nature portfolio

Peer Review File

Divergent impacts of explainable AI for dermatological diagnosis on clinicians versus lay people

Corresponding Author: Dr Marzyeh Ghassemi

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
A. Summary of Key Results
The paper evaluates how different explainable AI (XAI) methods affect diagnostic accuracy, deference to AI, and trust in AI systems among laypeople and primary care physicians in dermatological tasks. Results indicate that:
- LLMs improve accuracy for laypeople but also increase automation bias and overreliance, especially when the AI is incorrect.
- PCPs benefit from AI assistance without being misled by erroneous outputs.
- Human-first vs AI-first decision paradigms influence deference patterns, with AI-first inducing more anchoring bias.

B. Originality and significance
While the study is well-executed and thoroughly documented, I find that it lacks conceptual novelty. The core contributions—such as automation bias among lay users, greater AI resilience among experts, and the double-edged nature of explainability—have already been extensively documented in prior literature. However, framing the introduction to show that the study serves as a dermatology-specific replication of these findings rather than a novel advancement in human-AI interaction or explainable AI theory can help.

C. Data & methodology: validity of approach, quality of data, quality of presentation
From a methodological standpoint, the paper is rigorous. However, how confounding factors were controlled is unclear.

D. Appropriate use of statistics and treatment of uncertainties
The statistical analyses are generally appropriate. Use of paired and unpaired t-tests, effect size measures, and interaction terms is standard. However, given the volume of hypothesis testing and subgroup analysis, the authors should consider corrections for multiple comparisons or more formal hierarchical modeling. With many comparisons across subgroups and conditions, there is a high risk of false positives (Type I errors). Without adjustments or more structured modeling, such as multilevel models that account for repeated measures across images and participants, the reported p-values may overstate the reliability of the observed effects, especially when differences are marginal.

E. Conclusions: robustness, validity, reliability
Conclusions are well supported by the data but do not yield novel theoretical insights. The notion that semantic (LLM) explanations can mislead less experienced users while aiding calibration among experts is expected and aligns with prior studies on automation bias and cognitive trust in AI systems. The claim that LLMs are a “double-edged sword” is rhetorically strong but not conceptually new. A key limitation is that no real design interventions or mitigation strategies are proposed—such as counterfactual explanations, progressive disclosure, or user-adaptive interfaces—that could operationalize the authors’ findings into safer, more effective human-AI collaborations.

F. Suggested improvements:
Few critical confounds limit the interpretability of key findings, particularly those related to automation bias and user trust.

First, the experimental design does not provide participants, especially those in the general public cohort, with feedback about the correctness of their diagnostic decisions (please clarify if there was any feedback). This absence of performance feedback may lead participants to default to AI suggestions not due to overtrust or automation bias, but due to self-doubt or

uncertainty, especially when tasks are perceived as difficult. Thus, the observed AI deference might reflect perceived competence gaps, not necessarily trust miscalibration.

Second, it is inclear if and how prior familiarity with LLMs or AI-assisted tools was assessed and controlled for. Exposure to tools like ChatGPT, could significantly shape interpretive tendencies independent of explanation quality.

Third, the authors are encouraged to report the time taken by participants to complete the diagnostic tasks. In web-based experimental designs, fast task completion can indicate inattentiveness or satisficing behavior. Although attention check questions were included, it remains unclear whether any behavioral metrics (e.g., time on task) were used to assess participant engagement or attention. Without such controls, there is a risk that some participants, particularly lay users, clicked through tasks rapidly, which could artificially inflate observed patterns of AI deference.

(Remarks on code availability)

(Remarks on figshare data availability)

Reviewer #2

(Remarks to the Author)
Summary

The manuscript addresses how different types of AI explanations influence primary care physician and general public decision-making. It investigates deference to AI in dermatology settings using AI confidence, saliency maps, similar image retrieval, and LLM-based explanations. Two studies are presented with decent sample sizes. The general public deferred to AI more often even when AI was wrong, but physicians were more resilient to wrong AI predictions.

Strengths

The paper fills a gap in understanding how users interact with different types of AI explanations. Sample sizes are decent for both studies. The manuscript is well-structured and easy to follow. The inclusion of extensive subgroup analyses adds depth to the findings.

Major Concerns

The physicians were asked to choose from 445 skin diseases while the AI model only predicted from 34. This severely limits physician diagnostic accuracy without AI assistance and in turn the results. The physicians obviously performed better with AI assistance as the AI had less chances to be wrong (34 diseases compared to 445). If the physicians were informed about the 34 possible diagnoses or if the AI had to predict from 445 diseases, the deference numbers could have looked very different. In my opinion, this is the biggest limitation of study 2. If the physicians were informed, the conclusions drawn from the results could have been more valid.

Minor Concerns

The manuscript mentions "high quality labels" in the test set, but it's unclear how these were established. Were lesions biopsy-verified?

An overall AI accuracy is provided, but there is no information on AI accuracy per disease. How does the AI perform on the less common diseases? Were the participants mostly shown the common classes or were they also shown rare diseases? This information is important. If they were mainly shown common diseases, the AI should do quite well since there are more of these images to learn from. But the participants have to also consider rare diseases from a pool of 446 in their options.

Why AUC and not weighted AUC reported for study 2?

Participants were randomly assigned to one explanation type. This opens the possibility that observed differences stem from personality traits or decision-making styles (some tend to follow AI more and others tend to use own decisions more) rather than the explanation method. A more robust approach would have been to expose each participant to all 4 types of explanations, which would account for variability physician personalities.

What happens when the AI predicts melanoma on a symmetrical lesion, does the LLM explain it as asymmetrical even if it's clearly symmetrical? How are inconsistencies between model prediction and the LLM explanations handled?

Each participant evaluated only 12 images which is a bit on the low side for a study like this.

The discussion does not contain reasons why different explanation types had different deference. The authors should provide some reasoning (even if speculative) as to how the different explanation types impacted the participants. In my opinion, the LLMs had high deference for PCPs as it's intuitive to cross check things like "symmetry", "color irregularity", etc. compared to saliency maps, which highlight important regions but give no information on interpreting the regions. On the other hand, the LLM explanations provided more information on how to interpret the explanation. This may point to better usability of these concept-style explanations which can be discussed.

Other

Line 695 of the manuscript states that STab 3 provides the list of disease labels, but this should be STab4.

(Remarks on code availability)

(Remarks on figshare data availability)
See review text.

Reviewer #3

(Remarks to the Author)
This manuscript presents a significant advancement in elucidating human-AI collaboration in dermatology, through two large-scale reader studies that examine the differential impacts of various explainable AI (XAI) methods on diagnostic performance across populations with varying expertise levels. The work offers valuable insights into large language models (LLMs) as a "double-edged sword" in medical AI, highlighting novel patterns of automation bias and AI deference. The comparison between the general public and primary care physicians (PCPs) represents a major strength, providing a nuanced understanding of how expertise modulates AI assistance effects. However, several methodological, analytical, and interpretive issues must be addressed to ensure the robustness and clarity of the findings prior to publication.

1. A primary concern is the potential confounding between the fairness-constrained algorithms (CDANN) used in model training and the effects of the XAI methods evaluated. Both reader studies employ CDANN, which complicates the attribution of observed improvements—such as balanced accuracy across skin tones and reduced diagnostic disparities—to the XAI interventions rather than the underlying fairness training. For example, the abstract attributes these benefits to AI assistance via explanations, but they may stem from CDANN itself. The Model Training section on page 4 should explicitly clarify whether Study 2 also utilizes CDANN, as this is essential for interpreting the experimental design. Additionally, the manuscript should explain why different backbone models were employed in Study 1 and Study 2, and discuss how this choice might influence the results.

2. The selection of specific XAI methods—GradCAM, content-based image retrieval (CBIR), and multimodal LLMs—requires stronger justification, particularly in light of the extensive XAI literature in medical imaging and recent dermatology-specific advances, such as those in Chanda et al. [1] and Kim et al. [3]. Why were these approaches chosen over other cutting-edge methods, and are they sufficiently representative of existing XAI technologies? The information provided by these methods varies substantially (e.g., visual heatmaps in GradCAM versus textual reasoning in LLMs), raising questions about how their comparisons yield solid conclusions in a clinical setting. Furthermore, the manuscript notes that LLMs do not inherently analyze images; this should be elaborated upon, potentially through discussion of vision-language model (VLM) frameworks like LAVA, to clarify the multimodal integration.

3. The claim that well-trained AI can mitigate diagnostic biases while improving overall accuracy (page 6), with performance gaps reduced from 3.2% to 1.7%, contrasts with findings from Groh et al. [2], which reported that deep learning-based decision support exacerbates disparities in non-specialists' accuracy across skin tones. A more thorough engagement with this contradictory evidence is necessary, including explanations for the divergent conclusions. The current explanation for bias reduction appears shallow, focusing on CNNs or ViTs trained with fairness algorithms on large datasets without deeper analysis. The authors should discuss whether foundation models, such as self-supervised learning approaches like PanDerm [4] or vision-language models like MONET [3], could address fairness issues through domain-specific weight initialization and fine-tuning on similar datasets. Incorporating benchmarking results against these models would enhance reader understanding and guide future study designs.

4. Figure 2 indicates that all XAI methods improve diagnostic accuracy for the general public, yet this finding contradicts Chanda et al. [1], which showed no additional benefit from XAI over AI predictions alone. Further analysis is required to reconcile this discrepancy and identify conditions under which XAI provides value.

5. The manuscript reports that lay users were "misled by seemingly plausible reasons" from incorrect LLM advice, which decreased performance most notably (page 7). This critical claim needs direct substantiation: what renders an incorrect explanation "plausible"? Does it involve hallucinated clinical signs or accurate feature identification with flawed conclusions? Including examples of such explanations in the main text or supplement, along with a qualitative analysis, would be informative. Additionally, the stochasticity in LLM explanation generation introduces a confounder; the prompt in Supplementary Table 3, which instructs the model to "list criteria that you are certain [of]," may induce overconfidence, potentially driving the observed automation bias. The authors should discuss this limitation, clarify whether explanations were generated once or multiple times (and if cherry-picked), and address reproducibility. Recommendations for mitigation, such as uncertainty quantification, confidence calibration, or other safety measures, would strengthen the work's practical implications.

6. The PCP cohort reports an average of 6 years of experience (range 1-25), but no analysis examines how this variance affects results. Beyond skin tone, other major demographic factors (e.g., age, gender, or socioeconomic status) should be considered and reported if analyzed.

7. Task design simplifications warrant attention. For the general public, the binary melanoma-versus-nevus task may inflate AI's perceived impact, as laypersons typically assess lesions for worrisomeness rather than specific diagnoses. The ABCD rule for melanoma, referenced in the study, is not typically used in secondary care assessments. For PCPs, diagnoses rely on broader context (e.g., patient history, physical exams), which is absent here. These ecological validity concerns should be consolidated in a dedicated "Limitations" subsection. The attribution of higher deference in the AI-First condition to anchoring bias is plausible but overlooks alternatives like cognitive load or consistency bias. Reporting and analyzing decision times across conditions and rounds could disentangle these effects; if data are unavailable, they should be discussed.

8. While the manuscript provides insights into human-AI collaboration, it lacks formal, actionable recommendations for future studies and deployments. A structured framework, akin to Table 1 in Johri et al. [5], would enhance impact. The conclusion —that LLMs are a "double-edged sword" requiring "careful design"—is accurate but overly general. It should be expanded with bolder, concrete principles, such as: (1) preserving independent human judgment through staged workflows; (2) calibrating trust via uncertainty indicators; (3) adapting explanations to user expertise; and (4) highlighting AI-human disagreements to prompt re-evaluation.

In summary, this manuscript should resolve two core issues: (1) methodological confounders, including fairness algorithm effects, XAI method selection, and LLM stochasticity, which currently obscure the attribution of benefits; and (2) insufficient engagement with contradictory literature [1,2] and shallow analyses of discrepancies, risking misleading guidance on XAI's role in reducing diagnostic disparities. Addressing these, along with enhanced limitations discussion and practical recommendations, would position the work as a foundational contribution to medical AI.

References
[1] Chanda T, et al. Dermatologist-like explainable AI enhances trust and confidence in diagnosing melanoma. Nature Communications 2024.
[2] Groh M, et al. Deep learning-aided decision support for diagnosis of skin disease across skin tones. Nature Medicine 2024.
[3] Kim C, et al. Transparent medical image AI via an image–text foundation model grounded in medical literature. Nature Medicine 2024.
[4] Yan S, et al. A multimodal vision foundation model for clinical dermatology. Nature Medicine 2025.
[5] Johri, et al. An evaluation framework for clinical use of large language models in patient interaction tasks. Nature Medicine 2025.

(Remarks on code availability)

(Remarks on figshare data availability)

Version 2:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
Authors have addressed all my concerns. I have no further comments.

(Remarks on code availability)

(Remarks on figshare data availability)

Reviewer #2

(Remarks to the Author)
Comments on “Explainable AI as a Double-Edged Sword in Dermatology: The Impact on Clinicians versus The Public”

Summary
The authors present a methodologically strong analysis of the influence of XAI explanations on users of different experience levels. Their insights on the behavioral aspects among different user groups helps to further shape the application of X(AI) in clinical care and it underscores the point that there is no one-size-fits-all solution. The approach explicitly takes fairness and the performance among different skin types into account, a factor many studies fail to address.

Main Concern - Model Performance:
The models the authors use for their experiments perform poorly. For study 2, the authors report a balanced accuracy of 47.8% for the four main diseases and a balanced accuracy of 14.4% for the remaining 30 diseases. The decision to split the classification task into a primary and a secondary model further obfuscates the performance of the overall classification system and does not allow for an overall assessment of the system’s performance, including the propagation of errors between the subsystems.

These results show how these models are far from clinical readiness and cannot support the claim of providing ‘state-of-the-art AI suggestions’ as stated in the Discussion.

Post-hoc XAI methods, such as GradCAM or CBIR, directly depend on features generated by the underlying models. Without sufficiently well-performing models, the resulting explanations are unreliable and may not reflect meaningful, decision-relevant features.

Deliberately sampling images to match a predefined ratio of correct and incorrect predictions obscures the true performance of the system and evaluates XAI methods under conditions that do not reflect realistic model behavior

Although developing a state-of-the-art AI system was not the goal of this study, these points do weaken the authors’ argument regarding the ‘double-edged’ nature of XAI methods, as the quality and validity of the presented explanations are not sufficiently ensured.

Secondary Concern 1 - XAI quality evaluation:
A metric to evaluate XAI-explanations is introduced but not precisely defined. The authors chose to generate explanations based on the image and the label predicted by the base model, no matter the correctness of this label.
Does the evaluation metric define informativeness and correctness of explanations based on the ground truth label or the predicted labels? Did the dermatologists evaluating the explanations know the ground truth?

I.e. is a GradCAM heatmap that highlights a nevus, although the model predicts a melanoma (second example in SFig. 7) correct and/or informative? Similarly, is the explanation for the misclassified dysplastic nevus which is mentioned in the chapter on ‘Misplaced trust in LLM explanations for the general public’ considered a high or low quality explanation?

A more detailed explanation of the XAI-quality evaluation, and an acknowledgement of limitations would strengthen the comparison between high and low XAI Quality (EFig. 2) and subsequently all comparisons between XAI strategies.

An in-depth, model-agnostic analysis of explanation quality would also mitigate the effect of the impact of the main concern regarding low model performance.

Secondary Concern - 2 Choice of Task for Study 2:
The authors report a top-1 accuracy of 11.5% and a top-3 accuracy of 16.1% for PCPs without AI support. In combination with the similarly weak accuracies of AI models, this sparks doubt, whether the task at hand is solvable based on a single image, with or without AI.

Reporting misleading results for a solvable task would argue for a stronger case of the “double-edged” nature of XAI.

Secondary Concern 3 - Carry-Over Effects:
Each participant reviews the same 12 cases twice. All subsequent analysis stems from the comparison of the diagnostic accuracy per-image between the two runs. Carry-over effects apply. Other reviewers have pointed that out in-depth in the

Secondary Concern 4 - Small Sample Size:
Small n: Just 12 cases per participant. Other reviewers have pointed that out in-depth

Secondary Concern 5 - Human-First Performance Comparison:
Do the results shown in Fig. 2a) and 3a) only include results from the Human-First subgroup? Mentioning Rd.1 as ‘initial’ in

the caption of Fig. 2 points in that direction. The fact could be pointed out more clearly.

(Remarks on code availability)

(Remarks on figshare data availability)
The figshare link had expired.

Reviewer #3

(Remarks to the Author)
The authors have satisfactorily addressed my concerns raised in the previous round. The introduction of the "dose-response" framework to reconcile findings with Groh et al., the shift to Linear Mixed Models, and the "cognitive firewall" distinction are significant improvements. Below are my specific comments:

While the use of CDANN is methodologically sound, the manuscript (particularly the Abstract and Discussion) must carefully distinguish between benefits derived from the fairness-constrained model predictions (which affect all conditions, including Basic) versus benefits derived specifically from the XAI modalities. The reduction in diagnostic disparity appears to be primarily driven by the underlying CDANN training rather than the explanations themselves. Please ensure the text reflects this nuance to avoid overstating the independent role of XAI in equity.

Regarding the selection of XAI methods, the rationale for selecting GradCAM and CBIR as representative baselines is accepted. However, characterizing Concept-Based methods (e.g., Concept Bottlenecks) merely as 'novelty' is slightly reductive in the context of dermatology. Concept-based XAI represents the computational analog to standard clinical training. By contrast, the 'Multimodal LLM' arm in your study represents 'Unstructured Narrative' explanations. Please refine the Discussion to explicitly distinguish between Structured Concept-Based XAI (which aligns with standard dermatological criteria like ABCD) and the Unstructured Narrative Explanations (LLM) used in this study. It is important to clarify that the observed "Double-Edged Sword" effect may be driven specifically by the persuasive rhetoric of the LLM's narrative, which is a distinct mechanism from the structured concept scoring often used in clinical training.

To fully reconcile your findings with Chanda et al., please ensure the Results section explicitly states that, in the aggregate view, the advantage of advanced XAI over Basic AI is limited/mixed. The text should emphasize that the primary novelty lies in the conditional performance: the LLM modality drives high impact in correct-prediction scenarios but carries significant risk in error scenarios. This framing prevents readers from assuming a universal superiority of XAI.

Minor comments:
- Line 233 (Statistical Notation): The phrase "delta of detla" contains a typo and is slightly colloquial. Please revise to "difference in differences" or "differential impact."
- Line 521: Duplicate word: "explanations explanations."

(Remarks on code availability)

(Remarks on figshare data availability)

Response to Review #1

R1-Comment1 (R1-C1). Originality and Significance

While the study is well-executed and thoroughly documented, I find that it lacks conceptual novelty. The core contributions—such as automation bias among lay users, greater AI resilience among experts, and the double-edged nature of explainability—have already been extensively documented in prior literature. However, framing the introduction to show that the study serves as a dermatology-specific replication of these findings rather than a novel advancement in human-AI interaction or explainable AI theory can help.

We appreciate the reviewer's feedback. We agree that concepts like automation bias and expertise-based resilience are established in the broader HCI literature. Our intention is not to claim conceptual novelty regarding automation bias or expertise-driven differences, but to clarify our contribution in an area where prior findings have been inconsistent. As we describe in the manuscript and elaborate below, prior work has reached conflicting conclusions about how expertise affects resilience to incorrect AI predictions, and how different types of explanations help or harm diagnostic decision-making.

To directly reflect this, we revised our Introduction to clarify the two-fold motivation of our work. First, our study provides a **robust empirical confirmation** of well-established phenomena in human–AI interaction in clinical contexts, including automation bias and differential AI resilience [1, 2] (note that we add reference below point-by-point for easier readability). Second, we highlight that prior literature contains **contradictory conclusions** around the "double-edged" nature of AI explanations. Some studies show that expertise strengthens users' ability to resist incorrect model predictions [3], while others find that higher expertise does not reliably protect users from misleading AI recommendations [4], and that poorly designed explanations can further magnify errors [5]. These inconsistencies motivated us to simultaneously examine both the general public and medical experts under the same experimental framework.

Another key contribution of our work is that we systematically evaluate **LLM-based explanations**, which differ fundamentally from traditional feature-based XAI (e.g., heatmaps) in their semantic persuasiveness and have not been extensively studied in clinical human–AI collaboration. Our findings demonstrate that LLM explanations benefit both groups when the AI is correct, but only medical experts can resist misleading LLM-generated rationales. This divergence adds new empirical clarity to conflicting prior claims. It demonstrates that the "double-edged" nature of XAI is not uniform but is modulated by the type of explanation (semantic vs. visual) interacting with user expertise.

Finally, we expanded the Discussion section in line with R3's suggestion (R3-C8) by providing actionable design implications (e.g., adapting explanations to user expertise, preserving independent judgment, and adding safeguards against user deference, see more details in R1-C4). These revisions clarify the significance of our work in supporting safer and more effective human–AI collaboration.

Local References:
[1] Khera, R., Simon, M. A., & Ross, J. S. (2023). Automation bias and assistive AI: risk of harm from AI-driven clinical decision support. *JAMA*, 330(23), 2255–2257.
[2] Groh, M., Badri, O., Daneshjou, R., Koochek, A., Harris, C., Soenksen, L. R., ... & Picard, R. (2024). Deep learning-aided decision support for diagnosis of skin disease across skin tones. *Nature Medicine*, 30(2), 573–583.
[3] Quinn, T. P., Senadeera, M., Jacobs, S., Coghlan, S., & Le, V. (2021). Trust and medical AI: the challenges we face and the expertise needed to overcome them. *JAMIA*, 28(4), 890–894.
[4] Tschandl, P., Rinner, C., Apalla, Z., Argenziano, G., Codella, N., Halpern, A., ... & Kittler, H. (2020). Human–computer collaboration for skin cancer recognition. *Nature Medicine*, 26(8), 1229–1234.
[5] Rosenbacke, R., Melhus, Å., McKee, M., & Stuckler, D. (2024). How explainable artificial intelligence can increase or decrease clinicians' trust in AI applications in health care: systematic review. *JMIR AI*, 3, e53207.

R1-C2. Data & Methodology — Confounding Factors

From a methodological standpoint, the paper is rigorous. However, how confounding factors were controlled is unclear.

Thank you for this comment. This concern resonates with R2 (R2-C5) and R3 (R3-C6). We updated our analysis method and revised the Methods to explicitly describe how confounding factors were handled. Specifically, we updated all analyses from t-tests to more rigorous linear mixed models (LMMs) that incorporate relevant individual-level covariates. These include demographic variables (age, gender, race), prior human–AI collaboration experience (HAI scores), open-mindedness (AOT), critical thinking (CRT), prior skin condition experience, years of medical experience, and medical expertise level. We also clarified that our study excluded invalid responses using both attention-check items (SFig. 5c) and time-check filters (removing respondents with <10 seconds or >5 minutes per image on average). We added all LMM results in the supplementary materials and updated the reported statistics throughout the paper.

These revisions ensure that our modeling more transparently accounts for participant-level heterogeneity. Overall, the major findings and conclusions of the paper remain consistent after adjusting for these confounding factors.

R1-C3. Statistics & Multiple Comparisons

The statistical analyses are generally appropriate. Use of paired and unpaired t-tests, effect size measures, and interaction terms is standard. However, given the volume of hypothesis testing and subgroup analysis, the authors should consider corrections for multiple comparisons or more formal hierarchical modeling. With many comparisons across subgroups and conditions, there is a high risk of false positives (Type I errors). Without adjustments or more structured modeling, such as multilevel models that account for repeated measures across images and

participants, the reported p-values may overstate the reliability of the observed effects, especially when differences are marginal.

We appreciate this important feedback regarding statistical rigor. As detailed in our response to R1-C2, we have implemented formal hierarchical mixed-effect modeling as our primary analytical framework and replaced the previous paired/unpaired t-tests with LMMs. By explicitly modeling random intercepts and slopes for participants and image instances, this approach accounts for the repeated-measures structure of the data and the non-independence of observations. This transition inherently mitigates the risk of Type I error inflation associated with multiple independent pairwise comparisons.

Regarding specific pairwise contrasts, we utilized Estimated Marginal Means (EMMs) to conduct post-hoc analyses. This method provides a more robust basis for comparison by adjusting for the other covariates in the model. We also ensured structural validity by tailoring the model inclusion criteria to the specific hypothesis (therefore leading to multiple LMMs). For example, when analyzing metrics specific to XAI quality, the "Basic AI" condition was excluded (as it lacks XAI components), whereas it was included for overall diagnostic performance comparisons.

These rigorous statistical updates are reflected throughout the revised manuscript and detailed in the Supplementary Materials (STable 7–37). Our major findings and conclusions of the paper remain consistent under the new analysis framework. And we are confident that the shift to mixed-effect modeling and the use of EMMs effectively addresses the risk of false positives and satisfies the reviewer's concern regarding multiple comparisons.

R1-C4. Conclusions & Design Implications

Conclusions are well supported by the data but do not yield novel theoretical insights. The notion that semantic (LLM) explanations can mislead less experienced users while aiding calibration among experts is expected and aligns with prior studies on automation bias and cognitive trust in AI systems. The claim that LLMs are a "double-edged sword" is rhetorically strong but not conceptually new. A key limitation is that no real design interventions or mitigation strategies are proposed—such as counterfactual explanations, progressive disclosure, or user-adaptive interfaces—that could operationalize the authors' findings into safer, more effective human-AI collaborations.

We thank the reviewer for this helpful suggestion. We agree that identifying the problem is only the first step. In the revised Discussion ("Design Implications"), we have added a comprehensive set of design implications (merging this feedback with R3-C8) to operationalize our findings into safer and more effective human-AI workflows.

Our new implications paragraphs in the Discussion ("Design Implications") argue against a "one-size-fits-all" approach and propose the following tailored interventions:

- For the General Public: Systems should prioritize safety over persuasiveness. This includes explicitly displaying uncertainty/fallibility, framing suggestions as "preliminary information" rather than diagnoses, and favoring non-narrative XAI modalities (like CBIR) that encourage active comparison rather than passive acceptance of an LLM's narrative.
- For Clinicians: Systems should focus on "expert augmentation" rather than simple guidance. We propose adaptive explanations that offer concise rationales for routine cases but reveal in-depth evidence (such as similar cases or conflicting features) when the clinician's initial diagnosis differs from the AI.
- Workflow Design: We advocate for a "Human-First" paradigm in clinical settings. Our data shows that showing AI suggestions first anchors users; a staged workflow preserves independent human reasoning. Furthermore, we suggest using disagreement as a trigger for a "moment of reflection," transforming the AI interaction into an evidence-based dialogue.
- Adaptive Safeguards: Finally, we propose systems that dynamically infer user deference patterns. For highly deferential users, the system could trigger mandatory "second opinion" workflows or require confidence prompts before finalizing a decision.

By addressing the reviewer's concerns, we believe these new discussions enhance the contribution of our work.

R1-C5. Lack of Feedback During Experiment

First, the experimental design does not provide participants, especially those in the general public cohort, with feedback about the correctness of their diagnostic decisions (please clarify if there was any feedback). This absence of performance feedback may lead participants to default to AI suggestions not due to overtrust or automation bias, but due to self-doubt or uncertainty, especially when tasks are perceived as difficult. Thus, the observed AI deference might reflect perceived competence gaps, not necessarily trust miscalibration.

We have revised the manuscript to clarify that our studies were explicitly designed to mirror real-world diagnostic scenarios. In practice (whether a layperson using an app or a clinician using an EHR), users rarely receive immediate "ground truth" feedback on a case-by-case basis.

Providing feedback after each trial would have introduced significant learning effects, confounding our measurement of deference with the participant's rapid learning curve during the experiment. By withholding feedback, we captured the participants' natural state of uncertainty. We acknowledge in the Discussion ("Limitation") that while this mirrors ecological validity, future work could leverage feedback loops specifically to study how quickly trust can be calibrated over time.

R1-C6. Prior LLM Familiarity

Second, it is inclear if and how prior familiarity with LLMs or AI-assisted tools was assessed and controlled for. Exposure to tools like ChatGPT, could significantly shape interpretive tendencies independent of explanation quality.

We agree that prior exposure to AI tools is a relevant factor. Participants' prior AI-related experience is partially captured by the HAI (Human-AI Collaboration Experience) score, which was measured at the end of the task using a validated scale [1]. As described in R1-C2, in the revised analysis, the HAI measure has been included in all LMMs as a covariate to control for its impact.

We observed that, in some cases, the HAI score indeed had a small yet consistent positive impact on diagnostic performance (e.g., STab 7 and STab 10). However, accounting for this factor did not alter our main findings regarding the differential impact of LLM explanations on laypeople versus experts. We also acknowledge in the Discussion ("Limitations") that future studies would benefit from more targeted measures specifically regarding familiarity with Generative AI tools like ChatGPT.

Local References:
[1] Jian, J.-Y., Bisantz, A. M. & Drury, C. G. Foundations for an Empirically Determined Scale of Trust in Automated Systems. Int. J. Cogn. Ergon. (2000)

R1-C7. Time-on-Task Metrics

Third, the authors are encouraged to report the time taken by participants to complete the diagnostic tasks. In web-based experimental designs, fast task completion can indicate inattentiveness or satisficing behavior. Although attention check questions were included, it remains unclear whether any behavioral metrics (e.g., time on task) were used to assess participant engagement or attention. Without such controls, there is a risk that some participants, particularly lay users, clicked through tasks rapidly, which could artificially inflate observed patterns of AI deference.

We agree with the reviewer on the importance of monitoring engagement. We employed two specific quality control measures involving time. First, we clarify that we used both attention check and time check in our study, which can be found in the Method "Human Subject Study Design and Protocol" part. Specifically, we utilized attention check questions in the middle of the study (see an example in SFig. 5c) and we also filtered out results with abnormal answering time (i.e., less than 10 seconds or more than 5 minutes per image on average).

Second, response time is a particularly important confounding factor when comparing Human-First and AI-First paradigms (i.e., participants with the tendency towards AI deference – deferential participants – can have shorter response time). Therefore, we also included response time in the LMMs to control its effect. Our results indicate that response time had a significant impact on accuracy for PCPs (STab 35), but no effect for the general public or other outcomes in both the general public and PCP studies. Crucially, controlling for time did not change the main conclusions of our paper.

We have updated the Results (“Putting AI Before Human Decisions Amplifies Deference Across Expertise Levels”) and Methods (“Human Subject Study Design and Protocol”) sections to reflect these quality control steps and analyses.

Response to Reviewer #2

R2-Comment1(R2-C1): Major Concerns:

The physicians were asked to choose from 445 skin diseases while the AI model only predicted from 34. This severely limits physician diagnostic accuracy without AI assistance and in turn the results. The physicians obviously performed better with AI assistance as the AI had less chances to be wrong (34 diseases compared to 445). If the physicians were informed about the 34 possible diagnoses or if the AI had to predict from 445 diseases, the deference numbers could have looked very different. In my opinion, this is the biggest limitation of study 2. If the physicians were informed, the conclusions drawn from the results could have been more valid.

We appreciate the reviewer's comment and have clarified these points in the revised manuscript. In Study 2, clinicians performed a free-text diagnostic task. The 445-disease list served only as an autocomplete aid, not a restriction. Participants were free to enter any diagnosis in their own words, and many entered multi-condition or nuanced answers. To ensure accuracy in coding, our team manually reviewed all responses to avoid errors due to spelling or synonyms. To avoid confusion, we removed the sentence suggesting a "0.22% random baseline" accuracy.

To maintain ecological validity, our main analyses focused on four common dermatological conditions (atopic dermatitis, pityriasis rosea, Lyme, and CTCL), which comprised 10 of the 12 images shown. Each participant saw a balanced mix of images with correct and incorrect AI predictions. The remaining two images were drawn from a pool of 30 other conditions to reflect real-world diagnostic variability.

Although the AI solved a classification task among 34 classes and clinicians completed an open-ended diagnostic task, our purpose was not to compare human versus AI diagnostic accuracy. The goal of Study 2 was to examine how clinicians use AI predictions and explanations, consistent with prior reader study designs such as Groh et al. [1]. We have clarified this point in the revised text.

Local Reference:

[1] Groh, M., Badri, O., Daneshjou, R., Koochek, A., Harris, C., Soenksen, L. R., ... & Picard, R. (2024). Deep learning-aided decision support for diagnosis of skin disease across skin tones. Nature Medicine, 30(2), 573-583.

R2-C2: Minor Concern 1

The manuscript mentions 'high quality labels' in the test set, but it is unclear how these were established. Were lesions biopsy-verified?

Thank you for pointing this out. We have clarified the definition of "high-quality labels" directly in the Methods ("AI Models and Explanations"). We employed a strict three-step inclusion criteria:

- Resolution: Images must meet a minimum resolution threshold (400x400 pixels) to ensure visual features are discernible for human review.
- Verification: Diagnosis labels were verified. Since we are using existing datasets, the specific verification methods depend on the dataset. For datasets like DDI, this involved pathology reports from biopsies reviewed by board-certified dermatologists and dermatopathologists.
- Metadata: All Fitzpatrick skin-tone labels were human-annotated, rather than machine-generated, ensuring valid fairness evaluations.

Note that only a subset of images can fulfill all these criteria among images in the public datasets (STab 4), and we ensured that our model evaluation and user studies could leverage these high-quality images.

These details have been added to the Methods to ensure full transparency about how diagnostic and metadata labels were established in the test set.

R2-C3: Minor Concern 2

An overall AI accuracy is provided, but there is no information on AI accuracy per disease. How does the AI perform on the less common diseases? Were the participants mostly shown the common classes or were they also shown rare diseases? This information is important. If they were mainly shown common diseases, the AI should do quite well since there are more of these images to learn from. But the participants have to also consider rare diseases from a pool of 446 in their options.

Thank you for raising this important point. We have added Weighted AUROC and Per-disease diagnostic performance metrics to the Methods section (“AI Models and Explanations”) and Supplementary Tables 5–6. These additions provide a transparent view of model performance across both common and less common diseases.

In addition, related to R2-C1, we have clarified the disease distribution and task structure in both the Methods and Supplement.

First, in Study 2, clinicians completed an open-ended free-text diagnostic task, not a constrained multiple-choice classification. The list of 445 diseases functioned only as an autocomplete aid while typing. Participants were not restricted by the list and could enter any diagnosis in their own words, including comorbidities. To ensure coding accuracy, multiple authors manually reviewed all responses to account for spelling variations and synonyms. To avoid confusion, we removed the sentence referring to a “0.22% random baseline,” since it was not directly relevant to the behavioral analysis.

Second, our behavioral analyses focused on four main dermatological conditions—atopic dermatitis, pityriasis rosea, Lyme disease, and CTCL—which accounted for 10 of the 12 images shown to each participant. As described in the Study Design section, each participant saw:

- 2 images of the four main conditions with *correct* AI predictions (8 in total)
- 2 images of the same four main conditions with *incorrect* AI predictions (2 in total)
- 2 additional images randomly sampled from 30 other dermatologic diseases (not all 445), chosen to reflect real-world diagnostic diversity. These two additional images were included to emulate realistic clinical uncertainty rather than to evaluate rare-disease performance.

Third, although the AI solved a 34-class classification task while clinicians answered an open-ended diagnostic question, our goal in this study was not to compare human vs. AI diagnostic accuracy. Instead, consistent with prior work such as [1], our focus was on how participants interact with AI predictions and explanations.

Local Reference:
[1] Groh, M., Badri, O., Daneshjou, R., Koochek, A., Harris, C., Soenksen, L. R., ... & Picard, R. (2024). Deep learning-aided decision support for diagnosis of skin disease across skin tones. Nature Medicine, 30(2), 573-583.

R2-C4: Minor Concern 3

Why report AUC and not weighted AUC?

Thank you for this suggestion. We have now added weighted AUROC to complement the existing AUC metric. These values are included in the revised Methods (“AI Models and Explanations”) and STab 5–6. This provides a more complete representation of model performance across diseases with differing prevalence.

R2-C5: Minor Concern 4

Participants were randomly assigned to one explanation type. This opens the possibility that observed differences stem from personality traits or decision-making styles (some tend to follow AI more and others tend to use own decisions more) rather than the explanation method. A more robust approach would have been to expose each participant to all 4 types of explanations, which would account for variability physician personalities.

We appreciate this concern. While we agree with the reviewer’s point that within-subject design offers higher statistical power per participant and can account for individual difference, we opted for a between-subjects design to prevent the significant fatigue effects (lower engagement and response quality) and carryover learning effects (e.g., learning to read a Saliency Map might influence how one subsequently interprets a CBIR retrieval) that would occur if a participant evaluated 48 images (12 images x 4 methods).

To address the reviewer’s concern about personality and individual differences (resonating with R1-C2 and R3-C6), we updated our analysis method and revised the Methods to explicitly describe how confounding factors were handled. Specifically, we updated all analyses from t-

tests to more rigorous linear mixed models (LMMs) that incorporate relevant individual-level covariates. In particular, for Study 2 with physicians, we included the following factors in our LLMs:

- Demographics: age, gender, race.
- HAI: Prior Human–AI collaboration experience.
- Cognitive Traits: Open-mindedness (AOT) and Critical Thinking (CRT).
- Expertise: Prior skin-condition experience and years of practice.

By including these factors in the model, we statistically account for the variability in decision-making styles, mitigating the risk that our results are driven by personality traits rather than the explanation methods themselves.

R2-C6: Minor Concern 5

What happens when the AI predicts melanoma on a symmetrical lesion, does the LLM explain it as asymmetrical even if it's clearly symmetrical? How are inconsistencies between model prediction and the LLM explanations handled?

LLM explanations were generated directly from the model's prediction and the input image without human correction, even when the prediction was *incorrect*. This is a central feature of our study design. LLM is not prompted to perform predictions, but only prompted to provide explanations. If the model made an incorrect prediction, the LLM was prompted to explain *that* specific diagnosis. For instance, the prompt to the LLM in Study 2 reads: "*...This image is recognized as '{given_label}'. Explain to a general medical expert with basic dermatology knowledge the reason for this diagnosis…*" (see the complete prompt structure in STab 3). We deliberately did not edit these "hallucinations" because understanding how participants respond to both correct and incorrect rationales is central to our research question.

Three dermatologists rated each explanation for correctness and informativeness, and these expert ratings reflect cases where the LLM referenced incorrect features or clinically irrelevant cues. Such examples are shown in EFig. 8 - 10, and SFig. 9, 12. As expected, some LLM-generated explanations are rated as incorrect or not informative (therefore, marked as "low-quality" explanations). Understanding the impact of the quality of these explanations is one of the research questions in our paper, as detailed in the revised Results and Discussion. For instance, EFig. 2c shows that while high-quality explanations helped, low-quality LLM explanations were particularly dangerous for laypeople, who trusted them regardless of the visual evidence. In contrast, experts were able to spot these inconsistencies, effectively using their domain knowledge as a "cognitive firewall" against the LLM's hallucination.

R2-C7: Minor Concern 6

Each participant evaluated only 12 images which is a bit on the low side for a study like this.

Thank you for raising this concern. We acknowledge this limitation. Our selection of 12 images was a trade-off to balance ecological validity with participant burden, aligning with sample sizes

in similar reader studies (e.g., 10 cases in [1], 8 cases in [2], 15 cases in [3]). Our online study required 10–15 minutes to complete, and using substantially more images may risk reduced engagement and lower data quality. On the other hand, we do agree with the reviewer that a larger image set for each participant could support more granular analyses (e.g., adopting a within-subject design as discussed in R2-C5). We note this as a limitation in the Discussion ("Limitation").

Local Reference:
[1] Groh, M., Badri, O., Daneshjou, R., Koochek, A., Harris, C., Soenksen, L. R., ... & Picard, R. (2024). Deep learning-aided decision support for diagnosis of skin disease across skin tones. Nature Medicine, 30(2), 573-583.
[2] Gaube, Susanne, et al. "Do as AI say: susceptibility in deployment of clinical decision-aids." NPJ digital medicine 4.1 (2021): 31.
[3] Chanda, Tirtha, et al. "Dermatologist-like explainable AI enhances trust and confidence in diagnosing melanoma." Nature Communications 15.1 (2024): 524.

R2-C8: Minor Concern 7

The discussion does not contain reasons why different explanation types had different deference. The authors should provide some reasoning (even if speculative) as to how the different explanation types impacted the participants. In my opinion, the LLMs had high deference for PCPs as it's intuitive to cross check things like "symmetry", "color irregularity", etc. compared to saliency maps, which highlight important regions but give no information on interpreting the regions. On the other hand, the LLM explanations provided more information on how to interpret the explanation. This may point to better usability of these concept-style explanations which can be discussed.

We thank the reviewer for this helpful suggestion. We have expanded the Discussion to explicitly theorize why LLM-based explanations induced different behaviors compared to visual methods (GradCAM/CBIR) :

- Narrative Persuasion (the general public): We argue that LLMs provide a "cohesive and seemingly complete diagnostic story". Unlike GradCAM (which requires interpreting an abstract heatmap) or CBIR (which requires active visual comparison), the LLM reduces cognitive effort by connecting features to a conclusion with authoritative prose. This "narrative ease" likely anchors lay users, creating an illusion of understanding even when the AI is wrong.
- Cognitive Firewall (Experts): For PCPs, the semantic nature of LLMs allows for easier verification. An expert can instantly cross-check a specific claim (e.g., "The lesion has multiple colors") against their observation. If the text explanation contradicts their structured medical knowledge, they can confidently reject it. This may explain why experts showed resilience to LLMs specifically, while laypeople did not.

We added these points as explicit discussion content in the Discussion section.

R2-C9: Other Concern:

Line 695 of the manuscript states that STab 3 provides the list of disease labels, but this should be STab4.

We thank the reviewer for pointing out our editing inconsistencies. We have double-checked and corrected all table and figure labels.

Response to Reviewer #3

R3-Comment1 (R3-C1):

A primary concern is the potential confounding between the fairness-constrained algorithms (CDANN) used in model training and the effects of the XAI methods evaluated. Both reader studies employ CDANN, which complicates the attribution of observed improvements—such as balanced accuracy across skin tones and reduced diagnostic disparities—to the XAI interventions rather than the underlying fairness training. For example, the abstract attributes these benefits to AI assistance via explanations, but they may stem from CDANN itself. The Model Training section on page 4 should explicitly clarify whether Study 2 also utilizes CDANN, as this is essential for interpreting the experimental design. Additionally, the manuscript should explain why different backbone models were employed in Study 1 and Study 2, and discuss how this choice might influence the results.

Choice of Algorithms and Backbones

Thank you for raising this critical point regarding experimental controls. We clarify that both Study 1 and Study 2 employed fairness-aware training pipelines. We systematically evaluated multiple algorithms (ERM, DANN, CDANN, GroupDRO, EMA) and selected the final model based on a "fairness-first" criterion: maximizing worst-group AUROC, followed by optimizing overall performance trade-offs. Consequently, CDANN was selected and utilized for the final models in both studies. We now report the specific model backbones and hyperparameters used in each study and clarify their rationale and consistency. Please see the updated Methods (“AI Models and Explanations”) text, e.g., Study 1: ViT-B/32 (ImageNet-21k) with CDANN; Study 2 Primary Model: DenseNet-121 (ImageNet-1k) with CDANN, together with per-disease metrics included in STab 5 & 6.

Confounding between fairness algorithms and explanations

We agree that the training algorithm (CDANN) shapes the model’s internal representations, which inevitably influences the post-hoc explanations. However, this algorithm was held constant across all experimental conditions within each study. Therefore, while CDANN influences the generation of the explanation, it does not confound the comparison between XAI modalities (GradCAM vs. CBIR vs. LLM), as the underlying model behavior was identical for all participants. Ultimately, any study evaluating post-hoc XAI methods must be interpreted in the context of the underlying model it seeks to explain. The complex interaction between a model’s training objective (accuracy, fairness, or robustness) and the behavior of its explanations is a general limitation of the field. We have added a statement in the Discussion (“Limitations”) section acknowledging that post-hoc XAI evaluations are inherently tied to the specific model architecture and training objectives employed.

R3-C2:

The selection of specific XAI methods—GradCAM, content-based image retrieval (CBIR), and multimodal LLMs—requires stronger justification, particularly in light of the extensive XAI literature in medical imaging and recent dermatology-specific advances, such as those in Chanda et al. [1] and Kim et al. [3]. Why were these approaches chosen over other cutting-edge methods, and are they sufficiently representative of existing XAI technologies? The information provided by these methods varies substantially (e.g., visual heatmaps in GradCAM versus textual reasoning in LLMs), raising questions about how their comparisons yield solid conclusions in a clinical setting. Furthermore, the manuscript notes that LLMs do not inherently analyze images; this should be elaborated upon, potentially through discussion of vision-language model (VLM) frameworks like LAVA, to clarify the multimodal integration.

We appreciate the reviewer's insightful comment regarding the selection of our XAI methods. We have revised the Introduction and Methods to provide a stronger justification for our choices and to clarify the technical nature of the multimodal LLM used.

Rationale for XAI Method Selection

We carefully selected GradCAM, CBIR, and Multimodal LLMs not to benchmark the absolute "best" performing algorithms, but to investigate the three fundamental paradigms of explanation currently available to clinicians. Our goal was to maximize the ecological validity and generalizability of our findings regarding human–AI interaction behaviors:

- Representativeness over Novelty: While we acknowledge recent advances in dermatology-specific XAI (such as concept bottlenecks or specific segmentation approaches), GradCAM and CBIR remain the "canonical" representatives of Feature Attribution (visual heatmaps) and Case-Based Reasoning (visual analogy), respectively. A recent systematic review confirms that these remain the two most commonly used XAI techniques in dermatology [1]. By using these standard methods, our findings on human behavior (e.g., automation bias) are more likely to generalize to other clinical systems than if we had used "cutting-edge" methods.
- Distinct Modalities: We aimed to cover the spectrum of cognitive scaffolding: (1) GradCAM: Provides spatial localization ("Where to look"); (2) CBIR: Provides visual evidence ("What it looks like"); (3) LLM: Provides semantic reasoning ("Why it is so"). This distinction allows us to draw solid conclusions about how different types of information (visual vs. textual) impact clinical trust, rather than comparing minor architectural variations within the same modality.

Clarification on Multimodal LLMs

We appreciate the opportunity to clarify the technical implementation of the multimodal LLM condition. We apologize for any confusion caused by the terminology. In our study, we utilized GPT-4V, which was one of the state-of-the-art Vision-Language Models (VLMs) capable of

inherently analyzing images. We have revised the Methods and Introduction to explicitly state that the model processed both the visual data (the clinical image) and textual prompts. It was not a text-only model relying solely on metadata. As explained in R3-C1, to ensure a controlled experimental design, we did not ask the VLM to predict the diagnosis from scratch (which would introduce a confounding variable of varying AI accuracy across conditions). Instead, we prompted the VLM to provide an explanation for the diagnosis predicted by our primary DenseNet model. This ensured that the "AI Accuracy" variable remained consistent across GradCAM, CBIR, and LLM conditions, allowing us to isolate the effect of the explanation modality itself.

We have standardized the terminology to "Multimodal LLM" (which is a widely used term in computational fields [2,3]) in the Introduction and Methods sections to reflect the underlying VLM framework, while retaining "LLM" in the Results section to emphasize that the output modality experienced by the user was natural language.

Local Reference
[1] Hauser, Katja, et al. "Explainable artificial intelligence in skin cancer recognition: A systematic review." European Journal of Cancer 167 (2022): 54-69.
[2] Wu, Shengqiong, et al. "Next-gpt: Any-to-any multimodal llm." Forty-first International Conference on Machine Learning. 2024.
[3] Wu, Jiayang, et al. "Multimodal large language models: A survey." 2023 IEEE International Conference on Big Data (BigData). IEEE, 2023.

R3-C3:

The claim that well-trained AI can mitigate diagnostic biases while improving overall accuracy (page 6), with performance gaps reduced from 3.2% to 1.7%, contrasts with findings from Groh et al. [2], which reported that deep learning-based decision support exacerbates disparities in non-specialists' accuracy across skin tones. A more thorough engagement with this contradictory evidence is necessary, including explanations for the divergent conclusions. The current explanation for bias reduction appears shallow, focusing on CNNs or ViTs trained with fairness algorithms on large datasets without deeper analysis. The authors should discuss whether foundation models, such as self-supervised learning approaches like PanDerm [4] or vision-language models like MONET [3], could address fairness issues through domain-specific weight initialization and fine-tuning on similar datasets. Incorporating benchmarking results against these models would enhance reader understanding and guide future study designs.

Thank you for raising this crucial point. We have expanded the Discussion (second paragraph) and thoroughly engaged with the apparent contradiction between our findings on bias reduction and those of Groh et al [1].

We argue that these findings are not contradictory but complementary. The finding in [1] that AI exacerbated disparities was specifically linked to a "control" model with moderate accuracy

(47%). However, the same study noted that a higher-accuracy "treatment" model (84% accuracy) did not exacerbate disparities. Our model, with a similarly high accuracy (79.2% expected accuracy in Study 2), aligns with their high-accuracy condition. We have synthesized these findings to propose a critical "dose-response" relationship between AI accuracy and diagnostic equity: Low-Accuracy AI acts as a source of distracting noise, amplifying uncertainty and errors in challenging cases (often dark skin images), thereby worsening disparities; High-Accuracy AI functions as a reliable decision support tool that overcomes baseline uncertainty, thereby neutralizing or reducing disparities. We added these insights in the Discussion (the third paragraph).

Regarding the suggestion to benchmark against foundation models like PanDerm or MONET: Similar to our response in R3-C2, while we agree this is a vital technical question, our study's primary focus is on human behavioral responses to XAI, not SOTA model benchmarking. We prioritized controlling the AI Correctness and skin conditions variable to ensure internal validity for the user study, as shown in Results (“Study Design”). Consequently, a full benchmarking of various model architectures was beyond the scope of this human-subjects study. We acknowledge this as a limitation and agree with the reviewer that exploring how advanced models can improve fairness through better domain-specific feature extraction is a critical direction for future work. We have expanded our Discussion (“Limitation”) to highlight this.

Local Reference:
[1] Groh, M., Badri, O., Daneshjou, R., Koochek, A., Harris, C., Soenksen, L. R., ... & Picard, R. (2024). Deep learning-aided decision support for diagnosis of skin disease across skin tones. Nature Medicine, 30(2), 573-583.

R3-C4:

Figure 2 indicates that all XAI methods improve diagnostic accuracy for the general public, yet this finding contradicts Chanda et al. [1], which showed no additional benefit from XAI over AI predictions alone. Further analysis is required to reconcile this discrepancy and identify conditions under which XAI provides value.

Thank you for this observation. Upon closer inspection, our findings are actually consistent with Chanda et al. [1]. Our Basic AI method (Prediction + Confidence) provides minimal explanations with model confidence, which is close to the AI condition in [1]. As shown in Fig. 2b and STab 7, while GradCAM, CBIR, and LLM showed higher average improvements than the Basic AI condition, these differences were not always statistically significant in the aggregate view. The same for Fig. 3b with the results in STab 16. This aligns with Chanda et al.'s finding that more advanced XAI does not outperform simple AI predictions.

Our contribution lies in the granular breakdown (AI Correct vs. AI Incorrect). The novel finding is not that "XAI is always better," but that semantic (LLM) explanations act as a double-edged sword: they provided significantly higher improvement when the AI was correct (+13.4%), but

caused the steepest performance drop when the AI was wrong (-21.1%), significantly more so than visual methods. We have clarified this nuance in the Discussion (first three paragraphs).

Local Reference:
[1] Chanda, Tirtha, et al. "Dermatologist-like explainable AI enhances trust and confidence in diagnosing melanoma." Nature Communications 15.1 (2024): 524.

R3-C5:

The manuscript reports that lay users were "misled by seemingly plausible reasons" from incorrect LLM advice, which decreased performance most notably (page 7). This critical claim needs direct substantiation: what renders an incorrect explanation "plausible"? Does it involve hallucinated clinical signs or accurate feature identification with flawed conclusions? Including examples of such explanations in the main text or supplement, along with a qualitative analysis, would be informative. Additionally, the stochasticity in LLM explanation generation introduces a confounder; the prompt in Supplementary Table 3, which instructs the model to "list criteria that you are certain [of]," may induce overconfidence, potentially driving the observed automation bias. The authors should discuss this limitation, clarify whether explanations were generated once or multiple times (and if cherry-picked), and address reproducibility. Recommendations for mitigation, such as uncertainty quantification, confidence calibration, or other safety measures, would strengthen the work's practical implications.

Plausibility of Incorrect Explanations

We have added specific examples and clarifications to the Results (Study 1) and Discussion (third paragraphs) to clarify and discuss "plausibility".

As illustrated in EFig 8-10 (which we now explicitly reference in the text), the multimodal LLM rarely hallucinated non-existent clinical signs (i.e., the visual analytic capability of GPT-4V is solid). Instead, it frequently referenced ambiguous dermatologic criteria (such as "irregular border," "color variation," or "asymmetry") even when these features were only subtly present or visually debatable. This created an illusion of understanding where the authoritative narrative tone prompted non-experts to reinterpret the ambiguous visual data to match the text. We added a specific case study from EFig 10 – a dysplastic nevus with wrong AI prediction as melanoma, and thus wrong LLM explanations – at the end of the Results (Study 1) section to illustrate a case with accurate feature identification but flawed conclusions. Besides, we also added new paragraphs to describe this mechanism explicitly in the Discussion (the third paragraph).

Generation procedure

To clarify our methodology, each LLM-based explanation was generated a single time via the GPT-4V API for each clinical case and was used directly without cherry-picking. We avoided

merging multiple outputs to prevent inconsistency, since the model's stochasticity naturally produces response variation across generations. We agree with the reviewer that the stochasticity and the potential overconfidence introduce potential confounders. The inherent variability in LLM generation means that the specific tone, phrasing, or framing of the single explanation presented to participants could have influenced their perception and decision-making, acting as an uncontrolled variable.

On the one hand, we mitigated this by having three dermatologists rate the explanations for quality, ensuring we accounted for the output's validity in our analysis. On the other hand, we did not account for subtle rhetorical differences between other potential generations. This stochasticity presents a challenge for direct replication, and we acknowledged it as a limitation in Discussion ("Limitation"). To ensure reproducibility of our analysis results, we will open-source the experimental data, including all multiumodal LLM-generated explanations.

We also appreciate the reviewer's suggestion on uncertainty quantification and confidence calibration. We have added paragraphs in Discussion ("Design Implications") to include this aspect.

R3-C6:

The PCP cohort reports an average of 6 years of experience (range 1–25), but no analysis examines how this variance affects results. Beyond skin tone, other major demographic factors (e.g., age, gender, or socioeconomic status) should be considered and reported if analyzed.

This important concern resonates with other reviewers (R1-C2, R2-C5). We updated our analysis method and revised the Methods to explicitly describe how confounding factors were handled. Specifically, we updated all analyses from t-tests to more rigorous linear mixed models (LMMs) that incorporate relevant individual-level covariates. In particular, for Study 2 with physicians, we included the following factors in our LLMs:

- Demographics: age, gender, race.
- HAI: Prior Human–AI collaboration experience.
- Cognitive Traits: Open-mindedness (AOT) and Critical Thinking (CRT).
- Expertise: Prior skin-condition experience and years of practice.

By including these factors in the model (e.g., STab 16), we statistically account for the variability in decision-making styles, mitigating the risk that our results are driven by personality traits rather than the explanation methods themselves.

R3-C7:

Task design simplifications warrant attention. For the general public, the binary melanoma-versus-nevus task may inflate AI's perceived impact, as laypersons typically assess lesions for worrisomeness rather than specific diagnoses. The ABCD rule for

melanoma, referenced in the study, is not typically used in secondary care assessments. For PCPs, diagnoses rely on broader context (e.g., patient history, physical exams), which is absent here. These ecological validity concerns should be consolidated in a dedicated "Limitations" subsection. The attribution of higher deference in the AI-First condition to anchoring bias is plausible but overlooks alternatives like cognitive load or consistency bias. Reporting and analyzing decision times across conditions and rounds could disentangle these effects; if data are unavailable, they should be discussed.

Limited Ecological Validity

We appreciate this comment and agree with the reviewer. Although we designed our user studies to be close to real-world scenarios, the ecological validity of the studies is still limited. However, it's hard to control these aspects in the online user study setup. We have consolidated these aspects into a dedicated discussion point in Discussion ("Limitation") in the Discussion.

Analysis of Decision Time when Comparing Human-First and AI-First

We thank the reviewer for this helpful suggestion. In Results ("Putting AI Before Human Decisions Amplifies Deference Across Expertise Levels"), we also included response time in the LMMs to control its effect. Our results indicate that response time had a significant impact on accuracy for PCPs (STab 35), but no effect for the general public or other outcomes in both the general public and PCP studies. Overall, controlling for time did not change the main conclusions of our paper.

R3-C8:

While the manuscript provides insights into human-AI collaboration, it lacks formal, actionable recommendations for future studies and deployments. A structured framework, akin to Table 1 in Johri et al. [5], would enhance impact. The conclusion—that LLMs are a "double-edged sword" requiring "careful design"—is accurate but overly general. It should be expanded with bolder, concrete principles, such as: (1) preserving independent human judgment through staged workflows; (2) calibrating trust via uncertainty indicators; (3) adapting explanations to user expertise; and (4) highlighting AI-human disagreements to prompt re-evaluation.

We deeply appreciate the reviewer for this insightful advice on different angles on design principles. In the revised Discussion ("Design Implications"), we have added a comprehensive set of design implications (merging this feedback with R1-C4) to operationalize our findings into safer and more effective human-AI workflows.

Our new implications paragraphs in the Discussion ("Design Implications") argue against a "one-size-fits-all" approach and propose the following tailored interventions:

- For the General Public: Systems should prioritize safety over persuasiveness. This includes explicitly displaying uncertainty/fallibility, framing suggestions as "preliminary

information" rather than diagnoses, and favoring non-narrative XAI modalities (like CBIR) that encourage active comparison rather than passive acceptance of an LLM's narrative.

- For Clinicians: Systems should focus on "expert augmentation" rather than simple guidance. We propose adaptive explanations that offer concise rationales for routine cases but reveal in-depth evidence (such as similar cases or conflicting features) when the clinician's initial diagnosis differs from the AI.
- Workflow Design: We advocate for a "Human-First" paradigm in clinical settings. Our data shows that showing AI suggestions first anchors users; a staged workflow preserves independent human reasoning. Furthermore, we suggest using disagreement as a trigger for a "moment of reflection," transforming the AI interaction into an evidence-based dialogue.
- Adaptive Safeguards: Finally, we propose systems that dynamically infer user deference patterns. For highly deferential users, the system could trigger mandatory "second opinion" workflows or require confidence prompts before finalizing a decision.

By adopting the advice from the reviewer, we believe these new discussions enhance the contribution of our work.

R3-C9:

In summary, this manuscript should resolve two core issues: (1) methodological confounders, including fairness algorithm effects, XAI method selection, and LLM stochasticity, which currently obscure the attribution of benefits; and (2) insufficient engagement with contradictory literature [1,2] and shallow analyses of discrepancies, risking misleading guidance on XAI's role in reducing diagnostic disparities. Addressing these, along with enhanced limitations discussion and practical recommendations, would position the work as a foundational contribution to medical AI.

We want to reiterate our appreciation of R3's constructive feedback. In response to the two core issues:

(1) Methodological Confounders: We have clarified that the fairness algorithm (CDANN) was held constant across all conditions to isolate XAI effects, justified our selection of XAI methods to ensure generalizability, and addressed LLM stochasticity by controlling for explanation quality and committing to open-source data. We also replace paired/unpaired t-tests with more robust LMMs to control other confounding factors during data analysis.

(2) Engagement with Literature: We have reconciled apparent contradictions with Groh et al., Chanda et al., and more literature. For instance, we add the analysis on the "dose-response" relationship between AI accuracy and fairness, and highlight the "double-edged" nature of LLMs stems from their specific interaction with user expertise, rather than a generic performance advantage.

Response to Reviewer #1

Authors have addressed all my concerns. I have no further comments.

We appreciate all previous comments proposed by R1, which fixed potential loopholes and lifted this work to the next level. We believe that our work has been improved significantly with the help of R1's helpful comments.

Response to Reviewer #2

R2-Comment1 (R2-C1) Model Performance and Task Evaluation

The models the authors use for their experiments perform poorly. For study 2, the authors report a balanced accuracy of 47.8% for the four main diseases and a balanced accuracy of 14.4% for the remaining 30 diseases. The decision to split the classification task into a primary and a secondary model further obfuscates the performance of the overall classification system and does not allow for an overall assessment of the system's performance, including the propagation of errors between the subsystems. These results show how these models are far from clinical readiness and cannot support the claim of providing 'state-of-the-art AI suggestions' as stated in the Discussion. Post-hoc XAI methods, such as GradCAM or CBIR, directly depend on features generated by the underlying models. Without sufficiently well-performing models, the resulting explanations are unreliable and may not reflect meaningful, decision-relevant features. Deliberately sampling images to match a predefined ratio of correct and incorrect predictions obscures the true performance of the system and evaluates XAI methods under conditions that do not reflect realistic model behavior. Although developing a state-of-the-art AI system was not the goal of this study, these points do weaken the authors' argument regarding the 'double-edged' nature of XAI methods, as the quality and validity of the presented explanations are not sufficiently ensured.

A metric to evaluate XAI-explanations is introduced but not precisely defined. The authors chose to generate explanations based on the image and the label predicted by the base model, no matter the correctness of this label. Does the evaluation metric define informativeness and correctness of explanations based on the ground truth label or the predicted labels? Did the dermatologists evaluating the explanations know the ground truth? I.e. is a GradCAM heatmap that highlights a nevus, although the model predicts a melanoma (second example in SFig. 7) correct and/or informative? Similarly, is the explanation for the misclassified dysplastic nevus which is mentioned in the chapter on 'Misplaced trust in LLM explanations for the general public' considered a high or low quality explanation? A more detailed explanation of the XAI-quality evaluation, and an acknowledgement of limitations would strengthen the comparison between high and low XAI Quality (EFig. 2) and subsequently all comparisons between XAI strategies. An in-depth, model-agnostic analysis of explanation quality would also mitigate the effect of the impact of the main concern regarding low model performance.

We appreciate the reviewer's feedback. We agree that the AI models used in Study 2 show moderate performance and are not intended to claim clinical readiness. In the current manuscript, we report the balanced accuracy for the 5-class primary model (0.478) and the 30-class secondary model (0.141), together with AUROC/weighted AUROC and per-disease metrics (Methods: AI Models and Explanations; Supplementary Tab. 5–6). To avoid any ambiguity introduced by the hierarchical ("primary + secondary") setup, we have now added the end-to-end performance of the combined system on the full held-out test set, computed using the same routing rule used in the experiment (secondary model invoked only when the primary model predicts "other"). We further report overall accuracy, AUROC, and weighted AUROC for

the combined system as 0.197, 0.734, and 0.755, including the same subgroup breakdowns already used in Supplementary Tab. 5–6, and present these results in Methods - AI Models and Explanations.

We also agree with the reviewer that post-hoc explanations (GradCAM/CBIR/LLM) depend on model predictions and learned representations, so model errors can reduce explanation reliability. For this reason, the paper does not treat explanations as inherently valid; instead, we explicitly measure explanation quality using independent dermatologist ratings and use these ratings to stratify downstream analyses (e.g., Extended Data Fig. 2). In the Methods (Evaluation Metrics), we define two separate constructs, correctness ("how accurate the explanation is based on the skin image") and informativeness ("how much this explanation can support diagnosis"), each rated by three dermatologists, and we define high vs. low XAI quality using threshold rules (high if correctness ≥ 3 and informativeness ≥ 3 and low if otherwise) (Methods: Evaluation Metrics; Extended Data Fig. 2). We also provide representative examples where both the ground-truth label, AI-predicted label, and expert scores are shown (Supplementary Fig. 7–12), which is the basis for the high/low-quality grouping used throughout the analysis (Extended Data Fig. 2; Supplementary Fig. 7–12).

We share R2's concerns that the explanations are not reliable when the overall model performance is moderate. Therefore, we investigated the explanation quality more closely, especially when AI model prediction is correct (we won't expect the explanation to be reliable when prediction is wrong). Assuringly, our summary statistics show that expert-rated explanation quality is substantially higher in AI-correct cases than AI-incorrect cases. For Study 1 (general public), correctness is higher when AI is correct (3.63±1.05) versus incorrect (2.75±1.13; $t=5.28$, $p<0.001$, Cohen's $d=0.83$, two-sided unpaired t-test), and informativeness is higher when AI is correct (3.25±1.05) versus incorrect (2.42±0.96; $t=5.09$, $p<0.001$, Cohen's $d=0.80$). For Study 2 (PCPs), correctness is higher when AI is correct (3.32±1.43) versus incorrect (2.19±0.95; $t=11.32$, $p<0.001$, Cohen's $d=0.96$), and informativeness is higher when AI is correct (3.14±1.44) versus incorrect (1.91±0.84; $t=12.64$, $p<0.001$, Cohen's $d=1.07$). These results support the intended interpretation of our experiments: model errors do degrade explanations, but when the model is correct, the explanations are rated as reasonably correct and informative by experts, enabling meaningful comparisons of how different XAI strategies affect users under AI-correct versus AI-incorrect conditions (e.g., Fig. 2d–f and Extended Data Fig. 2 for the general public; the same analyses for PCPs).

Regarding the key clarification about labels: the dermatologists rated explanation correctness and informativeness with access to the ground truth, and their ratings are anchored to the true label, not the predicted label. We clarified this in the Methods (Evaluation Metrics). Also, as we described in Methods (Model Training) and Method (AI Models and Explanations), all post-hoc AI explanations were based on the same trained base model. We further emphasized that part in the paper to make it more clear and explicit. The differences between human evaluation and post-hoc explanation resolve the ambiguity raised by examples such as Supplementary Fig. 7: when the AI predicts melanoma but the image is a nevus, a heatmap can still be partially informative if it highlights clinically relevant regions, but it should receive low correctness if it

supports the wrong diagnosis—exactly the motivation for separating correctness and informativeness and stratifying analyses by high vs. low explanation quality (Methods: Evaluation Metrics; Extended Data Fig. 2; Supplementary Fig. 7–12).

Finally, we acknowledge that instead of directly utilizing all model predictions, we intentionally sampled test images to achieve a controlled mixture of AI-correct and AI-incorrect cases, which is described in the Methods (AI Models and Explanations). We indeed agree with R2 that this approach does not reflect realistic model behavior. Our purpose is to mimic an envisioned future where AI models have reasonable performance (which we currently do not have yet). We expanded our limitation paragraph to discuss both the implicit impact of model performance (and therefore, representations) on explanation quality. We also noted that behavioral effects should be interpreted as conditional on AI correctness under our controlled sampling, rather than as a direct estimate of population-level impact in real-world deployment.

R2-Comment2 (R2-C2) Performance of Study 2 (differential diagnosis task)

The authors report a top-1 accuracy of 11.5% and a top-3 accuracy of 16.1% for PCPs without AI support. In combination with the similarly weak accuracies of AI models, this sparks doubt, whether the task at hand is solvable based on a single image, with or without AI. Reporting misleading results for a solvable task would argue for a stronger case of the “double-edged” nature of XAI.

We appreciate the reviewer’s concern that the PCP baseline performance (11.5%±1.4% top-1; 16.1%±1.8% top-3) together with moderate AI model performance could raise doubt about whether single-image differential diagnosis is feasible. We would like to clarify that these numbers are consistent with prior work using the same clinical-image, single-image setup. In our Results (“PCPs Reliably Leverage Accurate AI Guidance…”), we now explicitly note that the PCP baseline accuracy is “aligned with the prior work” [1] when reporting the 11.5%/16.1% results: PCPs achieved ~14% top-1 and ~19% top-3 on a similar single clinical-image differential diagnosis task. This supports that the task is challenging but not ill-posed, and our observed baseline is within the expected range for this setting.

Importantly, our study is designed to reflect common real-world workflows. In the revised version, we now added clearer description in the Results (Study Design) section (and Fig. 1a), both studies use clinical images to resemble common outside-clinic scenarios: (i) a lay person may take a photo and consult online tools, and (ii) clinicians may provide differential diagnoses from a clinical image in asynchronous communication (e.g., telehealth consulting or EHR messaging). (Results: Study Design; Fig. 1a). In these realistic settings, our results show that AI suggestions can meaningfully improve clinician performance despite the task difficulty: with AI suggestions, PCP top-1 accuracy increases by +21.5% (β=0.258, 95%CI=[0.170, 0.260], $p<0.001$; Fig. 3a; Supplementary Tab. 16) and top-3 accuracy increases by +43.5% (β=0.450, 95%CI=[0.352, 0.548], $p<0.001$; Extended Data Fig. 6a; Supplementary Tab. 16). These

improvements indicate that the task is sufficiently informative for us to explore systematic benefits and harms of different XAI strategies, which is the central goal of our work.

Finally, we agree that single-image diagnosis is not ideal and is not generally recommended as a stand-alone medical workflow (e.g., clinicians usually have access to richer information such as patients' prior medical records and family history), but it is increasingly common for the general public to seek image-based advice from general-purpose AI systems, especially with the recent booming of generative AI tools (e.g., ChatGPT Health, Claude for Healthcare). Our manuscript motivates this behavioral reality in the Study Design section by noting that lay users resort to online tools for single images [2, 3]. We now have emphasized that part in the Discussion to explicitly acknowledge this as a limitation of the single-image-based Q&A setting and to cite recent evidence on public use of LLMs for image-based health questions [2, 3].

[1] Groh, M., Badri, O., Daneshjou, R., Koochek, A., Harris, C., Soenksen, L. R., ... & Picard, R. (2024). Deep learning-aided decision support for diagnosis of skin disease across skin tones. Nature Medicine, 30(2), 573-583.
[2] Jain, A., Way, D., Gupta, V., Gao, Y., de Oliveira Marinho, G., Hartford, J., ... & Liu, Y. (2021). Development and assessment of an artificial intelligence–based tool for skin condition diagnosis by primary care physicians and nurse practitioners in teledermatology practices. JAMA network open, 4(4), e217249-e217249.
[3] Lee, P., Bubeck, S., & Petro, J. (2023). Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. New England Journal of Medicine, 388(13), 1233-1239.

R2-Comment3 (R2-C3) Carry-Over Effect

Each participant reviews the same 12 cases twice. All subsequent analysis stems from the comparison of the diagnostic accuracy per-image between the two runs. Carry-over effects apply. Other reviewers have pointed that out in-depth.

We appreciate the reviewer's concern that evaluating the same 12 cases twice could introduce carry-over effects. We designed our study design to take care of this potential confounder. Specifically, the study explicitly separates when AI is provided across two paradigms (Methods; Fig. 1b): in the AI-First paradigm, participants receive AI support in Round 1 (their first exposure to each image together with AI suggestions), whereas in the Human-First paradigm, participants receive AI support in Round 2 (after an initial unaided judgment). Note that in the AI-First paradigm, AI suggestions were removed in Round 2 (so that human participants can review & reflect for the second time independently). Our existing analysis focuses on AI-First Round 2 vs Human-First Round 2 (in Results Fig. 5). This is to control the number of decision rounds that humans could make (i.e., both groups would make two rounds of decisions).

As suggested by R2, to directly test whether carry-over meaningfully biases the AI-assisted outcomes, we now conducted an additional analysis comparing AI-assisted performance in AI-First Round 1 versus AI-assisted performance in Human-First Round 2, two settings where participants have AI support, but AI is introduced either during or after prior exposure to the image. We found no significant difference between these two AI-assisted conditions (the

general public: β=0.002, 95%CI=[-0.018, 0.021], p=0.873; PCPs: β=-0.013, 95%CI=[-0.116, 0.090], p=0.803, Supplementary Tab. 38-39). This suggests that the main conclusions about explanation strategies are not driven by carry-over from repeating the same cases, but reflect the impact of AI support and explanation design under the same two-round decision process. We have added these results in Supplementary Materials and mentioned this briefly in the Results section (see Results: Putting AI Before Human Decisions Amplifies Deference Across Expertise Levels).

R2-Comment4 (R2-C4) Small Experiment Trials

Small n: Just 12 cases per participant. Other reviewers have pointed that out in-depth.

We acknowledge the concern about the per-participant image count. In our protocol, each participant completes 12 cases to balance statistical power with participant burden and data quality in an online, attention-sensitive setting (Study Design; Fig. 1b). In Study 1, the general public evaluates 12 clinical images sampled from a larger pool (Methods: Study 1 design), and in Study 2, PCPs also evaluate 12 cases with a structured composition across the four main skin conditions plus “other” conditions (Methods: Study 2 design). This choice is consistent with prior study from Groh et al. [1] which had 10 cases per participant.This is similar to other studies that explore a similar image-based setup beyond dermatology (e.g., 8 cases in [2], 15 cases in [3]). As mentioned in our response in the last round, for an online study like ours (our current studies took 10-15 minutes to finish), using substantially more images may risk reduced engagement and lower data quality. We have now added this as an explicit limitation in the Discussion, and further clarified in the revision that increasing cases per participant is a natural next step for future work to enable more granular subgroup analyses.

[1] Groh, M., Badri, O., Daneshjou, R., Koochek, A., Harris, C., Soenksen, L. R., ... & Picard, R. (2024). Deep learning-aided decision support for diagnosis of skin disease across skin tones. Nature Medicine, 30(2), 573-583.
[2] Gaube, Susanne, et al. "Do as AI say: susceptibility in deployment of clinical decision-aids." NPJ digital medicine 4.1 (2021): 31.
[3] Chanda, Tirtha, et al. "Dermatologist-like explainable AI enhances trust and confidence in diagnosing melanoma." Nature Communications 15.1 (2024): 524.

R2-Comment5 (R2-C5) Clarification of HAI Paradigm Used in Each Analysis

Do the results shown in Fig. 2a) and 3a) only include results from the Human-First subgroup? Mentioning Rd.1 as ‘initial’ in the caption of Fig. 2 points in that direction. The fact could be pointed out more clearly.

The results in Fig. 2a and Fig. 3a report outcomes from the Human-First paradigm only, where Round 1 is the initial (unaided) diagnosis and Round 2 is the updated diagnosis after showing AI support (Study Design; Fig. 1b). The AI-First paradigm is analyzed only in comparisons that explicitly involve the human–AI decision paradigm, including Fig. 5, Extended Data Fig. 4m, and

Extended Data Fig. 6m. We appreciate R2's suggestion and agree this could be clearer. In this revision, we modified the main text and the captions of Fig. 2 and Fig. 3 to explicitly state that these panels present Human-First only, and that AI-First results appear only in the paradigm-focused analyses (Fig. 5; Extended Data Fig. 4m; Extended Data Fig. 6m).

R2-Comment6 (R2-C6) Expired Figshare Link

The figshare link had expired.

Thanks for pointing this out. We have updated the link in the revised version.

Response to Reviewer #3

R3-Comment1 (R3-C1) Model Performance and Explanation

While the use of CDANN is methodologically sound, the manuscript (particularly the Abstract and Discussion) must carefully distinguish between benefits derived from the fairness-constrained model predictions (which affect all conditions, including Basic) versus benefits derived specifically from the XAI modalities. The reduction in diagnostic disparity appears to be primarily driven by the underlying CDANN training rather than the explanations themselves. Please ensure the text reflects this nuance to avoid overstating the independent role of XAI in equity.

We agree with the reviewer's point and have now revised the Abstract and Discussion to clearly separate (i) benefits that arise from the fairness-constrained predictions produced by CDANN (which are applied in all study conditions, including Basic) from (ii) any additional effects attributable to the XAI modalities themselves. In our current Results ("Advanced AI improves the general public's performance"), the reduction in skin-tone disparity is reported as a benefit associated with AI assistance overall, rather than with any specific explanation type: "With the help of AI model humans achieved a more balanced diagnosis performance across patient skin tones (Round 1: β=0.033, 95%CI=[0.009, 0.057], p=0.007; Round 2: β=0.017, 95%CI=[-0.007, 0.041], p=0.166; Fig. 2c; Supplementary Tab. 11)."

Consistent with the reviewer's interpretation, we did not observe interactive effects of XAI methods in fairness gains ($p>0.05$ for all conditions in both general public and PCPs, Supplementary Tab. 11-22), indicating that the fairness improvement is primarily driven by the underlying CDANN-trained model outputs rather than by explanation methods. We have now added clarification in the Abstract, Results, and Discussion to avoid implying that XAI independently improves equity, and instead state that XAI mainly modulates how users interpret and act on AI suggestions, while the equity-related improvement in the AI-supported conditions is attributable to the fairness-constrained model predictions.

R3-Comment2 (R3-C2) Effect of the Nature of XAI

Regarding the selection of XAI methods, the rationale for selecting GradCAM and CBIR as representative baselines is accepted. However, characterizing Concept-Based methods (e.g., Concept Bottlenecks) merely as 'novelty' is slightly reductive in the context of dermatology. Concept-based XAI represents the computational analog to standard clinical training. By contrast, the 'Multimodal LLM' arm in your study represents 'Unstructured Narrative' explanations. Please refine the Discussion to explicitly distinguish between Structured Concept-Based XAI (which aligns with standard dermatological criteria like ABCD) and the Unstructured Narrative Explanations (LLM) used in this study. It is important to clarify that the observed "Double-Edged Sword" effect may be driven specifically by the persuasive rhetoric of the LLM's narrative, which is a distinct mechanism from the structured concept scoring often used in clinical training.

We thank R3 for this important nuance and agree that “concept-based” explanations are closely aligned with how clinicians are trained to reason in dermatology (e.g., structured criteria such as ABCD). Our intent in the Discussion was not to dismiss concept-based XAI as mere novelty, but to note that it represents a distinct design family from commonly deployed post-hoc visual retrieval/saliency baselines (GradCAM/CBIR) and from free-form generative explanations. We revised the Discussion to make this distinction explicit and to clarify the mechanism of the observed “double-edged sword” effect in our setting.

In our study, GradCAM and CBIR serve as representative post-hoc visual baselines, and the multimodal LLM produces free-form narrative explanations. We agree that structured concept-based XAI (e.g., concept bottlenecks or explicit ABCD-style concept scoring) is better viewed as a computational analog of clinical training rather than simply novel. We have now revised the Discussion to distinguish (i) structured, concept-based explanations that map to standardized dermatologic criteria from (ii) unstructured narrative explanations (our LLM arm). We also clarified that the “double-edged” effect we observe is most consistent with mechanisms tied to how narrative explanations can increase perceived plausibility and trust, rather than implying the same mechanism would necessarily apply to structured concept scoring.

R3-Comment3 (R3-C3) Revision of Expression

The phrase "delta of detla" contains a typo and is slightly colloquial. Please revise to "difference in differences" or "differential impact." Duplicate word: "explanations explanations."

We appreciate R3’s great suggestions and have improved our wording accordingly.