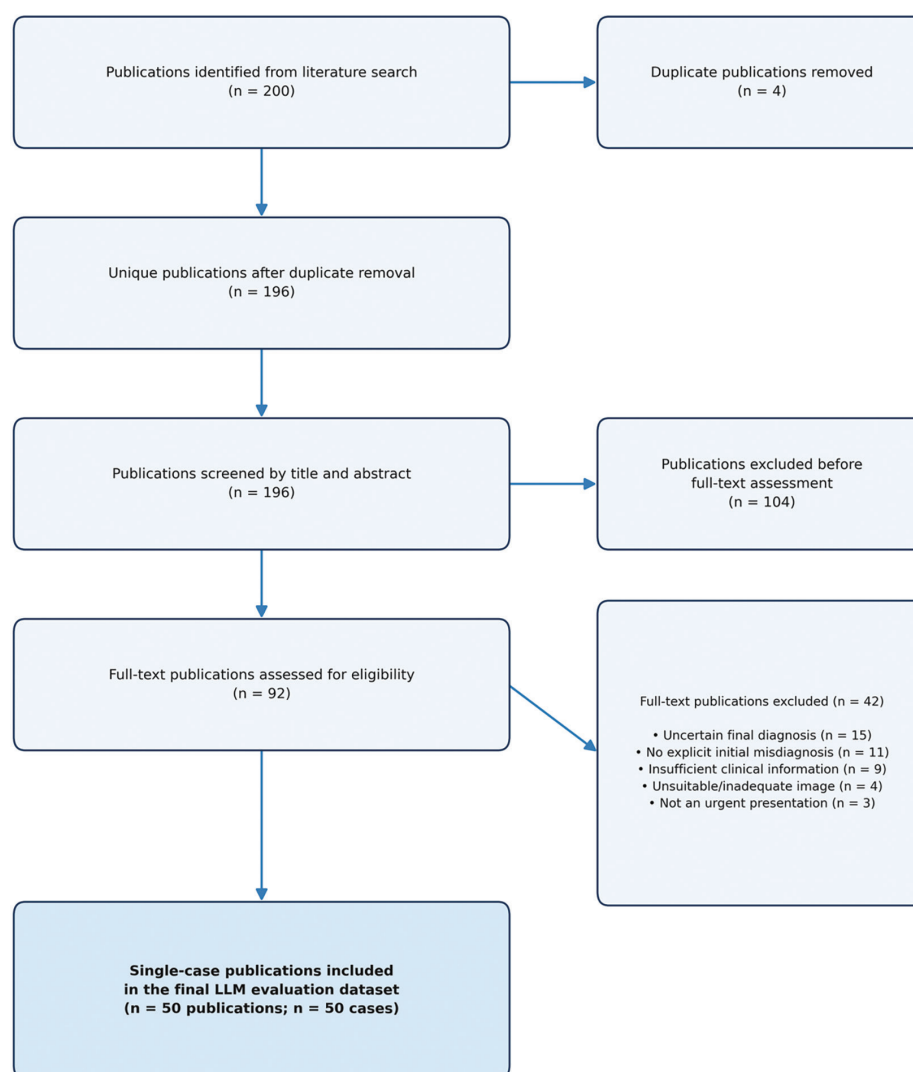


Figure 1. Flow diagram of the literature search and publication selection process. Publications were screened at the title/abstract and full-text stages. All 50 publications included in the final LLM evaluation dataset were single-patient case reports; therefore, the final dataset comprised 50 publications corresponding to 50 cases

LLM: Large language model

Prompting Protocol

In the text-only condition, each model received a clinical vignette prepared using the standardized construction framework and containing demographic, historical, systemic, and physical examination findings.

In the text-plus-image condition, the same prepared vignette and the clinical image from the original publication were provided in a separate conversation.

The standardized prompt was:

“You are an expert emergency physician. Analyze the following patient presenting to the emergency department with a dermatologic complaint using the provided clinical history (and image). Provide the most likely diagnosis, three differential diagnoses ranked in order of probability, and the recommended clinical approach. Use medical terminology.”

Clinical Vignette Construction

Clinical vignettes were manually prepared from the source publications using a standardized construction framework. The investigators aimed to include information considered to be available at the initial clinical presentation, including demographic characteristics, presenting symptoms and their duration, relevant medical and medication history, systemic symptoms, vital signs, and physical and dermatologic examination findings. The standardized construction framework specified that final diagnoses, diagnostic-confirmation findings, and post-diagnostic laboratory, histopathological, microbiological, imaging, therapeutic response, and follow-up information were not included in the vignettes. The initial incorrect diagnosis was not intentionally presented as a diagnostic label. The source material was paraphrased and summarized, and publication titles, citations, figure captions, annotations, and direct source-identifying statements were not included in the prompts. A consistent set of clinical fields was sought, although no fixed word-count requirement was applied, because the completeness of the published reports varied.

Response Evaluation

Responses were independently assessed by a dermatologist and an emergency medicine specialist, who were blinded to each other's scores during the initial evaluation phase. After completion of independent scoring, all discordant ratings were jointly reviewed and resolved through structured discussion and consensus. The final consensus ratings, rather than either assessor's individual ratings, were used in the main analyses. No third assessor was involved, and neither assessor's rating was used as a default decision.

The final diagnosis reported in the source publication was used as the diagnostic reference standard. Emergency management was evaluated according to current literature, clinical guidelines, and established practice principles.

Diagnostic Accuracy

Top-1 accuracy was defined as the first-ranked diagnosis being identical to or clinically equivalent to the reference diagnosis.

Top-4 accuracy was defined as a cumulative measure indicating that the reference diagnosis was included either as the most likely diagnosis or as one of the three ranked differential diagnoses. Therefore, a correct Top-1 response was also considered correct in the Top-4.

Patient-safety Outcomes

Risk-triage failure was defined as the failure to recognize or communicate urgency, including underestimating severity, omitting urgent referral or hospitalization, or failing to identify key red flags.

A harmful recommendation was defined as an active suggestion that was incorrect, contraindicated, falsely reassuring, potentially harmful, or capable of delaying urgent care.

These outcomes were not mutually exclusive. Assessment criteria were determined on a disease-specific basis. Relevant Turkish national treatment guidelines were prioritized when available. In their absence, widely accepted international clinical practice guidelines or consensus statements were used. When no applicable guideline or consensus statement was available, the relevant literature and the clinical judgment and experience of the dermatologist and the emergency medicine specialist were considered. The two assessors scored the responses independently, and disagreements were resolved by consensus.

Clinically Appropriate Emergency Management

Clinically appropriate emergency management was evaluated using a two-step approach. For brevity, this outcome is referred to as “management appropriateness” throughout the manuscript. Responses containing a risk-triage failure or a harmful recommendation were classified as inappropriate. Responses without these safety violations were assessed for adequate recognition of urgency, referral or hospitalization, specialist consultation, initial treatment, diagnostic work-up, and safety warnings.

Ethical Considerations

The literature search and dataset construction were conducted between May 1 and June 1, 2026, and model evaluations were performed on June 2, 2026. The study used only anonymized

clinical information and images obtained from published case reports and case series. No patient intervention was performed, no new clinical data were collected, no identifiable personal data were used, and no biological material sampling was performed.

Because the study was based on a secondary analysis of previously published anonymized data and did not involve the collection of new patient data or the inclusion of identifiable personal information, it was considered a study that did not require ethics committee approval or patient informed consent.

All case information was reviewed prior to analysis to remove elements that could identify patients. The study was conducted in accordance with the 2013 revision of the Declaration of Helsinki, ICMJE recommendations, and COPE guidelines. Clinical images obtained from original publications were used with appropriate citation of the relevant sources and in accordance with the licensing and copyright conditions of the publications.

Before submission to the model interfaces, clinical images were reviewed and cropped to include only the dermatologic lesion where possible. Faces, tattoos, unique markings, personal identifiers, publication captions, labels, arrows, and watermarks were removed when present. The images were processed using default consumer web interfaces in private (incognito) browser sessions.

Statistical Analysis

Top-1 accuracy and clinically appropriate emergency management were designated as primary outcomes, whereas cumulative Top-4 accuracy, risk-triage failure, and harmful recommendation were designated as secondary outcomes. This reporting hierarchy was not preregistered.

All 50 eligible cases identified during dataset construction were included; therefore, no formal a priori power calculation was performed. With 50 cases, a proportion of 50% corresponds to a Wilson 95% confidence interval (CI) of 36.6%-63.4%, indicating a maximum precision of approximately ± 13.4 percentage points.

Outcomes were summarized as n/N (%) with Wilson 95% CIs. Within each platform, text-only and text-plus-image conditions were compared using two-sided exact McNemar tests. Paired risk differences, calculated as text-plus-image minus text-only, were reported with 95% CIs calculated using the Newcombe hybrid-score method. Comparisons among the three platforms within each input condition were performed using Cochran's Q test. Paired between-platform risk differences, with 95% CIs, were additionally reported. When an omnibus comparison was significant, post-hoc pairwise comparisons were conducted using exact McNemar tests.

Multiplicity was addressed using the Holm procedure across the 25 principal comparisons, comprising 15 within-platform McNemar tests and 10 omnibus Cochran Q tests. Post-hoc pairwise p-values were Bonferroni-adjusted within each family of three comparisons. CIs were not adjusted for multiplicity and were interpreted as measures of estimation precision. Statistical significance was interpreted according to exact and Holm-adjusted p-values, rather than unadjusted CIs. Non-significant findings were not interpreted as evidence of equivalence or absence of an effect.

Category-specific analyses were descriptive because of the small subgroup sizes and were reported as n/N with whole-number percentages. Diagnostic-management concordance was evaluated using response-level cross-tabulations. Conditional outcome rates and risk differences with 95% CIs were calculated according to Top-1 correctness, with an exploratory analysis based on Top-4 accuracy. Inter-assessor agreement was calculated using the independent pre-consensus ratings. Final outcome counts and all inferential analyses were based on the final consensus ratings. The relationship between the independent ratings and the final consensus ratings is provided at the case level in Supplementary Data S1. Analyses were performed using IBM SPSS Statistics, version 29.0, and independently verified using the case-level dataset.

Results

A total of 50 initially misdiagnosed dermatologic emergency cases were evaluated across three consumer multimodal LLM platforms under text-only and text-plus-image conditions, corresponding to 300 response-level evaluations. Overall outcomes, with Wilson 95% CIs, are presented in Table 1. Paired input-mode comparisons (Table 2), omnibus between-platform comparisons (Table 3), and post-hoc pairwise comparisons (Table 4) are presented. The case-selection process is shown in Figure 1.

Overall Performance

Top-1 accuracy ranged from 16/50 (32.0%; 95% CI 20.8%-45.8%) to 25/50 (50.0%; 95% CI 36.6%-63.4%) under text-only input and from 18/50 (36.0%; 95% CI 24.1%-49.9%) to 29/50 (58.0%; 95% CI 44.2%-70.6%) under text-plus-image input. Top-4 accuracy ranged from 25/50 (50.0%; 95% CI 36.6%-63.4%) to 33/50 (66.0%; 95% CI 52.2%-77.6%) and from 28/50 (56.0%; 95% CI 42.3%-68.8%) to 37/50 (74.0%; 95% CI 60.4%-84.1%), respectively.

Clinically appropriate emergency management ranged from 19/50 (38.0%; 95% CI 25.9%-51.8%) to 27/50 (54.0%; 95% CI 40.4%-67.0%) under text-only input, and from 22/50 (44.0%; 95% CI 31.2%-57.7%) to 32/50 (64.0%; 95% CI 50.1%-75.9%) under text-plus-image input (Table 1).

Table 1. Overall outcomes by platform and input condition

Platform	Input	Top-1 accuracy	Top-4 accuracy	Appropriate management	Risk-triage failure	Harmful recommendation
ChatGPT	Text-only	23/50 (46.0%; 33.0-59.6)	31/50 (62.0%; 48.2-74.1)	25/50 (50.0%; 36.6-63.4)	13/50 (26.0%; 15.9-39.6)	5/50 (10.0%; 4.3-21.4)
ChatGPT	Text-plus-image	28/50 (56.0%; 42.3-68.8)	35/50 (70.0%; 56.2-80.9)	30/50 (60.0%; 46.2-72.4)	10/50 (20.0%; 11.2-33.0)	3/50 (6.0%; 2.1-16.2)
Gemini	Text-only	25/50 (50.0%; 36.6-63.4)	33/50 (66.0%; 52.2-77.6)	27/50 (54.0%; 40.4-67.0)	12/50 (24.0%; 14.3-37.4)	4/50 (8.0%; 3.2-18.8)
Gemini	Text-plus-image	29/50 (58.0%; 44.2-70.6)	37/50 (74.0%; 60.4-84.1)	32/50 (64.0%; 50.1-75.9)	8/50 (16.0%; 8.3-28.5)	2/50 (4.0%; 1.1-13.5)
Claude	Text-only	16/50 (32.0%; 20.8-45.8)	25/50 (50.0%; 36.6-63.4)	19/50 (38.0%; 25.9-51.8)	18/50 (36.0%; 24.1-49.9)	8/50 (16.0%; 8.3-28.5)
Claude	Text-plus-image	18/50 (36.0%; 24.1-49.9)	28/50 (56.0%; 42.3-68.8)	22/50 (44.0%; 31.2-57.7)	15/50 (30.0%; 19.1-43.8)	6/50 (12.0%; 5.6-23.8)

Values are presented as n/N (%; Wilson 95% confidence interval)

Table 2. Paired text-only versus text-plus-image comparisons

Platform	Outcome	Text	Image	T=1/I=0	T=0/I=1	RD (95% CI), pp	Exact p	Holm p
ChatGPT	Top-1	23	28	0	5	+10 (+1.8 to +17.8)	0.063	1.000
ChatGPT	Top-4	31	35	1	5	+8 (-1.3 to +17.2)	0.219	1.000
ChatGPT	Management	25	30	1	6	+10 (+0.1 to +19.4)	0.125	1.000
ChatGPT	Risk-triage failure	13	10	7	4	-6 (-19.0 to +7.0)	0.549	1.000
ChatGPT	Harmful recommendation	5	3	5	3	-4 (-16.3 to +8.1)	0.727	1.000
Gemini	Top-1	25	29	0	4	+8 (+0.6 to +15.1)	0.125	1.000
Gemini	Top-4	33	37	0	4	+8 (+0.3 to +15.9)	0.125	1.000
Gemini	Management	27	32	0	5	+10 (+1.7 to +18.0)	0.063	1.000
Gemini	Risk-triage failure	12	8	5	1	-8 (-18.3 to +1.7)	0.219	1.000
Gemini	Harmful recommendation	4	2	3	1	-4 (-14.3 to +5.2)	0.625	1.000
Claude	Top-1	16	18	2	4	+4 (-5.5 to +13.4)	0.688	1.000
Claude	Top-4	25	28	1	4	+6 (-2.4 to +14.2)	0.375	1.000
Claude	Management	19	22	2	5	+6 (-4.1 to +15.8)	0.453	1.000
Claude	Risk-triage failure	18	15	8	5	-6 (-19.6 to +7.9)	0.581	1.000
Claude	Harmful recommendation	8	6	6	4	-4 (-17.0 to +8.9)	0.754	1.000

Risk differences were calculated as the difference between the text-plus-image and text-only conditions. Unadjusted Newcombe hybrid-score 95% confidence intervals were used. Exact p values are derived from two-sided exact McNemar tests and may not be inferentially equivalent to Newcombe confidence intervals. Statistical significance was interpreted using exact and Holm-adjusted p values rather than unadjusted confidence intervals. Holm p values refer to the 25 principal comparisons

Paired Input-mode Comparisons

For all three platforms, Top-1 accuracy, Top-4 accuracy, and clinically appropriate management were numerically higher with text-plus-image input, while risk-triage failure rates and harmful recommendation rates were lower. The paired risk differences ranged from +4 to +10 percentage points for Top-1 accuracy, +6 to +8 percentage points for Top-4 accuracy, and +6 to +10 percentage points for management appropriateness. However, none of the 15 paired input-mode comparisons was statistically significant after Holm correction (all adjusted p=1.000; Table 2). These non-significant findings were not interpreted as demonstrating equivalence between input conditions.

Between-platform Comparisons

Significant between-platform differences remained after Holm correction for Top-1 accuracy under text-only input (Q=13.400, adjusted p=0.028) and text-plus-image input (Q=15.857, adjusted p=0.009). Under text-plus-image input, significant differences were also observed for Top-4 accuracy (Q=13.400, adjusted p=0.028) and for clinically appropriate management (Q=14.000, adjusted p=0.022). The remaining omnibus comparisons were not statistically significant after multiplicity correction (Table 3).

In Bonferroni-adjusted post-hoc comparisons, Gemini had higher Top-1 accuracy than Claude under text-only input, with a paired

Table 3. Omnibus between-platform comparisons within each input condition

Input	Outcome	ChatGPT	Gemini	Claude	Cochran Q	Raw p	Holm p
Text-only	Top-1	23/50	25/50	16/50	13.400	0.001	0.028
Text-only	Top-4	31/50	33/50	25/50	8.667	0.013	0.262
Text-only	Management	25/50	27/50	19/50	9.455	0.009	0.186
Text-only	Risk-triage failure	13/50	12/50	18/50	4.133	0.127	1.000
Text-only	Harmful recommendation	5/50	4/50	8/50	2.600	0.273	1.000
Text-plus-image	Top-1	28/50	29/50	18/50	15.857	<0.001	0.009
Text-plus-image	Top-4	35/50	37/50	28/50	13.400	0.001	0.028
Text-plus-image	Management	30/50	32/50	22/50	14.000	<0.001	0.022
Text-plus-image	Risk-triage failure	10/50	8/50	15/50	5.571	0.062	1.000
Text-plus-image	Harmful recommendation	3/50	2/50	6/50	2.889	0.236	1.000

Cochran Q tests accounted for the fact that the same 50 cases were evaluated by all three platforms. The Holm correction was applied across the 25 principal comparisons

Table 4. Post-hoc pairwise comparisons for Holm-significant omnibus outcome families

Input/outcome	Comparison	RD, pp	95% CI, pp	Raw p	Bonferroni p
Text-only Top-1	ChatGPT-Gemini	-4	-9.2 to +1.3	0.500	1.000
Text-only Top-1	ChatGPT-Claude	+14	+2.9 to +24.5	0.039	0.117
Text-only Top-1	Gemini-Claude	+18	+7.2 to +27.9	0.004	0.012
Text-plus-image Top-1	ChatGPT-Gemini	-2	-10.5 to +6.5	1.000	1.000
Text-plus-image Top-1	ChatGPT-Claude	+20	+7.4 to +31.3	0.006	0.019
Text-plus-image Top-1	Gemini-Claude	+22	+10.3 to +32.4	0.001	0.003
Text-plus-image Top-4	ChatGPT-Gemini	-4	-12.2 to +4.0	0.625	1.000
Text-plus-image Top-4	ChatGPT-Claude	+14	+4.3 to +23.3	0.016	0.047
Text-plus-image Top-4	Gemini-Claude	+18	+7.2 to +28.2	0.004	0.012
Text-plus-image management	ChatGPT-Gemini	-4	-11.6 to +3.7	0.625	1.000
Text-plus-image management	ChatGPT-Claude	+16	+4.4 to +26.7	0.021	0.064
Text-plus-image management	Gemini-Claude	+20	+8.7 to +30.2	0.002	0.006

Risk differences were calculated as the values from the first platform minus those from the second platform. Confidence intervals are unadjusted. Pairwise p values were Bonferroni-adjusted within each three-comparison family, CI: Confidence interval

risk difference of 18 percentage points (95% CI, 7.2-27.9; adjusted $p=0.012$). Under text-plus-image input, ChatGPT and Gemini had higher Top-1 accuracy than Claude by 20 percentage points (95% CI 7.4–31.3; adjusted $p=0.019$) and 22 percentage points (95% CI 10.3-32.4; adjusted $p=0.003$), respectively. For text-plus-image Top-4 accuracy, ChatGPT and Gemini exceeded Claude by 14 percentage points (95% CI 4.3-23.3; adjusted $p=0.047$) and 18 percentage points (95% CI 7.2-28.2; adjusted $p=0.012$), respectively. Regarding management appropriateness, Gemini exceeded Claude by 20 percentage points (95% CI 8.7-30.2; adjusted $p=0.006$). No post hoc comparisons between ChatGPT and Gemini were statistically significant (Table 4).

Category-specific Management

Category-specific management findings were descriptive. When given text-plus-image input, ChatGPT and Gemini each provided

appropriate management in 9/12 (75%) severe cutaneous adverse reaction cases. The appropriateness of management in vascular or vasculitic emergencies ranged from 3/9 (33%) to 5/9 (56%), depending on the platform and input condition. Because subgroup denominators ranged from four to twelve, these findings are considered exploratory and are presented as n/N with whole-number percentages in Figure 2.

Patient-safety Outcomes

Risk-triage failure ranged from 12/50 (24.0%; 95% CI 14.3%-37.4%) to 18/50 (36.0%; 95% CI 24.1%-49.9%) under text-only input and from 8/50 (16.0%; 95% CI 8.3%-28.5%) to 15/50 (30.0%; 95% CI 19.1%-43.8%) under text-plus-image input. Harmful recommendation rates ranged from 4/50 (8.0%; 95% CI 3.2%-18.8%) to 8/50 (16.0%; 95% CI 8.3%-28.5%) and from 2/50 (4.0%; 95% CI 1.1%-13.5%) to 6/50 (12.0%; 95% CI 5.6%-23.8%),

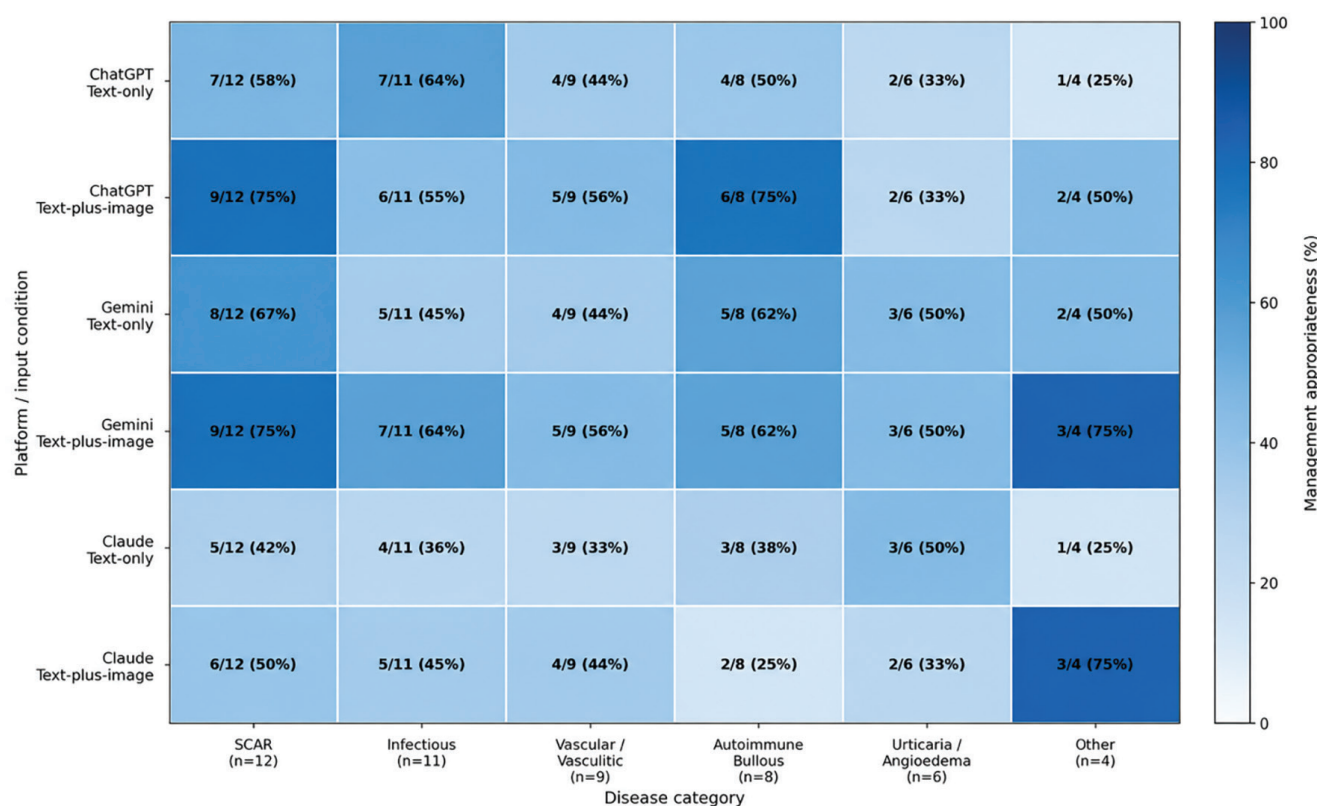


Figure 2. Clinically appropriate emergency management by disease category and input condition. Heat map showing n/N (whole-number percentage) for each platform, input condition, and disease category. A colour-blind-accessible, sequential blue palette is used, with darker shades indicating greater appropriateness of management

respectively. Neither outcome differed significantly between input conditions nor among platforms after the Holm correction (Tables 1-3).

Diagnostic-management Concordance

Diagnostic-management concordance was evaluated using a descriptive, response-level cross-tabulation across all 300 generated responses. For Top-1 accuracy, 139 responses were Top-1 correct with appropriate management, 0 were Top-1 correct with inappropriate management, 16 were Top-1 incorrect with appropriate management, and 145 were Top-1 incorrect with inappropriate management. Risk-triage failures and harmful recommendations occurred exclusively in Top-1 incorrect responses. For cumulative Top-4 accuracy, 155 responses were Top-4 correct with appropriate management; 34 were Top-4 correct with inappropriate management; 0 were Top-4 incorrect with appropriate management; and 111 were Top-4 incorrect with inappropriate management. These response-level descriptive findings indicate that a correct Top-1 diagnosis closely corresponds with appropriate management, whereas inclusion of the reference diagnosis within the Top-4 does not ensure safe management. Detailed response-level cross-tabulations and effect estimates are provided in the supplementary material.

Discussion

This single-run technical benchmark evaluated diagnostic and safety-related response performance across three three consumer multimodal LLM platforms using a selected publication-derived challenge set. Three findings emerged. First, text-plus-image input produced numerically more favorable outcomes, but none of the paired input-mode comparisons remained significant after multiplicity correction. Second, several between-platform differences remained significant, primarily reflecting lower observed Top-1, Top-4, and management performance for Claude than for Gemini in this dataset. Third, Top-1 diagnostic accuracy closely corresponded with appropriate management, whereas inclusion of the correct diagnosis among the Top-4 did not ensure safe management.

The response-level analysis clarified the relationship between diagnosis and management. All 139 Top-1-correct responses were accompanied by appropriate management, whereas only 16 of 161 Top-1-incorrect responses were. Risk-triage failures and harmful recommendations occurred only when the Top-1 diagnosis was incorrect. Nevertheless, 34/189 of the Top-4-correct responses had inappropriate management. The findings,

therefore, do not support a general dissociation between Top-1 diagnostic correctness and safe management; rather, they indicate that lower-ranked recognition of the correct diagnosis does not necessarily translate into appropriate clinical action.

Clinical images were associated with numerical improvements of 4-10 percentage points in Top-1 accuracy, 6-8 percentage points in Top-4 accuracy, and 6-10 percentage points in management appropriateness. Rates of risk-triage failure and harmful recommendations also declined numerically. However, none of these paired changes remained significant after Holm correction. Image resolution, illumination, lesion framing, skin tone, and the ability of a single photograph to represent lesion distribution may influence multimodal performance (7-9). Because image acquisition was not standardized, the results should be interpreted as reflecting performance under heterogeneous, publication-derived image conditions, not as evidence that image input is generally superior.

After multiplicity correction, between-platform differences remained for Top-1 accuracy under both input conditions, and for Top-4 accuracy and management appropriateness under the text-plus-image input. Post-hoc findings primarily reflected lower observed performance for Claude than for Gemini; ChatGPT also exceeded Claude in text-plus-image Top-1 and Top-4 accuracy. No significant post-hoc difference was observed between ChatGPT and Gemini. These comparisons are sample- and date-specific. They should not be interpreted as a stable ranking of the platforms because the exact backend versions were unavailable, each condition was queried only once, and consumer interfaces change over time.

Safety-related response outcomes provided information beyond the diagnostic rank. Harmful recommendations were less frequent than risk-triage failures, but represented active advice that could delay appropriate care. This distinction is relevant in time-sensitive conditions such as necrotizing soft tissue infection and severe cutaneous adverse reactions (4,11). Similar work in other acute-care and patient-facing LLM evaluations has separated diagnostic correctness from urgency, red-flag recognition, management, and risk of harm (12-14). Nevertheless, the present rates reflect only the content of generated responses and should not be interpreted as observed patient harm or real-world safety estimates.

Category-specific findings were heterogeneous and explicitly descriptive. With only 4-12 cases per category, a single response changed the percentage by approximately 8-25 percentage points. The comparatively favorable management outcomes for severe cutaneous adverse reactions and less favorable outcomes in some vascular or vasculitic cases may reflect differences in the

available clinical history, morphology, and systemic context, but the data cannot establish category-level superiority or causal explanations. Comparable dermatology-focused evaluations have likewise reported variable diagnostic performance across LLM platforms and evaluation settings (15-17).

The study's inferential scope must be separated into four domains. First, it directly assesses performance on a deliberately selected publication-derived challenge set. Second, because only one generation was obtained per condition, it does not assess output reproducibility or within-platform variability. Third, without a contemporaneous clinician comparator, it cannot determine the clinical usefulness as a decision-support tool, or whether use of the platforms improves clinician performance. Fourth, simulated responses to published cases cannot establish safety in real-world implementation. The safety-oriented endpoints, therefore, represent only a structured assessment of potentially unsafe content within the generated responses.

Automation bias and anchoring remain relevant considerations for future implementation research (18,19), but they were not directly evaluated here. Similarly, evidence from other diagnostic domains and human-AI studies cannot be used to infer benefit in this dataset (20,21). Future studies should use prospectively collected and representative cases, identifiable model versions, repeated generations, neutral and prespecified prompts, case-specific management standards, and contemporaneous clinician comparator groups. These design elements are necessary before conclusions about reproducibility, clinical benefit, or implementation safety can be made (22).

A strength of the study was the use of confirmed cases that posed documented diagnostic difficulties in their source reports and the separate evaluation of diagnostic and safety-related response endpoints. The same testing structure was applied across the three platforms, and case eligibility and response scoring involved both dermatology and emergency medicine expertise. These features support internal comparison within the challenge set but do not remove the limitations associated with retrospective case selection, single-run testing, and expert adjudication.

Study Limitations

This study employed a small, publication-derived challenge set, deliberately enriched with unusual and initially misdiagnosed dermatologic emergencies. It was neither representative nor consecutive, and therefore could not be used to estimate routine emergency department diagnostic accuracy. Subgroup sizes were small, and category-specific findings should be considered descriptive and hypothesis-generating.

Each platform-case-condition combination was queried once using a single standardized prompt on one testing date. The study therefore does not measure within-platform output variability, reproducibility across repeated generations, or longitudinal stability. The exact backend versions could not be verified through the consumer interfaces. Although the prompt did not identify the cases as dermatologic emergencies, the emergency-department setting and the role of expert emergency physicians may have encouraged urgency-oriented responses and may have limited comparisons with unprompted clinical use. In addition, the exact original vignette texts, complete verbatim model responses, query-specific timestamps, and complete query-level access metadata were not systematically archived. Although the retained case-level scoring dataset permits verification of the quantitative analyses, the original model responses cannot be independently re-adjudicated, and prompt-level auditability is limited.

All cases and images were derived from published reports. Training-data contamination, memorization, and recognition of distinctive clinical details or images cannot be excluded, even though web search functionality was not activated and source-identifying wording and image annotations were excluded when possible. Image acquisition and image quality were not standardized, and a publication-date sensitivity analysis could not be conducted because back-end knowledge cut-offs were unavailable. In addition, potential overlap with a previously published pilot study that used initially misdiagnosed dermatologic cases could not be definitively assessed because the pilot study did not provide a complete case-level source list or sufficient case identifiers. No vignettes, images, model outputs, scoring data, or statistical analyses were knowingly reused from that study; however, unrecognized overlap at the level of published source cases cannot be excluded.

Management appropriateness, risk-triage failure, and harmful recommendations were based on expert judgment. Although two specialists independently evaluated responses, and any disagreements were adjudicated by consensus, a formal case-specific reference-management checklist was not prospectively archived. Subjective outcome adjudication, therefore, remains a limitation.

No contemporaneous human-expert comparator was included. Initial misdiagnosis in the source reports was used only as a case-selection characteristic and did not represent a clinician–platform comparison. Accordingly, the study should be interpreted strictly as a technical benchmark of consumer interfaces using the selected dataset and as a partial analysis of safety-related response content. It does not establish reproducibility, clinical usefulness, clinician-support benefits, or safety in real-world implementation.

Conclusion

In this single-run technical benchmark of a selected publication-derived challenge set, the consumer multimodal LLM platforms demonstrated moderate diagnostic performance and incomplete safety-related performance in their responses. Text-plus-image input produced numerically more favorable outcomes; however, no pairwise input-mode comparison remained statistically significant after correction for multiple comparisons. Several between-platform differences were observed within this dataset, primarily reflecting lower performance for Claude than for Gemini; these findings do not establish stable platform superiority.

A correct Top-1 diagnosis closely corresponded with appropriate management, whereas inclusion of the correct diagnosis within the Top-4 did not ensure safe management. The findings apply only to the tested cases, consumer interfaces, testing date, single generation, and standardized prompting condition. They do not establish output reproducibility, clinical usefulness as decision support, or safety in real-world practice. Future studies should use representative prospective cases, identifiable model versions, repeated generations, predefined management standards, and contemporaneous clinician comparator groups.

Ethics

Ethics Committee Approval: The literature search and dataset construction were conducted between May 1 and June 1, 2026, and model evaluations were performed on June 2, 2026. The study used only anonymized clinical information and images obtained from published case reports and case series.

Informed Consent: No patient intervention was performed, no new clinical data were collected, no identifiable personal data were used, and no biological material sampling was performed.

Footnotes

Authorship Contributions

Concept: B.M.K., D.A., Design: B.M.K., D.A., Data Collection or Processing: B.M.K., D.A., Analysis or Interpretation: B.M.K., D.A., Literature Search: B.M.K., D.A., Writing: B.M.K.

Conflict of Interest: Demet Acar (Editor-in-Chief) is a member of the Editorial Board of the Eurasian Journal of Emergency Medicine. However, she was not involved in the editorial evaluation or decision-making process for this manuscript. The manuscript was evaluated by editors from different institutions. The other author declared no conflicts of interest.

Financial Disclosure: The authors declared that this study received no financial support.

Supplementary Table S1, S2: <https://d2v96fxpocvxx.cloudfront.net/1dda20c2-8b22-466b-a98e-85ecfb39ceaf/content-images/08fcb64-8248-4f15-8620-5f1aa9b01ed2.pdf>

Supplementary Data S1: <https://d2v96fxpocvxx.cloudfront.net/1dda20c2-8b22-466b-a98e-85ecfb39ceaf/content-images/65c87102-2765-44df-9efd-8e8de5049d5c.pdf>

References

- Podder I, Vasudevan B. Dermatological emergencies: what the term encompasses and key features in their diagnosis. In: Verma R, Vasudevan B, editors. *Dermatological Emergencies*. Boca Raton: CRC Press; 2019. p. 1-12.
- Chacon AH. Dermatologic emergencies. *Cutis*. 2015;95:E28-31.
- Gowda A, Christensen L, Polly S, Barlev D. Necrotizing neutrophilic dermatosis: a diagnostic challenge with a need for multi-disciplinary recognition, a case report. *Ann Med Surg (Lond)*. 2020;57:299-302.
- Alahmad MS, El-Menyar A, Abdelrahman H, Abdelrahman MA, Aurif F, Shaikh N, et al. Time to diagnose and time to surgery in patients presenting with necrotizing fasciitis: a retrospective analysis. *Eur J Trauma Emerg Surg*. 2025;51:140.
- Moola H, Visser WI. Erythroderma in the emergency department: a narrative review. *Emerg Care Med*. 2026;3:19.
- Preiksaitis C, Ashenburg N, Bunney G, Chu A, Kabeer R, Riley F, et al. The role of large language models in transforming emergency medicine: scoping review. *JMIR Med Inform*. 2024;12:e53787.
- Zhou J, He X, Sun L, Xu J, Chen X, Chu Y, et al. Pre-trained multimodal large language model enhances dermatological diagnosis using SkinGPT-4. *Nat Commun*. 2024;15:5649.
- Cirkel L, Lechner F, Henk LA, Krusche M, Hirsch MC, Hertl M, et al. Large language models for dermatological image interpretation: a comparative study. *Diagnosis (Berl)*. 2026;13:75-81.
- Ju EY, Sinharoy A, Edmonds N, Flowers RH. Diagnostic accuracy of three large language models on clinical images varies by skin tone: findings from the Stanford Diverse Dermatology Images Dataset. *Int J Dermatol*. 2026;65:1783-5.
- Cai ZR, Chen ML, Kim J, Novoa RA, Barnes LA, Beam A, et al. Assessment of correctness, content omission, and risk of harm in large language model responses to dermatology continuing medical education questions. *J Invest Dermatol*. 2024;144:1877-9.
- Dodiuk-Gad RP, Chung WH, Valeyrie-Allanore L, Shear NH. Stevens–Johnson syndrome and toxic epidermal necrolysis: an update. *Am J Clin Dermatol*. 2015;16:475-93.
- Mittal S, Aggarwal Y. Evaluation of large language models in the diagnosis, urgency triage, and initial management of ophthalmic emergencies. *Cureus*. 2026;18:e101433.
- Draeos RL, Afreen S, Blasko B, Brazile TL, Chase N, Desai DP, et al. Large language models provide unsafe answers to patient-posed medical questions. *NPJ Digit Med*. 2026;9:241.
- Agrawal M, Chen IY, Gulamali F, Joshi S. The evaluation illusion of large language models in medicine. *NPJ Digit Med*. 2025;8:600.
- Şen O. Performance of multimodal large language models in misdiagnosed dermatologic cases: a pilot study on diagnostic accuracy and human error replication. *Cutan Ocul Toxicol*. 2026.
- Liu X, Duan C, Kim MK, Zhang L, Jee E, Maharjan B, et al. Claude 3 Opus and ChatGPT with GPT-4 in dermoscopic image analysis for melanoma diagnosis: comparative performance analysis. *JMIR Med Inform*. 2024;12:e59273.
- Tekchandani N, Mukherjee A, Poonthottam N, Boussios S. Comparative analysis of large language models in dermatological diagnosis: an evaluation of diagnostic accuracy. *Cureus*. 2025;17:e92089.
- Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *J Am Med Inform Assoc*. 2012;19:121-7.
- Ly DP, Shekelle PG, Song Z. Evidence for anchoring bias during physician decision-making. *JAMA Intern Med*. 2023;183:818-23.
- McDuff D, Schaekermann M, Tu T, Palepu A, Wang A, Garrison J, et al. Towards accurate differential diagnosis with large language models. *Nature*. 2025;642:451-7.
- Krakowski I, Kim J, Cai ZR, Daneshjou R, Lapins J, Eriksson H, et al. Human–AI interaction in skin cancer diagnosis: a systematic review and meta-analysis. *NPJ Digit Med*. 2024;7:78.
- Omiye JA, Gui H, Rezaei SJ, Zou J, Daneshjou R. Large language models in medicine: the potentials and pitfalls: a narrative review. *Ann Intern Med*. 2024;177:210-20.