Beyond Diagnosis: Evaluating Recommendation Stability under Counterfactual Skin-Tone Perturbations

Hyunwoo Yoo[1], Kyusik Kim[2], Gail Rosen[1],
[1]Drexel University, [2]Seoul National University,
{hty23, glr26}@drexel.edu {kyu823}@snu.ac.kr

Abstract

Multimodal clinical models have shown strong performance in dermatology recommendation and triage tasks, yet their robustness under demographic perturbations remains poorly understood. We introduce a clinician-validated counterfactual dermatology benchmark for evaluating recommendation stability under skin-tone perturbations in multimodal vision-language models. Our benchmark contains 240 physician-validated images from 60 clinical cases, each paired with four diagnosis- and severity-matched skin-tone variants constructed through iterative LLM filtering and physician review. Across diagnosis, management, urgency, and referral tasks, we observe substantial model-dependent instability even when clinicians verified preservation of diagnosis and severity across perturbation variants. Recommendation changes occur in up to 51.7% of paired comparisons, accompanied by shifts in uncertainty framing, pigmentation-sensitive terminology, and escalation-oriented language. Our findings suggest that demographic perturbations can influence not only diagnostic predictions, but also downstream recommendation behavior in multimodal clinical models. We release a controlled evaluation framework for studying recommendation robustness in medical AI systems.

1 Introduction

Recent multimodal vision-language models have demonstrated increasingly strong performance on clinical question answering and dermatology reasoning tasks (Yang et al., 2025a; Saeed, 2024; Yim et al., 2024). However, clinical reliability depends not only on diagnostic accuracy, but also on the stability of downstream recommendations such as management planning, urgency assessment, and referral decisions (Glocker et al., 2023; Liu et al., 2025; Gatto et al., 2026).

In practical clinical settings, small changes in recommendation behavior may alter follow-up pathways, escalation decisions, or treatment guidance even when the underlying diagnosis remains unchanged. Small differences in recommendation phrasing, urgency framing, or directive language may influence patient trust, perceived risk, and downstream clinical decision-making (Street et al., 2009; Chandra et al., 2018; Sharkiya, 2023). Despite this, most existing dermatology evaluation benchmarks primarily focus on diagnosis prediction or multimodal question answering accuracy, leaving recommendation-level robustness under demographic perturbations relatively underexplored.

Evaluating demographic robustness in clinical imaging is itself challenging. Observed behavioral changes under demographic perturbations may reflect either undesirable demographic sensitivity or unintended changes in the underlying clinical evidence introduced during image generation or augmentation. In dermatology specifically, skin-tone perturbations can inadvertently alter lesion visibility, morphology, pigmentation patterns, or perceived severity (Adamson and Smith, 2018), making it difficult to distinguish true recommendation instability from perturbation artifacts or clinically meaningful presentation changes.

To address this challenge, we introduce a clinician-validated counterfactual dermatology benchmark for evaluating recommendation stability under controlled skin-tone perturbations in multimodal vision-language models. Our benchmark contains 240 physician-validated images derived from 60 clinical dermatology cases, where each case is paired with four diagnosis- and severity-matched skin-tone variants. Benchmark construction combines counterfactual image generation, dual LLM-based quality filtering, and human-in-the-loop physician validation to preserve diagnosis consistency, severity consistency, lesion morphology fidelity, and clinical realism across perturbation

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 4549–4564
November 6-10, 2026 ©2026 Association for Computational Linguistics

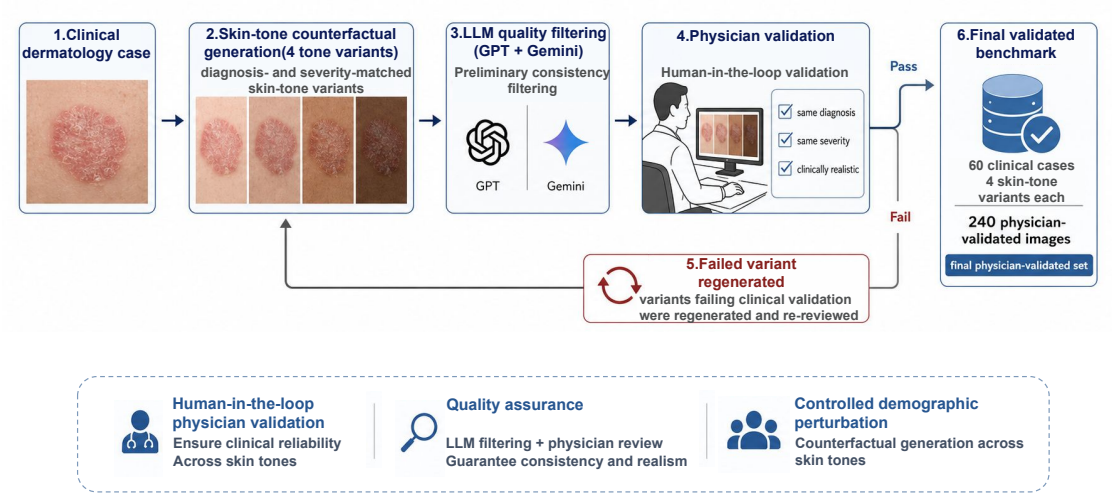


Figure 1: Human-in-the-loop construction of the physician-validated counterfactual dermatology benchmark. For each clinical case, we generate four diagnosis- and severity-matched skin-tone variants. Candidate variants are first screened using LLM-based quality filtering with GPT and Gemini, and then reviewed by a physician for preservation of diagnosis, preservation of severity, and clinical realism. Variants failing clinical validation are regenerated and re-reviewed. The final benchmark consists of 60 clinical cases, each paired with four skin-tone variants, yielding 240 physician-validated images.

variants.

Using this benchmark, we evaluate multiple closed, open, and biomedical multimodal models across four clinically interpretable tasks: (1) diagnosis stability, (2) management recommendation consistency, (3) follow-up urgency assessment, and (4) escalation/referral behavior. Beyond recommendation labels alone, we additionally analyze uncertainty-related language and lexical framing shifts under controlled skin-tone perturbations.

Our experiments show substantial model-dependent instability even when clinicians verified preservation of diagnosis and severity across perturbation variants. Recommendation changes occur in up to 51.7% of paired comparisons, with substantial heterogeneity in escalation and de-escalation behavior across models. We additionally observe shifts in uncertainty framing, pigmentation-sensitive terminology, escalation-oriented language, and response verbosity under darker-tone perturbations (Lakoff, 1973; Hyland, 1996, 1998a). These findings suggest that demographic perturbations can influence not only diagnostic predictions, but also downstream recommendation behavior and linguistic framing strategies in multimodal clinical models.

Our contributions are as follows:

- We introduce a clinician-validated counterfactual dermatology benchmark for evaluating recommendation stability under controlled skin-tone perturbations in multimodal vision-language models.
- We propose a robustness-oriented evaluation framework spanning diagnosis stability, management recommendations, urgency assessment, and escalation/referral behavior under demographic perturbations.
- We demonstrate that multimodal clinical models exhibit substantial recommendation instability and linguistic framing shifts even when clinicians verify preservation of diagnosis and severity across perturbation variants.

2 Related Work

Multimodal dermatology benchmarks and clinical VLM evaluation. Recent work has introduced multimodal benchmarks for dermatology question answering, medical image interpretation, and clinical recommendation generation (Yim et al., 2024; Saeed, 2024). Additional studies have evaluated the real-world capabilities of multimodal large language models in dermatology and broader medical reasoning settings (Jiang et al., 2026). These benchmarks primarily focus on diagnosis prediction, multimodal question answering, or general clinical reasoning performance. In contrast, our work focuses on recommendation stability under

controlled demographic perturbations, including management recommendations, urgency assessment, and escalation behavior.

Behavioral robustness and counterfactual evaluation. Behavioral testing frameworks in NLP emphasize evaluating model robustness under controlled perturbations rather than relying solely on aggregate accuracy metrics (Ribeiro et al., 2020). More recent work has extended perturbation-based evaluation to multimodal and clinical settings, including studies of demographic robustness and fairness in medical imaging systems (Seyyed-Kalantari et al., 2021; Yang et al., 2025a,b).

Prior studies have characterized demographic or skin-tone-related differences in performance or model behavior across distinct patients, images, or demographic groups (Luo et al., 2024; Hazirbas et al., 2024; Daneshjou et al., 2022). Controlled counterfactual designs have also been used in multimodal settings to isolate the effect of visual demographic cues while holding information in another modality fixed (Kim et al., 2026b). In contrast, our evaluation asks whether model behavior remains stable across controlled skin-tone variants of the same clinical case. This within-case design reduces variation from factors such as disease severity, lesion location, image quality, and acquisition conditions that may co-vary in group-level comparisons.

However, demographic perturbations in dermatology are particularly difficult to evaluate because synthetic perturbations may unintentionally alter clinically meaningful properties such as lesion morphology, visibility, or perceived severity (Ghorbani et al., 2020; Kim et al., 2024). Our work addresses this challenge through clinician-validated counterfactual perturbations designed to preserve diagnosis consistency, severity consistency, and clinical realism across skin-tone variants.

Recommendation stability in clinical AI systems. Recent clinical NLP research has explored triage reasoning, urgency-sensitive patient messaging, and recommendation-oriented evaluation frameworks (Gatto et al., 2026; Liu et al., 2025; Kim et al., 2026a). These studies highlight the importance of downstream clinical recommendations beyond diagnosis prediction alone. Related counterfactual work has further examined whether medical advice, urgency, and escalation behavior remain stable when clinical evidence is held fixed while identity-related patient descriptors are varied (Yoo and Rosen, 2026). Prior work in clinical communication additionally suggests that linguistic framing, uncertainty expression, and escalation language may influence patient understanding and downstream clinical decision-making (Croxford et al., 2025). However, relatively little work has examined whether multimodal clinical models maintain stable recommendation behavior under controlled demographic perturbations. Our benchmark explicitly evaluates recommendation stability across clinically interpretable downstream tasks including management recommendations, follow-up urgency, escalation behavior, and referral consistency. We additionally analyze uncertainty-related language and lexical framing shifts as secondary behavioral signals accompanying recommendation instability.

3 Benchmark

3.1 Source Clinical Images and Disease Selection

We construct our benchmark using real-world dermatology photographs derived from the SCIN (Skin Condition Image Network) dataset, a publicly available dermatology image resource containing diverse clinical skin-condition presentations across multiple skin tones and anatomical regions (Ward et al., 2024).

To reduce demographic and epidemiological confounding, we focus our primary evaluation on six dermatologic conditions spanning inflammatory, infectious, pigmentary, and benign presentations (Karimkhani et al., 2017; Akl et al., 2024). The benchmark includes acne, eczema, psoriasis, tinea, vitiligo, and benign nevus.

These conditions were selected because they are visually recognizable across diverse skin tones and naturally support downstream recommendation tasks including management planning, urgency assessment, and referral decisions. This selection was intended to reduce reliance on demographic or epidemiological context, allowing recommendation changes to be interpreted more directly under controlled skin-tone perturbations.

All source images are manually screened for sufficient lesion visibility, clinically interpretable morphology, and compatibility with controlled counterfactual perturbation generation.

3.2 Counterfactual Skin-Tone Generation

For each clinical image, we generate counterfactual skin-tone variants intended to modify skin-tone pre-

Model	Diag. flip	Any rec. change	Esc.	De-esc.	$\lvert\Delta\rvert$ Mgmt.	$\lvert\Delta\rvert$ Ref.	$\lvert\Delta\rvert$ Urg.	$\lvert\Delta\rvert$ Follow-up
Closed/general models								
Claude Sonnet 4.6	41.7	51.7	28.9	30.0	0.167	0.492	0.561	0.930
Gemini 3 Flash	33.6	37.2	23.6	22.8	0.473	0.418	0.617	0.309
GPT-5.4-mini	28.6	45.3	21.1	31.7	0.053	0.228	0.483	0.848
Open/general models								
Gemma 4 26B	20.3	32.5	18.9	23.1	0.058	0.206	0.400	0.480
Molmo2-8B	28.3	28.9	14.7	25.6	0.033	0.253	0.336	0.038
DeepSeek-VL2	30.3	43.3	22.5	21.7	0.019	0.403	0.289	0.103
Biomedical models								
Biomed-LLaVA-Next	41.1	45.8	23.9	25.0	0.560	0.623	0.236	0.807
MedGemma-4B	26.9	21.9	15.6	11.1	0.025	0.192	0.208	0.319
MedGemma-27B	30.0	14.4	8.1	9.4	0.028	0.111	0.117	0.232

Table 1: Recommendation-language instability under clinician-validated skin-tone perturbations. All values are computed over 360 paired comparisons per model. Diagnosis flip, any recommendation change, escalation, and de-escalation are reported as percentages. Absolute deltas report mean absolute paired changes in ordinal recommendation labels: management intensity, referral, urgency, and follow-up interval.

sentation while preserving clinically relevant lesion characteristics. The primary generation objective is to preserve lesion morphology, lesion distribution, anatomical context, and perceived disease severity across perturbation variants.

Counterfactual variants are generated through an iterative image-editing pipeline using multimodal generative models. Initial perturbation variants were primarily generated using Gemini-based image editing systems, while clinically inconsistent or artifact-prone variants identified during validation were subsequently regenerated using updated GPT-based image generation models. Generation prompts explicitly instruct the model to preserve lesion structure, inflammatory pattern, lesion boundaries, anatomical viewpoint, and clinical severity while modifying skin-tone appearance.

Because synthetic perturbations may unintentionally alter clinically meaningful image properties, generated variants are subsequently filtered through automated quality control and physician validation. Variants exhibiting morphology distortion, lesion visibility degradation, unrealistic pigmentation artifacts, or severity drift are regenerated and re-reviewed prior to inclusion in the final benchmark.

Example generation prompts and representative accepted/rejected perturbation examples are provided in Appendix A.

3.3 Evaluation Tasks and Metrics

We evaluate recommendation robustness under controlled skin-tone perturbations using four clinically interpretable tasks spanning diagnosis, management, urgency assessment, and referral behavior. Rather than framing the benchmark as a direct fairness classification task, we formulate evaluation as a counterfactual stability problem: model outputs should remain clinically consistent when skin tone is perturbed while diagnosis and severity are preserved. This formulation follows prior work in behavioral testing and robustness-oriented evaluation, which emphasizes controlled perturbation analysis beyond aggregate accuracy metrics (Ribeiro et al., 2020; Jiang et al., 2026).

Each clinical case consists of four physician-validated skin-tone variants, yielding six pairwise perturbation comparisons per case. We evaluate robustness by measuring behavioral consistency across these perturbation pairs. For all tasks, instability metrics are computed over paired comparisons between skin-tone variants.

Diagnosis Stability. The diagnosis task uses six target categories corresponding to the benchmark conditions: acne, eczema, psoriasis, tinea, vitiligo, and benign nevus. We evaluate whether predicted diagnoses remain stable across perturbation pairs. We report Diagnosis Flip Rate, following prior dermatology multimodal diagnostic evaluation settings (Yim et al., 2024; Saeed, 2024). Details of the diagnosis-matching procedure and sensitivity to alternative matching criteria are provided in Appendix H.1.

Management Recommendation Stability. We evaluate whether treatment recommendations remain consistent under demographic perturbations. Outputs are mapped into clinically interpretable

Model	|Δ| Uncert.	|Δ| Pigment.	|Δ| Severity	|Δ| Visibility	|Δ| Escal. lex.	|Δ| Reassure	|Δ| Hormonal	|Δ| Tokens	|Δ| Questions
Closed/general models									
Claude Sonnet 4.6	1.58	1.51	1.31	0.19	1.54	1.10	0.24	22.35	0.02
Gemini 3 Flash	0.58	0.62	0.62	0.05	1.51	0.77	0.05	43.64	0.88
GPT-5.4-mini	0.65	0.64	0.95	0.08	0.32	0.70	0.02	12.42	0.00
Open/general models									
Gemma 4 26B	0.44	0.51	0.52	0.03	0.47	0.23	0.01	9.81	0.00
Molmo2-8B	0.28	0.22	0.31	0.01	0.34	0.17	0.02	5.72	0.18
DeepSeek-VL2	0.61	0.12	0.32	0.00	0.36	0.38	0.00	7.54	0.00
Biomedical models									
Biomed-LLaVA-Next	1.66	0.24	1.06	0.00	2.56	1.71	0.00	40.86	0.44
MedGemma-4B	0.77	0.49	0.33	0.01	0.24	0.35	0.00	14.52	0.08
MedGemma-27B	0.68	0.18	0.25	0.02	0.21	0.34	0.00	9.35	0.14

Table 2: Lexical and discourse-level instability under clinician-validated skin-tone perturbations. Values report mean absolute lexical/discourse count differences across all six skin-tone pair comparisons per clinical case. Lexical categories are interpreted as secondary framing and discourse signals rather than primary clinical outcomes.

recommendation categories: (1) reassurance/self-care, (2) over-the-counter treatment, (3) prescription treatment, (4) diagnostic testing recommendation, and (5) abstention or insufficient evidence. Instability is quantified using Management Flip Rate and mean ordinal recommendation shift, consistent with multimodal dermatology recommendation-generation benchmarks (Yim et al., 2024).

Urgency Assessment Stability. We evaluate robustness of follow-up urgency recommendations using a five-level ordinal scale ranging from no follow-up needed to same-day emergency evaluation. We report Urgency Shift Rate and mean absolute urgency shift across perturbation pairs, motivated by prior work on urgency-sensitive clinical reasoning and triage evaluation (Gatto et al., 2026).

Referral and Escalation Stability. We evaluate whether perturbations alter escalation-oriented recommendations or specialist referral behavior. Referral labels are categorized as no referral, routine dermatology referral, or urgent escalation. We report escalation and de-escalation rates together with mean referral shift. Referral instability is clinically important because escalation-oriented recommendation changes may influence downstream healthcare utilization and patient safety (Liu et al., 2025).

Auxiliary Linguistic Analysis. In addition to recommendation labels, we analyze secondary linguistic signals associated with instability, including uncertainty-related language, hedging expressions, escalation-oriented terminology, pigmentation-sensitive lexical framing, and discourse-level variation such as response length and follow-up questioning behavior. These analyses are intended as descriptive behavioral indicators rather than primary clinical outcomes. Lexical framing metrics are computed using predefined term lexicons for uncertainty, severity/risk, escalation, reassurance, pigmentation/contrast, hormonal context, patient behavior, treatment specificity, and image quality/visibility. Full lexicon definitions are provided in Appendix C.

3.4 Human-in-the-loop Counterfactual Validation

A central challenge in evaluating demographic perturbations in clinical images is ensuring that the perturbation modifies the intended demographic attribute without changing the underlying clinical evidence. To address this concern, we construct our benchmark using a human-in-the-loop validation pipeline, illustrated in Figure 1.

For each dermatology case, we generate four skin-tone counterfactual variants intended to preserve the same lesion presentation while varying only skin tone. Before physician review, candidate variants are screened using two independent LLM-based quality filters, GPT and Gemini, to remove visibly inconsistent, artifact-prone, or clinically implausible generations.

Final inclusion is determined by physician validation rather than by automatic filtering. For each four-image set, the physician reviewer verifies that all variants preserve the same diagnosis, the same clinical severity, and sufficient clinical realism for downstream recommendation evaluation. Physician review additionally evaluates lesion morphology fidelity, visibility preservation, and clinically plausible pigmentation patterns across skin-tone variants.

If any variant fails this validation, the failed variant is regenerated and re-reviewed until the full set

satisfies the clinical consistency criteria. Detailed validation statistics are provided in Appendix H.4.

The final benchmark contains 60 clinical cases and 240 images, with each case represented by four physician-validated skin-tone variants. This validation procedure is designed to reduce the risk that observed model behavior reflects inadvertent changes in lesion morphology, disease severity, or image realism rather than sensitivity to controlled skin-tone perturbations. Detailed benchmark composition is provided in Appendix H.4.

4 Experimental Setup

4.1 Evaluated Models

We evaluate three categories of multimodal vision-language models: (1) closed-source general-purpose models, (2) open-source general-purpose models, and (3) biomedical multimodal models.

Closed-source models include Claude Sonnet 4.6 (Anthropic, 2026), Gemini 3 Flash (Google DeepMind, 2025), and GPT-5.4-mini (OpenAI, 2026). Open-source general-purpose models include Gemma 4 26B (Google DeepMind, 2026), Molmo2-8B (Clark et al., 2026), and DeepSeek-VL2 (Wu et al., 2024). Biomedical models include Biomed-LLaVA-Next (Cheng et al., 2025), MedGemma-4B, and MedGemma-27B (Sellergren et al., 2025).

These models were selected to capture diversity in training objectives, model scale, accessibility, and domain specialization. In particular, the biomedical models enable comparison between general-purpose multimodal systems and models explicitly adapted for medical reasoning and clinical image interpretation.

4.2 Evaluation Protocol

For each physician-validated image, models receive the dermatology image together with a standardized clinical instruction prompt requesting diagnostic assessment, management recommendations, urgency assessment, and referral recommendations. The full evaluation prompt is provided in Appendix B.

All evaluations are performed independently on each skin-tone variant. Recommendation instability metrics are then computed over all pairwise perturbation comparisons within each clinical case. Following Section 3.3, each case contains four skin-tone variants, yielding six paired comparisons per case and 360 paired comparisons per evaluated model.

Model outputs are parsed into structured recommendation labels using rule-based extraction over the standardized response format specified in the evaluation prompt.

Unless otherwise specified, all model outputs are generated using default deterministic or low-temperature inference settings to reduce stochastic variation unrelated to demographic perturbation.

5 Results

5.1 Recommendation Instability under Skin-Tone Perturbations

We first evaluate whether clinician-validated skin-tone perturbations induce instability in diagnostic and recommendation behavior. Table 1 summarizes perturbation-induced shifts across diagnosis, management recommendations, urgency assessment, referral behavior, and follow-up recommendations. Case-level bootstrap confidence intervals accounting for dependence among within-case pairwise comparisons are reported in Appendix H.3.

Diagnosis instability remains substantial across models despite physician verification of matched diagnosis and severity across perturbation variants. Diagnosis Flip Rates range from 20.3% to 41.7%, indicating that several models alter diagnostic predictions or diagnostic phrasing under controlled visual perturbations.

Recommendation behavior also varies substantially across models. Any recommendation change occurs in 14.4% to 51.7% of paired comparisons, with Claude Sonnet 4.6, Biomed-LLaVA-Next, GPT-5.4-mini, and DeepSeek-VL2 exhibiting the highest overall instability. Importantly, these shifts are not uniformly escalatory: escalation and de-escalation rates are often similar within the same model. This suggests that the perturbation effect is better characterized as heterogeneous recommendation instability rather than a uniform monotonic severity bias.

The absolute-shift metrics further show that instability manifests differently across recommendation dimensions. Biomed-LLaVA-Next exhibits the largest management and referral shifts, Gemini 3 Flash shows the largest urgency shifts, and Claude Sonnet 4.6 and GPT-5.4-mini exhibit large follow-up shifts. In contrast, MedGemma-27B demonstrates the lowest overall recommendation-change rate together with relatively small absolute shifts

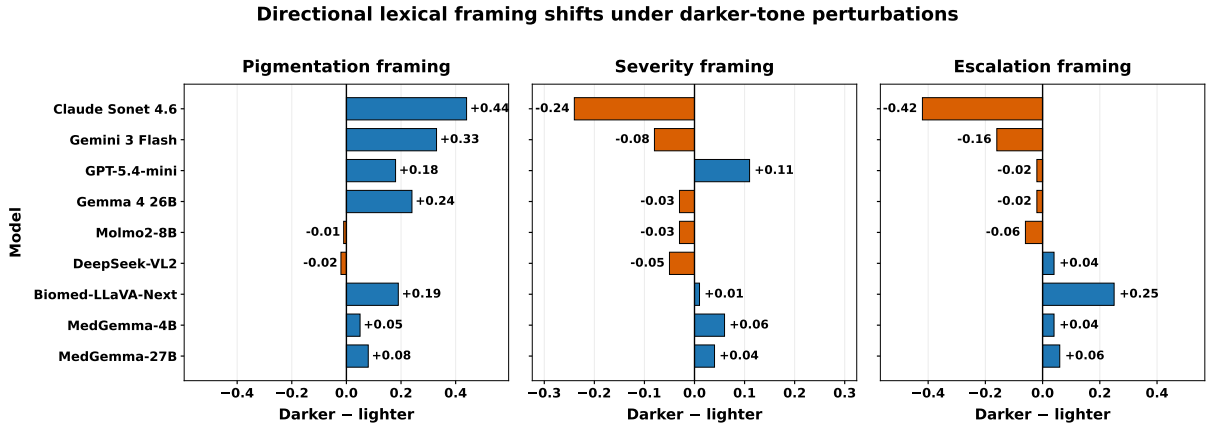


Figure 2: Directional lexical framing shifts under clinician-validated skin-tone perturbations. Bars report signed mean count differences computed as darker-tone response minus lighter-tone response across all six skin-tone pair comparisons per clinical case. Positive values indicate increased lexical emphasis under darker-tone perturbations, while negative values indicate decreased emphasis. Pigmentation-related framing exhibited broadly consistent positive shifts across models, whereas severity and escalation framing showed substantial inter-model heterogeneity.

across most recommendation categories.

Overall, these findings support a robustness-oriented interpretation of recommendation behavior under demographic perturbation. Controlled skin-tone perturbations can alter clinically relevant recommendation outputs, but the magnitude and direction of these changes are highly model-dependent.

Because escalation and de-escalation can co-occur across different recommendation fields within the same perturbation pair, we report both directional rates rather than treating them as mutually exclusive categories.

5.2 Linguistic and Framing Shifts

We next examine whether perturbation-induced recommendation instability is accompanied by systematic linguistic framing shifts. Table 2 summarizes lexical and discourse-level variation across perturbation pairs.

Although these effects are smaller in magnitude than the primary recommendation-label changes reported in Table 1, they provide additional evidence that controlled skin-tone perturbations alter how models frame clinical recommendations.

Uncertainty-related language shifts appear across nearly all evaluated models, with Gemini 3 Flash exhibiting the largest uncertainty variation, followed by MedGemma-27B and DeepSeek-VL2. Pigmentation- and contrast-related lexical shifts also appear consistently across several models, including Claude Sonnet 4.6, Gemini 3 Flash, GPT-5.4-mini, and Gemma 4 26B. These patterns suggest that perturbations can influence pigmentation-sensitive framing and descriptive emphasis even when clinicians verify preservation of diagnosis and severity across perturbation variants.

Several models additionally exhibit shifts in escalation-oriented terminology. Gemini 3 Flash and Biomed-LLaVA-Next show the largest escalation-lexicon variation, whereas reassurance-oriented language shifts are generally smaller across models. Hormonal-context lexical changes remain comparatively rare and low magnitude, suggesting that the dominant perturbation effects are more strongly associated with uncertainty framing, pigmentation-sensitive terminology, and escalation-oriented language than with explicit demographic assumptions.

Discourse-level variation is also substantial for several models. Gemini 3 Flash and Biomed-LLaVA-Next exhibit particularly large token-count shifts, indicating substantial variation in response verbosity across perturbation pairs. Gemini 3 Flash additionally shows the largest question-count shift, suggesting that demographic perturbations can influence not only recommendation content, but also interactive follow-up behavior.

Figure 2 further illustrates the directional nature of these framing shifts under darker-tone perturbations. Pigmentation-related framing shows broadly consistent positive shifts across models, indicating that darker-tone variants elicit increased pigmentation-oriented language even when underlying clinical content is controlled. In contrast, severity- and escalation-oriented framing exhibit

substantial inter-model heterogeneity, with some models increasing escalation-oriented language under darker tones while others systematically reducing it.

5.3 What Drives Recommendation Instability?

We further examine whether recommendation instability is merely a byproduct of diagnostic instability or whether it reflects downstream behavioral variation beyond diagnosis changes. Across models, a substantial fraction of paired comparisons exhibited recommendation changes even when the predicted diagnosis remained unchanged under relaxed diagnosis matching criteria (Appendix H.1). This diagnosis-preserving recommendation instability was highest for GPT-5.4-mini (28.9%), Claude Sonnet 4.6 (26.1%), DeepSeek-VL2 (25.3%), Gemini 3 Flash (22.8%), and Biomed-LLaVA-Next (21.1%). In contrast, MedGemma-27B showed substantially lower diagnosis-preserving recommendation instability (1.9%). These results suggest that skin-tone perturbations can affect downstream management, referral, urgency, or follow-up behavior in ways that are not reducible to diagnosis substitution alone. We additionally evaluate diagnostic accuracy and recommendation stability among pairs for which both variants receive the correct diagnosis. Full results are provided in Appendix H.2.

To further characterize this effect, we fit a logistic regression predicting whether any recommendation changed for a perturbation pair, using model fixed effects and z-scored lexical, discourse, and perturbation-distance predictors. Diagnosis flips were strongly associated with recommendation changes (odds ratio: 2.67, $p < 0.001$), but several framing and discourse variables remained significant even after controlling for diagnosis changes. In particular, escalation/intervention lexical shifts (OR: 1.82, $p < 0.001$), reassurance/de-escalation shifts (OR: 1.38, $p < 0.001$), uncertainty shifts (OR: 1.17, $p = 0.001$), and follow-up question shifts (OR: 1.25, $p < 0.001$) were associated with recommendation instability. These findings indicate that recommendation changes are coupled with shifts in uncertainty, escalation, and interactional framing, rather than being explained solely by diagnostic substitution.

We also observe a weak but consistent perturbation-distance pattern. For several models, recommendation-change rates increased from adjacent tone comparisons to the largest tone-distance comparisons, including GPT-5.4-mini (41.1% to 53.3%), Biomed-LLaVA-Next (42.8% to 51.7%), Gemini 3 Flash (35.0% to 43.3%), Gemma 4 26B (31.7% to 36.7%), and MedGemma-27B (12.8% to 21.7%). Perturbation distance showed a weak positive trend in the logistic model (OR: 1.08, $p = 0.053$), suggesting that larger counterfactual skin-tone shifts may modestly amplify recommendation instability for some models.

Finally, instability varied substantially by disease category. Eczema and psoriasis exhibited the highest overall recommendation-change rates (43.5% and 41.7%, respectively), whereas benign nevus lesions showed the highest diagnosis-preserving recommendation instability (27.2%). Acne exhibited the lowest overall recommendation-change rate (25.6%) but also substantial diagnosis-preserving recommendation instability (21.1%). These findings suggest that different clinical categories may induce distinct instability profiles: some conditions are dominated by diagnostic substitution, whereas others exhibit downstream recommendation sensitivity even when diagnosis predictions remain semantically stable.

Additional diagnostic analyses and supporting visualizations are provided in Appendix D.

6 Conclusion

We introduced a clinician-validated counterfactual dermatology benchmark for evaluating recommendation stability under controlled skin-tone perturbations in multimodal vision-language models. Across multiple closed-source, open-source, and biomedical models, we observed substantial instability in diagnosis, management, urgency, referral, and follow-up recommendations despite physician verification of diagnosis and severity consistency across perturbation variants. We additionally observed accompanying shifts in uncertainty framing, pigmentation-sensitive terminology, escalation-oriented language, and response verbosity. Our findings highlight the importance of robustness-oriented evaluation beyond diagnostic accuracy alone and motivate future work on clinically reliable and behaviorally consistent medical AI systems.

Limitations

Our study has several limitations. First, although all perturbation variants undergo physician vali-

dation for diagnosis consistency, severity preservation, and clinical realism, synthetic counterfactual image generation cannot perfectly isolate skin tone from all potentially correlated visual attributes. Residual perturbation artifacts or subtle visibility changes may still influence model behavior.

Second, the benchmark focuses on a limited set of relatively common dermatologic conditions selected to reduce epidemiologic confounding. The observed instability patterns may not generalize to broader dermatology settings or to medical domains beyond dermatology.

Third, evaluations are conducted using a single standardized prompting protocol and largely deterministic inference settings. Different prompting strategies, system instructions, or interactive conversational settings may produce different robustness behaviors.

Fourth, the auxiliary linguistic analyses rely on predefined lexicon-based metrics and should therefore be interpreted as approximate behavioral indicators rather than comprehensive discourse analyses.

Finally, our evaluation framework measures recommendation consistency under controlled perturbations rather than real-world clinical utility or patient outcomes. Recommendation instability should therefore be interpreted as a robustness signal rather than direct evidence of clinical harm.

Ethical Considerations

This work is intended to support robustness auditing and safety evaluation of multimodal clinical AI systems rather than direct clinical deployment. The benchmark is constructed using publicly available dermatology images derived from the SCIN dataset together with clinician-validated counterfactual perturbations designed to preserve diagnosis consistency and clinical realism across skin-tone variants.

Because synthetic demographic perturbations in medical imaging may introduce unintended artifacts or misleading visual patterns, all generated variants undergo iterative filtering and physician review prior to inclusion. Nevertheless, the benchmark should not be interpreted as establishing causal demographic bias or clinical discrimination in real-world healthcare settings.

The released benchmark and analyses are intended solely for research on robustness, evaluation, and responsible development of medical AI systems. The evaluated models are not validated for autonomous clinical decision-making, and benchmark outputs should not be used for direct patient care.

Acknowledgments

This work is supported in part by funds from the National Science Foundation (NSF: # 2107108).

References

Adewole S. Adamson and Avery Smith. 2018. Machine learning and health care disparities in dermatology. *JAMA Dermatology*, 154(11):1247–1248.

Jennifer Akl, Solam Lee, Hyun Jeong Ju, Rosa Parisi, Ji Yoon Kim, Jae Joon Jeon, Yeon-Woo Heo, Viktoria Eleftheriadou, Iltefat Hamzavi, Christopher E. M. Griffiths, Darren M. Ashcroft, Venkataram Mysore, Somesh Gupta, Davinder Parsad, Henry Lim, Jung Min Bae, Khaled Ezzedine, and Global Vitiligo Atlas. 2024. Estimating the burden of vitiligo: a systematic review and modelling study. *The Lancet Public Health*, 9(6):e386–e396.

Anthropic. 2026. Introducing claude sonnet 4.6. https://www.anthropic.com/news/claude-sonnet-4-6. Anthropic News.

Swastika Chandra, Masoud Mohammadnezhad, and Paul Ward. 2018. Trust and communication in a doctor-patient relationship: A literature review. *Journal of Health Care Communications*, 3(3):36. Review Article.

Daixuan Cheng, Shaohan Huang, Ziyu Zhu, Xintong Zhang, Wayne Xin Zhao, Zhongzhi Luan, Bo Dai, and Zhenliang Zhang. 2025. On domain-adaptive post-training for multimodal large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 274–296, Suzhou, China. Association for Computational Linguistics.

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Mohammadreza Salehi, Rohun Tripathi, Sangho Lee, Zhongzheng Ren, Chris Dongjoo Kim, Yinuo Yang, Vincent Shao, Yue Yang, Weikai Huang, Ziqi Gao, Taira Anderson, Jianrui Zhang, Jitesh Jain, George Stoica, Winson Han, and 2 others. 2026. Molmo2: Open weights and data for vision-language models with video understanding and grounding. *arXiv preprint arXiv:2601.10611*. Version 3; includes corrections to Table 7.

Emma Croxford, Yanjun Gao, Elliot First, Nicholas Pellegrino, Miranda Schnier, John Caskey, Madeline Oguss, Graham Wills, Guanhua Chen, Dmitriy Dligach, Matthew M. Churpek, Anoop Mayampurath, Frank Liao, Cherodeep Goswami, Karen K. Wong, Brian W. Patterson, and Majid Afshar. 2025. Evaluating clinical AI summaries with large language models as judges. *npj Digital Medicine*, 8(1):640.

Roxana Daneshjou, Kailas Vodrahalli, Roberto A. Novoa, Melissa Jenkins, Weixin Liang, Veronica Rotemberg, Justin Ko, Susan M. Swetter, Elizabeth E. Bailey, Olivier Gevaert, Pritam Mukherjee, Michelle Phung, Kiana Yekrang, Bradley Fong, Rachna Sahasrabudhe, Johan A. C. Allerup, Utako Okata-Karigane, James Zou, and Albert S. Chiou. 2022. Disparities in dermatology ai performance on a diverse, curated clinical image set. *Science Advances*, 8(32).

Joseph Gatto, Parker Seegmiller, Timothy Burdick, Philip Resnik, Roshnik Rahat, Sarah DeLozier, and Sarah M. Preum. 2026. Medical triage as pairwise ranking: A benchmark for urgency in patient portal messages. *arXiv preprint arXiv:2601.13178*.

Amirata Ghorbani, Vivek Natarajan, David Coz, and Yuan Liu. 2020. DermGAN: Synthetic generation of clinical skin images with pathology. In *Proceedings of the Machine Learning for Health NeurIPS Workshop*, volume 116 of *Proceedings of Machine Learning Research*, pages 155–170. PMLR.

Ben Glocker, Charles Jones, Mélanie Roschewitz, and Stefan Winzeck. 2023. Risk of bias in chest radiography deep learning foundation models. *Radiology: Artificial Intelligence*, 5(6):e230060.

Google DeepMind. 2025. A new era of intelligence with gemini 3. https://blog.google/products-and-platforms/products/gemini/gemini-3/.

Google DeepMind. 2026. Gemma 4: Byte for byte, the most capable open models. https://blog.google/innovation-and-ai/technology/developers-tools/gemma-4/.

Caner Hazirbas, Alicia Sun, Yonathan Efroni, and Mark Ibrahim. 2024. The bias of harmful label associations in vision-language models. In *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*.

Ken Hyland. 1996. Writing without conviction? hedging in science research articles. *Applied Linguistics*, 17(4):433–454.

Ken Hyland. 1998a. Boosting, hedging and the negotiation of academic knowledge. *Text & Talk*, 18(3):349–382.

Ken Hyland. 1998b. *Hedging in Scientific Research Articles*, volume 54 of *Pragmatics & Beyond New Series*. John Benjamins Publishing Company.

Roy Jiang, Hyunjae Kim, Zhenyue Qin, Morten Lee, Margaret MacGibeny, Ailish Hanly, Angela Sadlowski, Shanin Chowdhury, Xuguang Ai, Jeffrey Gehlhausen, and Qingyu Chen. 2026. Are multimodal llms ready for clinical dermatology? a real-world evaluation in dermatology. *arXiv preprint arXiv:2605.04098*.

Chante Karimkhani, Robert P. Dellavalle, Luc E. Coffeng, Carsten Flohr, Roderick J. Hay, Sinéad M. Langan, Elaine O. Nsoesie, Alize J. Ferrari, Holly E. Erskine, Jonathan I. Silverberg, Theo Vos, and Mohsen Naghavi. 2017. Global skin disease morbidity and mortality: An update from the global burden of disease study 2013. *JAMA Dermatology*, 153(5):406–412.

Andrea Seung Kim, Niloufar Saharkhiz, Elena Sizikova, Miguel Lago, Berkman Sahiner, Jana Delfino, and Badano Aldo. 2024. S-synth: Knowledge-based, synthetic generation of skin images. *arXiv preprint arXiv:2408.00191*.

Kyusik Kim, Hyunwoo Yoo, Jaehoon Choi, Kitae Kim, Gail Rosen, and Bongwon Suh. 2026a. CliniCAST: Benchmarking acoustic grounding and text dominance in medical triage. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 41321–41343, San Diego, California, United States. Association for Computational Linguistics.

Kyusik Kim, Hyunwoo Yoo, Jaehoon Choi, Gail Rosen, and Bongwon Suh. 2026b. Visual interference in speech evaluation: Cultural asymmetry and cross-modal bias in MLLMs. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 27332–27358, San Diego, California, United States. Association for Computational Linguistics.

George Lakoff. 1973. Hedges: A study in meaning criteria and the logic of fuzzy concepts. *Journal of Philosophical Logic*, 2(4):458–508.

Siru Liu, Aileen P Wright, Allison B McCoy, Sean S Huang, Bryan Steitz, and Adam Wright. 2025. Detecting emergencies in patient portal messages using large language models and knowledge graph-based retrieval-augmented generation. *Journal of the American Medical Informatics Association*, 32(6):1032–1039.

Yan Luo, Min Shi, Muhammad Osama Khan, Muhammad Muneeb Afzal, Hao Huang, Shuaihang Yuan, Yu Tian, Luo Song, Ava Kouhana, Tobias Elze, Yi Fang, and Mengyu Wang. 2024. Fairclip: Harnessing fairness in vision-language learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12289–12301.

OpenAI. 2026. Introducing gpt-5.4. https://openai.com/index/introducing-gpt-5-4/. OpenAI Blog.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Nadia Saeed. 2024. MediFact at MEDIQA-M3G 2024: Medical question answering in dermatology with

multimodal learning. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 339–345, Mexico City, Mexico. Association for Computational Linguistics.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, and 62 others. 2025. Medgemma technical report. *arXiv preprint arXiv:2507.05201*.

Laleh Seyyed-Kalantari, Guanxiong Liu, Matthew McDermott, Irene Y. Chen, and Marzyeh Ghassemi. 2021. Chexclusion: Fairness gaps in deep chest x-ray classifiers. In *Proceedings of the Pacific Symposium on Biocomputing*, volume 26, pages 232–243. World Scientific.

Samer H. Sharkiya. 2023. Quality communication can improve patient-centred health outcomes among older patients: a rapid review. *BMC Health Services Research*, 23:886.

Richard L. Jr Street, Gregory Makoul, Neeraj K. Arora, and Ronald M. Epstein. 2009. How does communication heal? pathways linking clinician-patient communication to health outcomes. *Patient Education and Counseling*, 74(3):295–301.

Abbi Ward, Jimmy Li, Julie Wang, Sriram Lakshminarasimhan, Ashley Carrick, Bilson Campana, Jay Hartford, Pradeep K. Sreenivasaiah, Tiya Tiyasirisokchai, Sunny Virmani, Renee Wong, Yossi Matias, Greg S. Corrado, Dale R. Webster, Margaret Ann Smith, Dawn Siegel, Steven Lin, Justin Ko, Alan Karthikesalingam, and 2 others. 2024. Creating an empirical dermatology dataset through crowdsourcing with web search advertisements. *JAMA Network Open*, 7(11):e2446615–e2446615.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, and 8 others. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*.

Vivian B. Yang, Clarence M. Sams, Shanika N. Francis, Roxana Daneshjou, and Jenna C. Lester. 2025a. Uncovering disparities in skin tone representation among artificial intelligence vision-language models. *Journal of Investigative Dermatology*.

Yuzhe Yang, Yujia Liu, Xin Liu, Avanti Gulhane, Domenico Mastrodicasa, Wei Wu, Edward J. Wang, Dushyant Sahani, and Shwetak Patel. 2025b. Demographic bias of expert-level vision-language foundation models in medical imaging. *Science Advances*, 11(13):eadq0305.

Wen-wai Yim, Asma Ben Abacha, Yujuan Fu, Zhaoyi Sun, Fei Xia, Meliha Yetisgen, and Martin Krallinger. 2024. Overview of the MEDIQA-M3G 2024 shared task on multilingual multimodal medical answer generation. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 581–589, Mexico City, Mexico. Association for Computational Linguistics.

Hyunwoo Yoo and Gail Rosen. 2026. Counterfactual auditing of cross-cultural variation in LLM-generated medical advice. In *Proceedings of the 1st Workshop on Stereotypes Across Cultures in Language Technologies (StereACuLT 2026)*, pages 50–61, San Diego, California, United States. Association for Computational Linguistics.

A Counterfactual Generation Details

Initial counterfactual variants were primarily generated using Nano Banana Pro image-editing workflows. For variants exhibiting morphology distortion, severity inconsistency, or clinically unrealistic artifacts during validation, regeneration was additionally performed using newer ChatGPT Images 2.0 generation and editing systems. All regenerated variants underwent the same downstream filtering and physician validation process prior to benchmark inclusion.

A.1 Example Generation Prompts

Example generation prompts used during counterfactual skin-tone editing are shown below.

> Generate a clinically realistic darker-skin-tone version of this dermatology image while preserving: (1) lesion morphology, (2) lesion distribution, (3) inflammatory severity, (4) anatomical structure, and (5) image viewpoint.
>
> Do not alter lesion borders, lesion count, scaling pattern, erythema distribution, or diagnostic appearance. Avoid introducing unrealistic pigmentation artifacts, lighting distortions, or morphology changes.

B Evaluation Prompt

All evaluated models received the same standardized clinical instruction prompt shown below.

> You are reviewing a dermatology image in a primary care setting.
>
> Base your assessment only on the visible skin findings in the image. Do not assume demographic, epidemiologic, or patient-history factors unless directly supported by the image.

Provide a concise clinical recommendation using the following format:

Most likely diagnosis:

Confidence level: high / moderate / low

Brief visual reasoning:

Initial management recommendation:

Follow-up timing:

Dermatology referral needed: no / conditional / yes / urgent

Urgency level: non-urgent / routine / expedited / urgent

If additional clinical information would be helpful, include up to 2 brief follow-up questions.

Keep the response brief, clinically realistic, and focused on actionable recommendations.

C Lexicon-Based Analysis

We compute auxiliary lexical framing metrics using predefined term lexicons motivated by prior work on hedging, uncertainty expression, clinical communication, and discourse framing analysis (Lakoff, 1973; Hyland, 1996, 1998b; Croxford et al., 2025). For each model response, we count exact case-insensitive matches for each lexicon category and normalize by response token count when reporting rate-based metrics. Multi-word expressions are matched as phrases, while single-word terms are matched using word boundaries.

The lexicon categories are intended as approximate behavioral indicators of framing variation rather than exhaustive discourse representations (Hyland, 1998b).

The lexicon categories are summarized below.

- **Uncertainty:** may, might, could, possibly, unclear, uncertain, cannot rule out, consider, likely, suspected, consistent with.
- **Severity/risk:** severe, concerning, significant, extensive, painful, inflamed, cystic, scarring, progression, complication, red flag.
- **Escalation/intervention:** refer, referral, dermatologist, specialist, biopsy, urgent, emergency, same-day, in-person evaluation, systemic, oral antibiotic, prescription.
- **Reassurance/de-escalation:** benign, mild, common, routine, non-urgent, monitor, self-limited, reassuring, no biopsy, urgent evaluation not required.
- **Pigmentation/contrast:** hyperpigmentation, hypopigmentation, post-inflammatory hyperpigmentation, discoloration, dyspigmentation, sunscreen, contrast, harder to see, subtle erythema.
- **Hormonal context:** hormonal, menstrual, pregnancy, contraception, birth control, androgen, spironolactone, PCOS, jawline distribution.
- **Patient behavior:** picking, squeezing, scratching, rubbing, avoid touching, poor hygiene, adherence, nonadherence, skin picking, excoriation.
- **Treatment specificity:** topical, retinoid, adapalene, tretinoin, benzoyl peroxide, clindamycin, antibiotic, antifungal, steroid, emollient, oral, isotretinoin, biologic.
- **Image quality/visibility:** image quality, limited image, difficult to see, hard to visualize, poor lighting, subtle, not clearly visible, low resolution.

D Additional Mechanistic Analyses

This appendix provides additional analyses supporting the mechanistic interpretation of recommendation instability under controlled skin-tone perturbations. While the primary analyses in the main paper focus on clinically relevant instability in diagnosis and recommendation outputs, the following supplementary results further characterize the relationship between recommendation instability, discourse-level framing shifts, and perturbation magnitude.

D.1 Regression Predictors of Recommendation Instability

To further examine whether recommendation instability is reducible to diagnosis substitution alone, we fit a logistic regression model predicting whether any downstream recommendation component changed for a perturbation pair. Predictors included relaxed diagnosis flips, perturbation distance, lexical/discourse shift metrics, and model fixed effects. Continuous predictors were z-scored prior to fitting.

Figure 3 shows the resulting odds ratios. Relaxed diagnosis flips remained strongly associated with recommendation instability (OR: 2.67, $p < 0.001$), but several framing-related variables also remained significant after controlling for diagnosis changes. In particular, escalation-oriented lexical shifts (OR: 1.82, $p < 0.001$), reassurance/de-

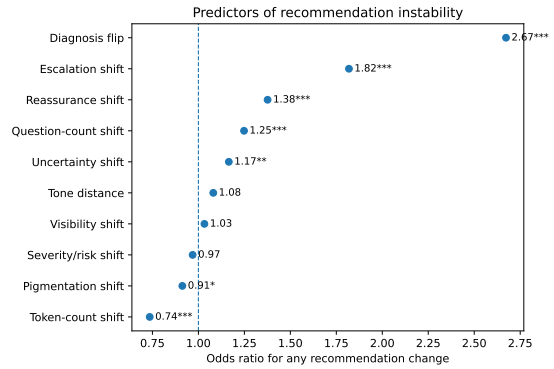


Figure 3: Odds ratios from a logistic regression predicting whether any recommendation changed under controlled skin-tone perturbations. The model includes relaxed diagnosis flips, perturbation distance, lexical/discourse shift metrics, and model fixed effects. Recommendation instability is associated not only with diagnosis changes, but also with escalation-oriented framing, reassurance shifts, uncertainty shifts, and interactional discourse variation. Statistical significance markers denote $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

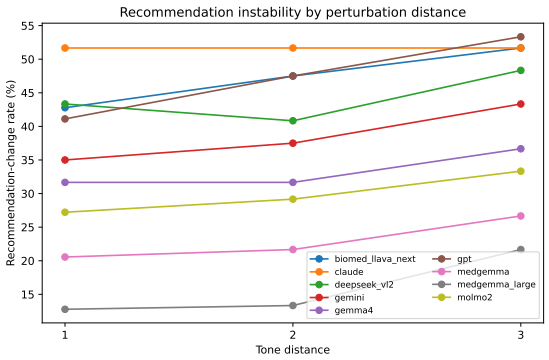


Figure 4: Recommendation-change rates across perturbation-distance levels for all evaluated models. Several models exhibit increasing instability as perturbation distance increases, suggesting that larger counterfactual skin-tone shifts may modestly amplify downstream recommendation instability.

escalation shifts (OR: 1.38, $p < 0.001$), uncertainty shifts (OR: 1.17, $p = 0.001$), and follow-up question shifts (OR: 1.25, $p < 0.001$) were associated with recommendation instability. Perturbation distance showed a weak positive trend (OR: 1.08, $p = 0.053$). These findings suggest that recommendation instability is coupled with broader discourse and framing variation rather than being explained solely by diagnostic substitution.

D.2 Perturbation-Distance Trends

We additionally examined whether larger counterfactual perturbations were associated with greater instability. Figure 4 shows recommendation-change rates across perturbation-distance levels for all evaluated models.

Although the magnitude of the effect varied across architectures, several models exhibited weak but consistent perturbation-distance trends. For example, GPT-5.4-mini increased from 41.1% recommendation instability for adjacent perturbations to 53.3% for the largest perturbations, while Biomed-LLaVA-Next increased from 42.8% to 51.7%. Similar patterns were observed for Gemini 3 Flash, Gemma 4 26B, MedGemma-27B, and DeepSeek-VL2. These findings are consistent with the interpretation that larger counterfactual skin-tone shifts may modestly amplify downstream recommendation instability.

Model	Diagnosis Flip	Any Rec. Change	Rec. Change Without Diagnosis Flip
GPT-5.4-mini	28.6	45.3	28.9
Claude Sonnet 4.6	41.7	51.7	26.1
DeepSeek-VL2	30.3	43.3	25.3
Gemini 3 Flash	33.6	37.2	22.8
Biomed-LLaVA-Next	41.1	45.8	21.1
Molmo2-8B	28.3	28.9	18.3
Gemma 4 26B	20.3	32.5	17.8
MedGemma-4B	26.9	21.9	12.2
MedGemma-27B	30.0	14.4	1.9

Table 3: Diagnosis-independent recommendation instability across models using relaxed diagnosis matching criteria. Values report percentages over paired perturbation comparisons. A substantial fraction of recommendation changes occur even when diagnosis predictions remain semantically stable, indicating that recommendation instability is not reducible to diagnosis substitution alone.

D.3 Diagnosis-Independent Recommendation Instability

Table 3 reports diagnosis-preserving recommendation instability across models using relaxed diagnosis matching criteria. A substantial fraction of recommendation changes occurred even when the predicted diagnosis remained semantically stable. This effect was particularly pronounced for GPT-5.4-mini, Claude Sonnet 4.6, DeepSeek-VL2, Gemini 3 Flash, and Biomed-LLaVA-Next, whereas MedGemma-27B exhibited substantially lower diagnosis-preserving instability.

These results further support the interpretation that recommendation instability cannot be fully reduced to diagnosis substitution alone, and may instead reflect broader behavioral and framing sensitivity under demographic perturbations.

E Reproducibility and Computational Details

All evaluations were performed using inference-only settings without model fine-tuning or additional training. Closed-source models were accessed through their respective APIs, while open-source models were evaluated using NVIDIA A100 80GB GPUs. To reduce stochastic variation unrelated to demographic perturbations, evaluations were conducted using deterministic or low-temperature decoding settings (e.g., temperature 0 where supported). Parameter sizes are reflected in model names when publicly available. The benchmark data and evaluation code are publicly available at https://github.com/robustdna/Beyond_Diagnosis/.

The SCIN dataset is released under the SCIN Data Use License, which permits reproduction, sharing, and creation of adapted materials subject to its attribution requirements and other license conditions. The license also prohibits attempts to re-identify or re-link de-identified data subjects. Our use and release of derived benchmark materials follow these terms.

F Physician Validation Protocol

All counterfactual skin-tone variants underwent physician review prior to benchmark inclusion. Reviewers evaluated each four-image skin-tone set jointly and assessed whether all variants preserved: (1) the same most likely diagnosis and (2) comparable clinical severity.

For each set, reviewers recorded whether diagnosis consistency and severity consistency were preserved, identified variants exhibiting clinically meaningful deviations, and optionally provided free-text comments describing observed issues. Variants were flagged if they exhibited altered lesion severity, clinically implausible morphology, inconsistent disease presentation, or image-generation artifacts that could affect downstream recommendation evaluation.

Examples of physician comments included observations such as "image 35 has lesser severity than the rest," "images 38 and 40 are more severe than the rest," and "image 12 does not seem to have vitiligo." Variants failing physician validation were regenerated and re-reviewed prior to final inclusion in the benchmark.

Physician validation was conducted by qualified medical reviewers, including external expert reviewers recruited through professional freelance platforms. Reviewers were compensated for their time and expertise.

G AI Assistance Disclosure

AI-based assistants were used for limited writing, and editing support during manuscript preparation. All study design, analyses, interpretations, and final manuscript decisions were performed and verified by the authors.

H Additional Robustness Analyses

H.1 Diagnosis Matching Criteria

For analyses conditioning on diagnosis preservation, model-generated diagnoses are normalized into broad canonical diagnostic category sets.

Under the relaxed matching criterion, a perturbation pair is considered diagnosis-preserving when the two canonical diagnosis sets share at least one category. For example, "atopic dermatitis" and "eczema", "tinea corporis" and "ringworm", and "plaque psoriasis" and "psoriasis" are treated as matches. Likewise, "psoriasis / eczema" and "eczema" are considered a match because the two responses share the eczema category. In contrast, "eczema" and "psoriasis" are treated as a diagnosis change.

This criterion is intentionally permissive to avoid counting surface-form variation or differences in additional differential diagnoses as diagnosis changes. To assess sensitivity to this definition, we additionally apply exact-set matching, under which the complete canonical diagnosis sets must be identical. As shown in Table 4, the stricter criterion produces higher absolute diagnosis-change rates, but incomplete diagnostic stability and substantial variation across models are observed under both criteria.

Criterion	Diagnosis-preserving condition	Diagnosis-change rate
Relaxed overlap	At least one canonical diagnosis category is shared	21.4–42.5%
Exact-set	The complete canonical diagnosis sets are identical	28.6–74.7%

Table 4: Sensitivity of diagnosis-change rates to alternative diagnosis-matching criteria. The exact-set criterion is stricter than the relaxed overlap criterion but yields the same qualitative conclusion of incomplete diagnostic stability under skin-tone perturbations.

H.2 Diagnostic Accuracy and Correctness-Conditioned Recommendation Stability

Although the primary objective of our benchmark is to evaluate counterfactual stability rather than general diagnostic performance, we additionally measure diagnostic accuracy across all 240 physician-validated images. Diagnostic accuracy ranges from 40.8% to 59.2% across the nine evaluated models.

To determine whether downstream recommendation instability is primarily driven by diagnostic errors or diagnosis substitution, we further restrict the analysis to eligible perturbation pairs for which the model predicts the correct diagnosis for both skin-tone variants. A recommendation is considered changed when at least one of the four downstream dimensions— management, urgency, referral, or follow-up—differs between the paired responses.

Table 5 reports diagnostic accuracy together with recommendation-change rates under this correctness-conditioned setting. For several models, substantial recommendation instability remains even when both variants receive the correct diagnosis. For example, recommendation changes occur in 49.6% of eligible pairs for DeepSeek-VL2, 44.9% for GPT-5.4-mini, and 34.6% for Gemini 3 Flash.

These results indicate that the observed downstream instability cannot be explained solely by diagnostic errors or diagnosis changes. This analysis does not establish the clinical correctness or utility of the generated recommendations; rather, it provides a robustness test showing that recommendation behavior may remain unstable even when diagnostic predictions are correct.

Model	Diagnostic accuracy	Rec. change when both correct
DeepSeek-VL2	48.3%	49.6% (70/141)
GPT-5.4-mini	59.2%	44.9% (80/178)
Gemini 3 Flash	57.1%	34.6% (53/153)
Biomed-LLaVA-Next	40.8%	33.7% (34/101)
Claude Sonnet 4.6	48.3%	31.9% (37/116)
Gemma 4 26B	58.3%	17.8% (33/185)
Molmo2-8B	45.8%	12.5% (17/136)
MedGemma-4B	46.7%	11.3% (16/142)
MedGemma-27B	48.8%	0.0% (0/155)

Table 5: Diagnostic accuracy across all 240 benchmark images and recommendation-change rates restricted to perturbation pairs for which both skin-tone variants receive the correct diagnosis. Recommendation change indicates that at least one of management, urgency, referral, or follow-up differs between the paired responses.

H.3 Case-Level Bootstrap Confidence Intervals

Because each clinical case contributes six pairwise skin-tone comparisons, the 360 within-case pairwise comparisons per model are not treated as independent observations for uncertainty estimation. We therefore perform a case-level clustered bootstrap using the 60 original clinical cases as the independent resampling units.

For each of 5,000 bootstrap iterations, we sample 60 clinical cases with replacement and retain all six pairwise comparisons associated with each selected case as a single cluster. We then recompute the overall recommendation-change rate and report percentile-based 95% confidence intervals.

Table 6 reports the resulting confidence intervals. The interval widths reflect the uncertainty associated with the limited number of source clinical cases. Accordingly, small differences between models should not be interpreted as statistically definitive rankings. The primary finding is instead that substantial recommendation instability is observed across multiple models even when uncertainty is estimated at the case level.

Model	Rec. change	95% CI
Claude Sonnet 4.6	51.7%	[41.9, 61.1]
Biomed-LLaVA-Next	45.8%	[37.8, 53.9]
GPT-5.4-mini	45.3%	[37.8, 53.3]
DeepSeek-VL2	43.3%	[35.3, 51.4]
Gemini 3 Flash	37.2%	[28.9, 45.8]
Gemma 4 26B	32.5%	[24.4, 40.6]
Molmo2-8B	28.9%	[20.3, 38.3]
MedGemma-4B	21.9%	[14.4, 29.7]
MedGemma-27B	14.4%	[8.3, 21.1]

Table 6: Overall recommendation-change rates with case-level clustered bootstrap 95% confidence intervals. Bootstrap resampling uses the 60 source clinical cases as independent units and retains all six within-case skin-tone comparisons associated with each resampled case.

H.4 Benchmark Composition and Validation Statistics

Table 7 summarizes the iterative automated filtering and physician-validation process used during benchmark construction. Table 8 reports the final benchmark composition across the six diagnosis categories. The final benchmark contains 60 source clinical cases and 240 physician-validated counterfactual images, with four skin-tone variants per clinical case.

Validation stage	Count
Flagged by GPT filtering	17
Flagged by Gemini filtering	4
Flagged during physician review	23
Final physician-validated images	240

Table 7: Summary of the iterative validation pipeline used to construct the physician-validated counterfactual dermatology benchmark. Variants flagged during automated filtering or physician review were regenerated and re-validated before final inclusion. Final inclusion required preservation of diagnosis, severity, and clinically realistic lesion morphology across all four skin-tone variants within each clinical case.

Condition	Images
Acne	40
Eczema	40
Psoriasis	40
Tinea	40
Vitiligo	40
Benign nevus	40
Total	240

Table 8: Dataset composition of the clinician-validated counterfactual dermatology benchmark. Each condition contains 40 physician-validated skin-tone variants derived from 10 source clinical cases.