nature medicine

Article

https://doi.org/10.1038/s41591-026-04553-w

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

Received: 6 June 2025

Accepted: 29 June 2026

Published online: 04 August 2026

Check for updates

A list of authors and their affiliations appears at the end of the paper

Artificial intelligence (AI) is increasingly permeating healthcare, from serving as a physician assistant to powering consumer applications. The opacity of AI algorithms makes the ability of humans to interact with AI algorithms challenging. To overcome this limitation, explainable AI (XAI) provides insight into AI decision-making, but evidence suggests that XAI can paradoxically induce bias in the human decision-making process. Here we present results from two large-scale experiments, involving 623 lay people and 153 primary care physicians (PCPs), respectively, in which a fairness-based AI model for dermatological diagnoses and different XAI-based explanations were combined to examine how XAI assistance, particularly multimodal large language models (LLMs), influences diagnostic performance. With fairness-constrained model training, assistance from an AI model that achieved balanced performance across skin tones improved final diagnostic accuracy and reduced skin-tone-related performance disparities among both lay people and PCPs. In this setting, LLM explanations yielded divergent effects: lay users showed higher automation bias—accuracy was boosted when the diagnoses provided by the AI model were correct but was reduced when the model erred—whereas experienced PCPs remained resilient, benefiting irrespective of the AI model's accuracy. In addition, presenting the AI model's diagnosis before human decision-making may lead to stronger anchoring bias. These findings highlight XAI's varying impacts based on human expertise and the timing of when the AI-based prediction is provided, underscoring the concept that LLMs can act as a 'double-edged sword' in medical AI and informing future human–AI collaborative system design.

AI has been intensively investigated as a tool to enhance clinical decision-making. Dermatology is one area with several developed and FDA-approved tools such as Nevisense[1] and DermaSensor[2], as skin conditions are primarily diagnosed through image assessment. In light of the national shortage of dermatologists[3,4], effective AI assistance could improve early detection and reduce unnecessary clinical visits. The development of AI-powered interfaces has also been proposed to assist the general public in making informed healthcare decisions (for example, self-diagnosis of skin diseases with Google Lens[5]).

XAI has been used in health generally to target AI usability and adoption[6,7]. In dermatology, physicians have clear features to look for, such as the ABCDs (asymmetry, irregular borders, multiple colors, diameter greater than a pencil eraser) of melanoma. Similarly, gradient-weighted class activation mapping (GradCAM)[8] and content-based image retrieval (CBIR)[9,10], the top two most commonly used XAI techniques in dermatology[11], have been used to highlight relevant image regions and retrieve similar cases, respectively. In addition, with the recent surge of generative AI[11], multimodal LLMs operating

e-mail: xx2489@cumc.columbia.edu; mghassem@mit.edu

https://doi.org/10.1038/s41591-026-04553-w

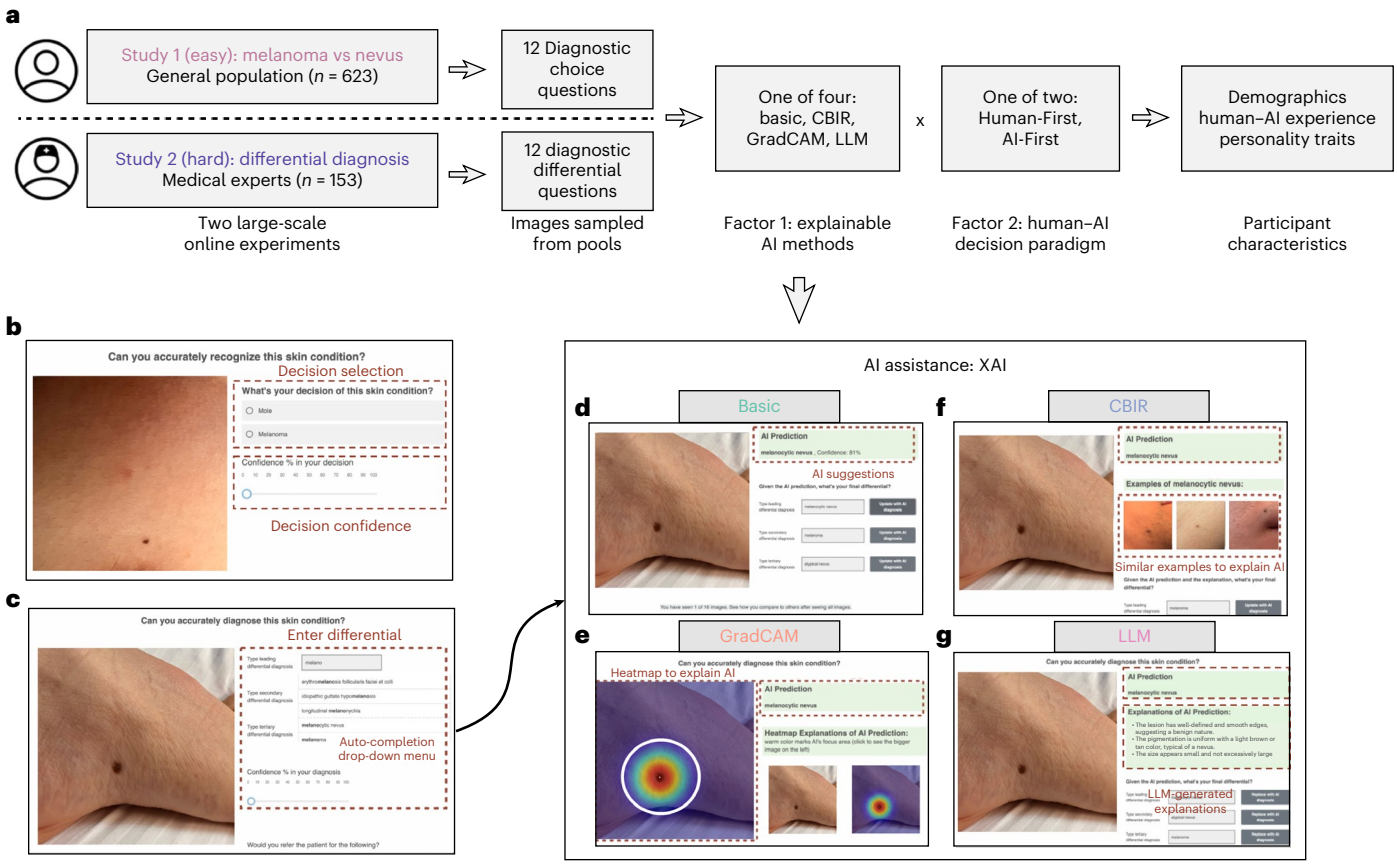


Fig. 1 | Study design and interface components of human–AI collaborative diagnostic system. a, Overview of the experimental workflow comprising two studies: study 1 recruited general public participants for melanoma versus nevus classification tasks; study 2 engaged medical experts for open-set differential diagnosis. Twelve diagnostic images were randomly drawn from a high-quality image database with an equal skin tone split. Participants completed either Human-First or AI-First diagnostic sequences followed by demographics, human–AI collaboration experience and personality assessment. **b**, Interface for the binary classification task showing a clinical image for melanoma versus nevus discrimination. **c**, Interface for the differential diagnosis task with top-3 free-text entry. To assist user input, the text box has a string-matching-based auto-completion function that covers a comprehensive set of 445 skin diseases. **d**–**g**, Four XAI assistance examples with basic AI providing straightforward predictions with model confidence (**d**), GradCAM highlighting relevant image regions (**e**), CBIR showing similar reference cases (**f**) and LLM providing natural language explanations (**g**). Skin images are not clinical images but are from the authors as placeholders for illustration. GradCAM and CBIR are synthetic and do not represent the real model performance.

over textual and visual modalities[12,13] have also been used to analyze dermatological images and explain AI decisions[14,15].

Unfortunately, previous work has shown that XAI can increase subject overreliance on AI in decision-making[16,17], and human–AI collaborative medical decisions do not always surpass those made by humans or AI alone[18,19]. To date, there is no consensus about whether medical experience improves human–AI collaborative diagnosis. Although some research indicates that medical knowledge is needed for better AI resilience and clinical decision-making[20], there is also research showing that it makes minimal difference[10], and, in dermatology, AI may even mislead humans and lead to worse outcomes in diagnosis[21,22]. As dermatological AI tools expand, it is crucial to understand how different XAI methods, especially with the vast spread of LLMs, affect skin disease diagnostic accuracy across both the general public and medical experts[23–25].

In the present study, we designed two large-scale experiments to systematically investigate how different XAI methods and human–AI decision paradigms impact diagnostic performance across expertise levels: the general public (n = 623) and PCPs (n = 153). We chose clinical-image-based skin condition diagnosis as a plausible real-world scenario (that is, ecological validity) given the influx of patient-facing and physician-facing diagnostic models in this space[22,26]. We investigated the overall effectiveness of XAI assistance in improving dermatological diagnostic accuracy[27,28], reducing the disparities across skin tones[29] and influencing accuracy–confidence calibration (that is, accuracy and confidence are consistent). We compared both correct and incorrect multimodal-LLM-based explanations to traditional XAI approaches in both scenarios. We also examined how individual differences in AI deference (that is, the propensity to follow AI regardless of accuracy, sometimes referred to as AI susceptibility[30]) impact diagnostic performance. Finally, we explored the impact of the human–AI decision paradigm (that is, the order of human or AI making decisions) on diagnostic outcomes[31,32], providing insights into optimal implementation strategies for clinical settings.

Results

Study design

We designed two complementary large-scale experiments to evaluate human–AI collaborative diagnostic performance across expertise levels (Fig. 1a). Study 1 engaged the general public (n = 623) in a binary classification task to distinguish melanoma from nevus. Study 2 engaged PCPs (n = 153) in a complex open-ended differential diagnosis task, focusing on four skin conditions previously identified as having potential diagnostic disparities across skin tones[22,33]: atopic dermatitis, pityriasis rosea, Lyme disease and cutaneous T cell lymphoma (CTCL). Studies 1 and 2 are not directly comparable as they emphasize different tasks. To measure the impact of medical training within the same task, for study 2, we recruited another cohort of medical students (n = 320) for comparison.

https://doi.org/10.1038/s41591-026-04553-w

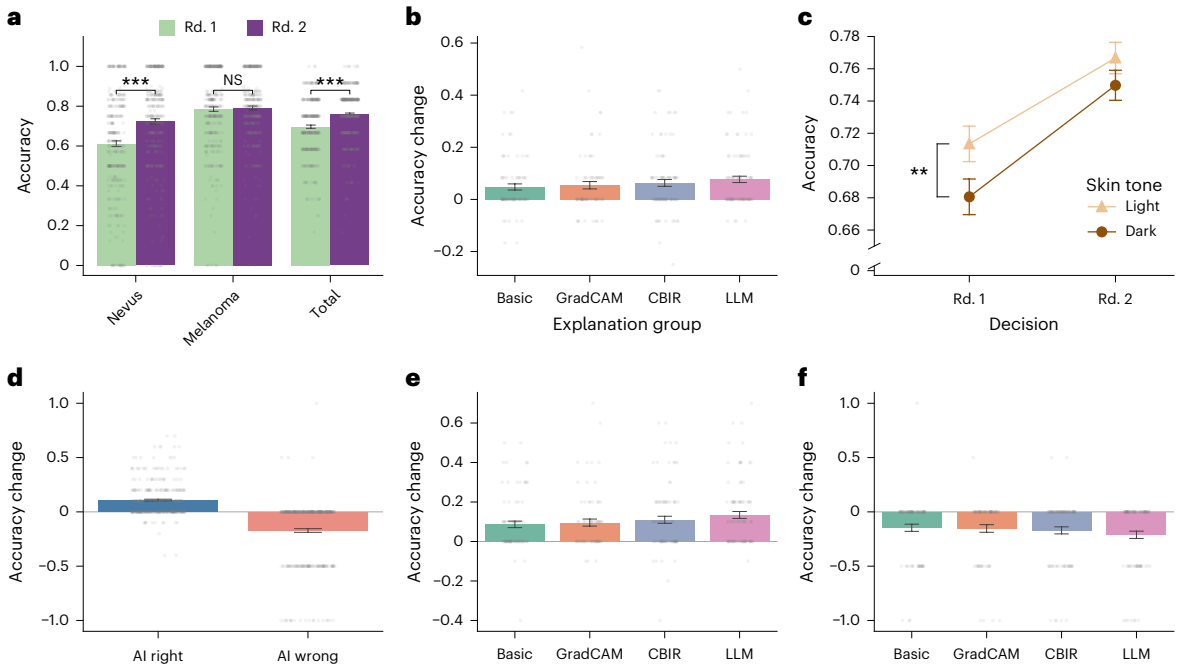


Fig. 2 | Overall diagnostic performance with AI assistance for the general public. a, Accuracy of the general public ($n = 335$) in detecting nevus and melanoma and the overall diagnosis in two rounds of testing: Rd. 1 tested initial human accuracy without AI assistance; Rd. 2 tested AI-assisted accuracy. **b**, Diagnostic accuracy changes between Rd. 2 and Rd. 1 across different AI explanations (Basic ($n = 82$), GradCAM ($n = 72$), CBIR ($n = 91$) and LLM ($n = 90$), same in **e** and **f**) in the Total condition of **a**. **c**, Diagnostic accuracy when patients are light-skinned (Fitzpatrick labels 1–4 on a scale of 6) or dark-skinned (Fitzpatrick labels 5–6) across two rounds in the Total condition of **a**. **d**, Overall accuracy changes ($n = 335$) between Rd. 1 and Rd. 2 under AI-right and AI-wrong conditions. **e**,**f**, Diagnostic accuracy changes between Rd. 1 and Rd. 2 of different AI explanations under AI-right (**e**) and AI-wrong (**f**) predictions. Each gray transparent scatter represents one measured sample. All presented bars or data points of a plot in this figure indicate the mean values of metrics measured. All error bars indicate the standard error of the mean. *P* values were calculated with post hoc marginal comparison after an initial analysis with a linear mixed model. $^{***}P < 0.001$, $^{**}P < 0.01$, $^{*}P < 0.05$. NS, not significant; Rd., round.

We employed a randomized between-subjects factorial design (4 × 2) across both studies. Participants were assigned to one of four AI assistance methods—basic (prediction and confidence), GradCAM (heatmap), CBIR (visual similarity) or multimodal LLM (textual explanation)—and one of two decision paradigms: Human-First (users make a decision first before reviewing AI suggestions) and AI-First (users review both images and AI suggestions before making the final decision). All participants evaluated 12 clinical images balanced by skin tone and pathology, utilizing outputs from fairness-constrained deep learning models.

Our fairness-constrained models (final architectures and hyperparameters reported in Methods) achieved strong overall performance with substantially reduced disparities across skin tones. For study 1, the binary classification model achieved a weighted area under the receiver operating characteristic (AUROC) of 0.930 (0.933/0.898 light/dark skin, $\Delta = 0.035$) and weighted balanced accuracy of 0.850 (0.852/0.831, $\Delta = 0.021$), narrowing the empirical risk minimization (ERM) baseline's skin tone gap (balanced accuracy 0.845, $\Delta = 0.091$) by 76.9%. For study 2, the primary five-class model achieved AUROC of 0.772 (0.782/0.691, $\Delta = 0.091$), weighted AUROC of 0.753 (0.725/0.693, $\Delta = 0.192$) and balanced accuracy of 0.478 on the four main diseases (0.487/0.431, $\Delta = 0.056$), outperforming ERM (0.457, $\Delta = 0.144$). The secondary 30-class model achieved AUROC of 0.728 (0.714/0.641, $\Delta = 0.073$), weighted AUROC of 0.752 (0.798/0.744, $\Delta = 0.053$) and balanced accuracy of 0.141 (0.157/0.092, $\Delta = 0.064$). Combined, the primary and secondary models achieved overall weighted accuracy of 0.197 and AUROC of 0.755. Per-disease performance is in Supplementary Tables 5 and 6.

In the study, images were intentionally sampled to have an overall AI accuracy of 83.3% (always 10 correct and two incorrect predictions) in study 1 and 79.2% (on average, 9.5 correct and 2.5 incorrect) in study 2 (see Methods for full details on model training, dataset curation and experimental protocol).

State-of-the-art AI improves the general public's performance due to AI deference and LLM explanations amplify such deference

Advanced AI improves the general public's performance. We first measured the general public's performance (study 1) without and with AI assistance. We found that AI improved average accuracy of the nevus versus melanoma detection task from 69.7 ± 0.8% to 75.8% ± 0.7% (effect of AI assistance: $\beta = 0.061$, 95% confidence interval (CI): 0.049–0.074, $P < 0.001$, linear mixed model on accuracy, with AI assistance, XAI methods and their interaction as the main factor, controlling gender, age, race, skin disease experience and the covariate of self-reported human–AI collaboration experience; see Supplementary Table 7 for details). Other study 1 statistical models below control the same set of confounders (unless noted differently), and most of the improvement came from nevus classification ($\beta = 0.111$, 95% CI: 0.091–0.132, $P < 0.001$; Fig. 2a and Supplementary Table 8), and the largest improvement in XAI came from multimodal LLM (see next section). Participants also had a modest increase in diagnosis confidence by 1.5% with AI assistance ($\beta = 0.018$, 95% CI: 0.010–0.019, $P < 0.001$; Extended Data Fig. 1a and Supplementary Table 10). With the help of AI, model humans achieved a more balanced diagnosis performance across patient skin tones (round 1: $\beta = 0.033$, 95% CI: 0.009–0.057, $P = 0.007$; round 2: $\beta = 0.017$, 95% CI: −0.007 to 0.041, $P = 0.166$; = 46.9%; Fig. 2c and Supplementary Table 11). As no interactive effects across different XAI methods, skin tones and decision rounds were found ($P > 0.05$ for all; Supplementary Table 11), the reduced diagnostic disparities were mainly contributed by the fairness-constrained training algorithm (conditional domain adversarial neural network (CDANN); see Model training section). These findings demonstrate that collaboration with well-trained AI can mitigate diagnostic biases while improving overall accuracy.

Performance improvement stems from AI deference and LLM-based explanations amplify such deference. We investigated the impact of the four XAI methods on diagnostic accuracy. The LLM explanations provided an improvement of +7.7% ($\beta = 0.077$, 95% CI: 0.053–0.101, $P < 0.001$), followed by CBIR (+6.3%, $\beta = 0.063$, 95% CI: 0.039–0.087, $P < 0.001$), GradCAM (+5.5%, $\beta = 0.054$, 95% CI: 0.028–0.081, $P < 0.001$) and the basic method (+4.8%, $\beta = 0.048$, 95% CI: 0.023–0.073, $P < 0.001$), as shown in Fig. 2b and Supplementary Table 7. However, the general public's diagnostic accuracy improved when AI provided correct predictions, whereas incorrect predictions reduced the performance significantly ($\beta = -0.233$, 95% CI: −0.307 to −0.158, $P < 0.001$; Fig. 2d and Supplementary Table 12). Correct LLM advice (+13.4%) enhanced performance more than other AI explanations (basic method +8.6%, GradCAM +9.5%, CBIR +10.9%; Fig. 2e), and incorrect LLM advice decreased performance most (−21.1%, basic −14.6%, GradCAM −15.3% and CBIR −17.0%; Fig. 2f). These findings indicate that LLM explanations amplify the general public's tendency to follow AI guidance, regardless of AI accuracy. Moreover, compared to basic explanation, LLM led to significantly more reduction of performance when AI becomes inaccurate (that is, difference in differences, $\beta = -0.048$, 95% CI: −0.093 to −0.003, $P = 0.035$; Supplementary Table 13).

Misplaced trust in LLM explanations for the general public. When AI predictions were correct, participants trusted LLM explanations more than other methods (Fig. 2e), and this was more noticeable when explanations were of low quality (post hoc pairwise estimated marginal means (EMMs) comparisons, LLM over GradCAM: $\beta = 0.117$, 95% CI: 0.041–0.192, $P = 0.002$; LLM over CBIR: $\beta = 0.092$, 95% CI: 0.021–0.163, $P = 0.011$; Extended Data Fig. 2c and Supplementary Table 15). When AI predictions were incorrect, LLM explanations negatively impacted the alignment between participants' confidence and accuracy ($z = -3.788$, $P < 0.001$, two-sided Fisher's r-to-z test to compare correlation difference; Extended Data Fig. 3c). These results further suggest that people struggle to assess LLM explanation reliability and can be easily misled by LLMs.

Study 1 results indicate that LLMs are a 'double-edged sword' in skin disease diagnosis for the general public with amplified AI deference. When AI was correct, LLM explanations boosted diagnosis performance, even when the quality of the explanation was low. However, when AI predictions were incorrect, the general public was misled by seemingly plausible reasons generated by LLM, whose explanations frequently referenced ambiguous dermatologic criteria even when these features were only partially present or visually unclear.

PCPs reliably leverage accurate AI guidance while resisting errors with LLM-based AI explanation

Basic AI improves performance of PCPs. We conducted a similar analysis of the PCP participants in study 2. This task was more challenging and required detailed dermatological knowledge. We found 11.5 ± 1.4% top-1 accuracy and 16.1 ± 1.8% top-3 accuracy in differential diagnoses of PCP without AI, which is aligned with previous work[22]. The performance was significantly improved with AI suggestions, with a +21.5% in top-1 accuracy ($\beta = 0.258$, 95% CI: 0.170–0.260, $P < 0.001$; Fig. 3a; linear mixed model on accuracy with AI assistance, XAI methods and their interaction as the main factor, controlling gender, age, race and medical expertise in skin, year of experience and personality traits; see Supplementary Table 16 for details). Other study 2 statistical models below control the same set of confounders (unless noted differently) and a +43.5% in top-3 accuracy ($\beta = 0.450$, 95% CI: 0.352–0.548, $P < 0.001$; Extended Data Fig. 4a and Supplementary Table 16). In contrast to the general public in study 1, for PCPs, basic AI assistance helped the most (top-1 accuracy $\beta = 0.258$, 95% CI:0.168–0.349, $P < 0.001$; Fig. 3b and Supplementary Table 16). Interestingly, we observed significant confidence increase only when PCPs were assisted by GradCAM ($\beta = 0.028$, 95% CI: 0.005–0.051, $P = 0.019$) and LLM explanations ($\beta = 0.035$, 95% CI: 0.012–0.058, $P = 0.003$) (Extended Data Fig. 1b and Supplementary Table 21). Improvements were significant in all four major diseases (+19.9–25.0%, all $P < 0.001$; Fig. 3a and Supplementary Tables 17–20). PCPs had disparate performance across skin tones as 4.6% ($\beta = 0.046$, 95% CI: 0.013–0.078, $P = 0.069$; Fig. 3c), which was reduced to 2.9% ($\beta = 0.029$, 95% CI: −0.018 to 0.076, $P = 0.248$; Supplementary Table 22) after AI assistance. Similar to the general public, no interactive effects of different XAI methods were found ($P > 0.05$ for all conditions; Supplementary Table 22), showing that the reduced disparities still resulted from CDANN.

PCPs are resilient to AI deference when AI is wrong. In contrast to the general public, incorrect AI predictions had minimal impact on the final decisions of PCPs across all XAI methods (Fig. 3f and Extended Data Fig. 2d–f; $\beta = 0$–0.021, $P = 0.328$–1.000). This suggests that PCPs relied on their own expertise and training rather than erroneous AI guidance. To control the effect of task difficulty between study 1 and study 2, we further compared the results of PCPs against the results of medical students on the same task. Extended Data Fig. 5 shows that medical students relied on AI more than PCPs, regardless of AI correctness. We term participants who answered correctly only when AI was correct—and, therefore, answered incorrectly when AI was wrong—as 'deferential participants'. We observed that the proportion of deferential participants was higher among medical students than PCPs (linear mixed model on the proportion of deferential participants, with medical role and medical expertise in skin as the main factor, controlling other confounders: main effect of medical role: $\beta = 0.067$, 95% CI: 0.004–0.130, $P = 0.037$; main effect of skin expertise: $\beta = 0.073$, 95% CI: 0.022–0.124, $P = 0.005$; Supplementary Table 25). These results indicate that higher expertise levels are associated with more careful AI adoption, which is supported by previous work[10].

LLM explanations do not aid in accuracy but in confidence calibration. Interestingly, for PCPs, LLM explanations were the least helpful method (+17.7%) and were 8.1% lower than the best improvement from the basic explanations ($\beta = -0.081$, 95% CI: −0.207 to 0.044, $P = 0.268$; Fig. 3b and Supplementary Table 16). This finding was consistent across explanation quality (Extended Data Fig. 2e,g and Supplementary Table 27) and top-3 accuracy (Extended Data Fig. 4c–f). These are opposite to the results of LLM's best improvement for the general public in study 1. Medical student data confirmed that the task was not biased toward certain explanations (Extended Data Fig. 5c), implying that expertise drove interactions: the general public overrelies on LLMs, whereas PCPs are more resilient to incorrect LLM suggestions.

However, LLM explanations did help improve the alignment between PCP participants' confidence and accuracy (correlation $r = 0.494$, $P = 0.010$; Extended Data Fig. 3d; similar findings in top-3 performance; Extended Data Fig. 6d) over No AI (correlation $r = 0.084$, $P = 0.415$, two-sided Fisher's r-to-z test comparing the two correlations: $z = 2.368$, $P = 0.018$; Extended Data Fig. 3d). This calibration benefit held even with incorrect AI predictions, where non-LLM explanations impaired alignment (Fisher's r-to-z test $z = 2.572$, $P = 0.010$; Extended Data Fig. 3f; similar in top-3 performance, $P = 0.059$; Extended Data Fig. 6f). PCP caution toward LLMs likely enforces cognitive engagement, thus enhancing diagnostic accuracy–confidence calibration regardless of correctness[34,35].

Overall, in contrast to the general public in study 1, PCPs maintained their performance under incorrect AI predictions across all XAI methods, including LLM.

Higher AI deference correlates with lower initial performance

We inspected the relationship between participants' deference toward AI suggestions and initial performance[30]. Among participants' final decisions that are correct (ranging from zero to 12, 12 images total), we visualize the number of images with AI suggestions that are correct (up to 10, shown in blue) and incorrect (up to two, shown in red) for each

https://doi.org/10.1038/s41591-026-04553-w

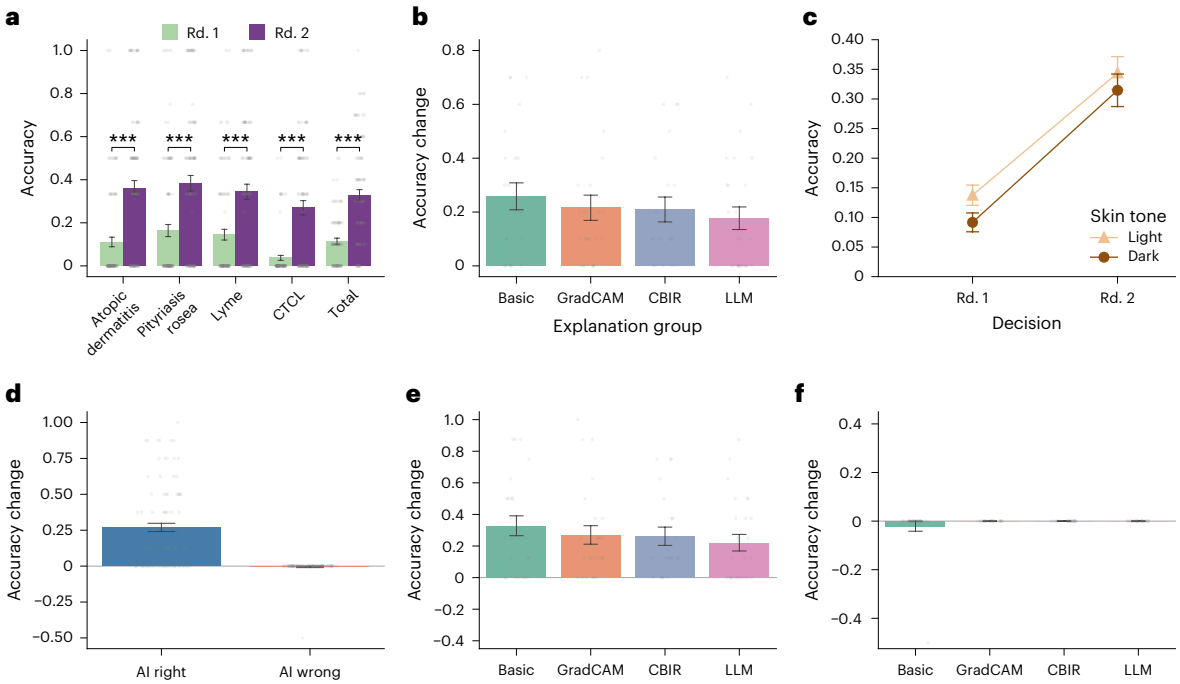


Fig. 3 | Overall diagnostic performance with AI assistance for PCPs. This figure shows similar analysis results as in Fig. 2 using the Human-Frst paradigm. **a**, Accuracy of PCPs ($n = 96$) in identifying atopic dermatitis, pityriasis rosea, Lyme disease and CTCL as well as the overall performance on four main diseases, with results presented for two rounds of testing. **b**, Diagnostic accuracy changes between Rd. 1 and Rd. 2 across different AI explanations (Basic ($n = 24$), GradCAM ($n = 25$), CBIR ($n = 21$) and LLM ($n = 26$), same as **e** and **f**). **c**, Diagnostic accuracy when patients are light-skinned (Fitzpatrick labels 1–4 on a scale of 6) or dark-skinned (Fitzpatrick labels 5–6) across the two rounds of testing. **d**, Overall accuracy ($n = 96$) changes between Rd. 1 and Rd. 2 under AI-right and AI-wrong conditions. **e**,**f**, Diagnostic accuracy changes between Rd. 1 and Rd. 2 of different AI explanations under AI-right (**e**) and AI-wrong (**f**) predictions. Each gray transparent scatter represents one measured sample. All presented bars or data points of a plot in this figure indicate the mean values of metrics measured. All error bars indicate the standard error of the mean. *P* values were calculated with post hoc marginal comparison after an initial analysis with a linear mixed model. ***$P < 0.001$, **$P < 0.01$, *$P < 0.05$. NS, not significant; Rd., round.

participant (Fig. 4a). As mentioned above, 'deferential participants' are those who got correct results only when AI was correct and always got incorrect results when AI was wrong (that is, a blue bar in Fig. 4a,d). Participants who got at least one correct outcome even when AI was incorrect are considered as 'non-deferential participants' (that is, a red bar on top of the blue bar in Fig. 4a,d).

Although the general public had similar performance after AI assistance, deferential participants had significantly lower initial diagnostic accuracy (65.8%) compared to non-deferential participants (72.5%) before receiving AI suggestions ($\beta = 0.082$, 95% CI: 0.074–0.111, $P < 0.001$, Cohen's $d = 0.462$; Fig. 4b; linear mixed model with deferential group as the main factor, controlling the same confounders as study 1 analysis) (Supplementary Table 28). Deferential PCPs also had a significantly lower performance in the initial round ($\beta = 0.194$, 95% CI: 0.079–0.310, $P < 0.001$, Cohen's $d = 1.578$; Fig. 4e and Supplementary Table 29), which could result from a lower level of critical thinking ($P = 0.028$, measured by critical thinking questionnaire[36]; see details in Methods).

LLM explanations led to the largest proportion of fully deferential participants in the general public (Fig. 4c), although no significance between LLM and others was observed ($P = 0.169$–0.433; Supplementary Table 30). By contrast, LLM resulted in the lowest proportion of deferential PCPs (Fig. 4f; $P = 0.268$–0.587; Supplementary Table 31). This is also aligned with our findings of the capability of experts in maintaining resilience against the misdirection of wrong semantic AI explanations[10].

Putting AI before human decisions amplifies deference across expertise levels

In addition to XAI methods, a practical design factor for human–AI collaboration systems is the decision-making order, either Human-First or AI-First paradigms. Both the general public and PCPs had significantly better performance in the first round with AI-First (all $P < 0.001$; Fig. 5a,d), which is not surprising due to superior AI performance. In the second round, after humans received the same amount of information, no difference was observed between Human-First and AI-First in either study (round 2, general public: $\beta = 0.005$, 95% CI: −0.017 to 0.026, $P = 0.650$; PCPs: $\beta = -0.020$, 95% CI: −0.091 to 0.050, $P = 0.572$; linear mixed models on accuracy with human–AI collaboration paradigm, decision round and their interaction as the main factors, controlling decision-making time and other confounders; see Supplementary Tables 32 and 34 for details). This indicates that the decision order may not influence the final performance. To exclude the influence where participants would be biased in the second round due to the prior exposure of the disease image, we compared AI-assisted performance in AI-First round 1 versus performance in Human-First round 2 and found no differences ($P > 0.05$ for both general public and PCPs; Supplementary Tables 38 and 39), indicating that the diagnosis strategies were not driven by the carryover effect.

With the AI-First paradigm, non-deferential participants still had better performance, especially after reviewing the examples again without AI (general public: $\beta = 0.031$, 95% CI: 0.000–0.062, $P = 0.049$; Supplementary Table 33; PCPs: $\beta = 0.218$, 95% CI: 0.009–0.426, $P = 0.041$; Supplementary Table 35). This is similar to the results in the Human-First paradigm in Fig. 4. By contrast, putting AI suggestions ahead increased the proportion of deferential participants in most cases (Fig. 5c,f). For the general public, the proportion was increased across all XAI methods (average $\Delta = +8.4\%$, although no significance was observed after controlling all confounders, $P = 0.067$–0.170; Fig. 5c and Supplementary Table 36). For PCPs, the largest deference increase was observed from LLM explanations ($\Delta = +19.0\%$; Fig. 5f,

https://doi.org/10.1038/s41591-026-04553-w

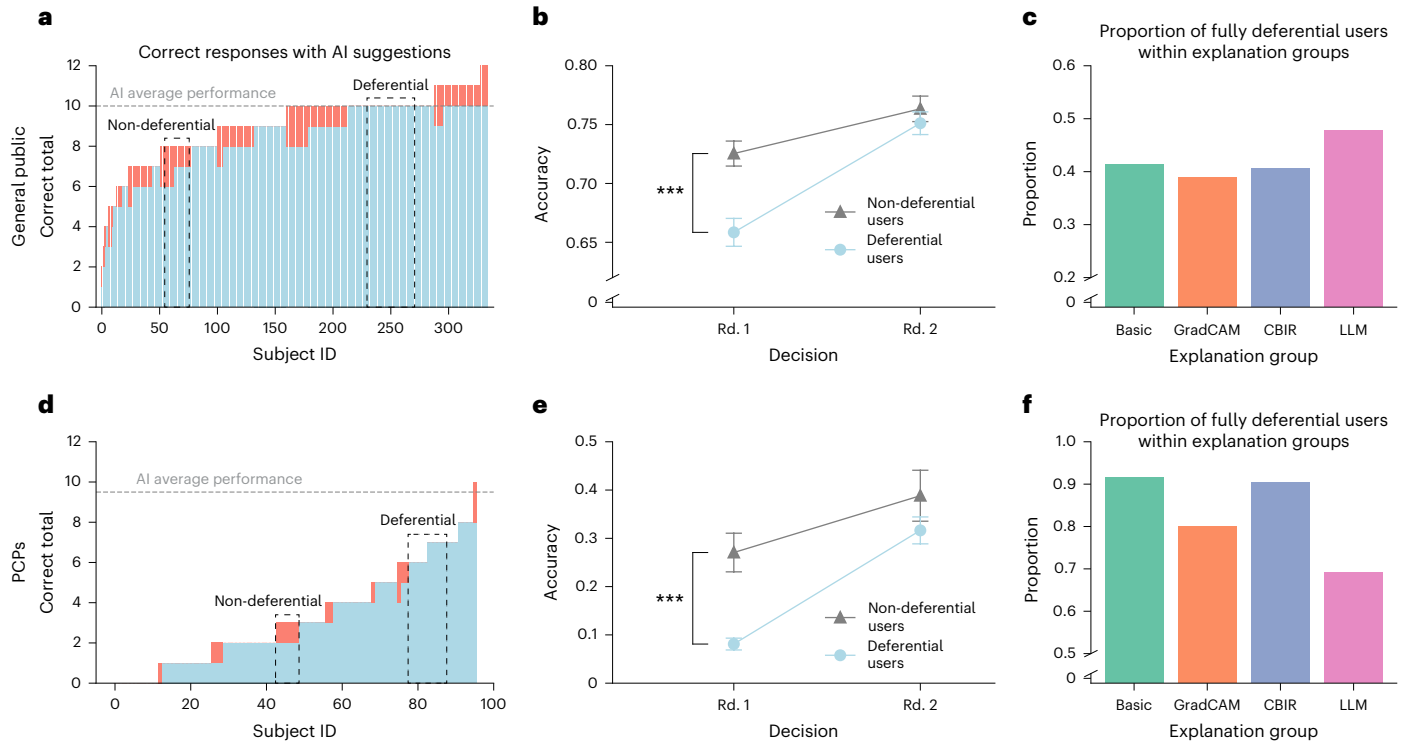


Fig. 4 | Deference patterns toward AI in human–AI collaboration. **a**,**b**, Performance comparison between deferential (fully compliant with AI suggestions) and non-deferential general populations. **a**, Distribution of correct responses with AI assistance. The blue bars represent instances where both AI's predictions and participants' predictions were correct; the red overlays indicate instances where participants' responses were correct while AI predictions were incorrect. **b**, Accuracy trajectories of deferential users ($n = 142$) and non-deferential users ($n = 193$) found from **a** across two decision rounds. **c**, Proportion of participants who defer to the output of each type of XAI (Basic ($n = 82$), GradCAM ($n = 72$), CBIR ($n = 91$) and LLM ($n = 90$)). **d**–**f**, Equivalent results for PCP participants as in **a**–**c** (in **e**, deferential PCPs $n = 79$, non-deferential PCPs $n = 17$; in **f**, Basic ($n = 24$), GradCAM ($n = 25$), CBIR ($n = 21$) and LLM ($n = 26$)). All presented bars or data points of a plot in this figure indicate the mean values of metrics measured. All error bars indicate the standard error of the mean. *P* values were calculated with post hoc marginal comparison after an initial analysis with a linear mixed model. ***$P < 0.001$, **$P < 0.01$, *$P < 0.05$. NS, not significant; Rd., round.

Extended Data Fig. 6m and Supplementary Table 37). This indicates that putting AI ahead may lead to stronger anchoring bias. It also suggests that, although PCPs showed resistance against AI's mislead in the Human-First paradigm, providing LLM-based explanations ahead of human choices can still cause more bias than other XAI methods and introduce risks of overreliance, even for PCPs.

Human–AI collaboration to combine each side s strength

Although AI deference risks misleading participants when AI makes mistakes, deference may lead some humans to improved performance. Figure 6 visualizes cases where either humans or AI routinely outperform each other. We found that AI tends to outperform humans in cases where the presentation of the disease is subtle but struggles with atypical symptoms or unexpected features in the image (see Supplementary Table 40 for example information). These qualitative examples provide some initial directions for future work in understanding the complementary strengths of humans and AI in dermatological diagnosis.

Discussion

Through two large-scale experiments, our work addresses how various XAI methods and human–AI decision paradigms impact collaborative diagnosis across expertise levels (general public versus PCPs). Our work reveals both the potential and challenges of human–AI collaboration in dermatology.

Our study demonstrated that state-of-the-art AI suggestions can create significant diagnostic improvements regardless of participants' expertise levels and task difficulty. Congruent with previous research[22], PCPs' diagnostic top-1 accuracy and top-3 accuracy are increased by 21.4% and 43.5% from a baseline of 11.5% and 16.1%, respectively. The general public's performance also increased by 6.1% from 69.7% to 75.8%. Our fairness-constrained AI model helped humans reduce diagnostic disparities on skin tones, with a relative reduction of 46.9% (accuracy disparity from 3.2% to 1.7%) among the general public and 35.6% (from 4.5% to 2.7%) among PCPs. Extended Data Fig. 7 further shows the correlation between the performance improvement on the subjective ratings on human–AI collaboration experience ($r = 0.35$, $P < 0.001$). In line with our results, Groh et al.[22] found similar disparities across skin tones without AI and similar reductions when PCPs accessed a similarly accurate AI tool (79.2% accuracy in the present study and 84.0% accuracy in the treatment deep learning system in Groh et al.[22]). However, when the AI assistance tool exhibited moderate performance (the 47.0% accuracy in the control deep learning system in Groh et al.[22]), PCP diagnostic disparities across skin tones increased. Together, these two studies reveal that the dose–response relationship between AI accuracy and its impact on diagnostic accuracy and equity deeply matters. As AI assistance accuracy decreases, physicians may be prone to incorporating incorrect AI predictions into their diagnoses especially for more challenging cases, such as images of dark skin.

We found interesting nuances of different XAI methods on different populations, especially LLMs. Although a direct comparison of the final diagnosis performance across XAI methods does not reveal significance (aligned with previous findings in Chanda et al.[37]), our breakdown analysis across AI correctness and XAI quality setups provides more nuanced insights. LLM acted as a 'double-edged sword' for the general public. Although it provided the highest accuracy improvements

https://doi.org/10.1038/s41591-026-04553-w

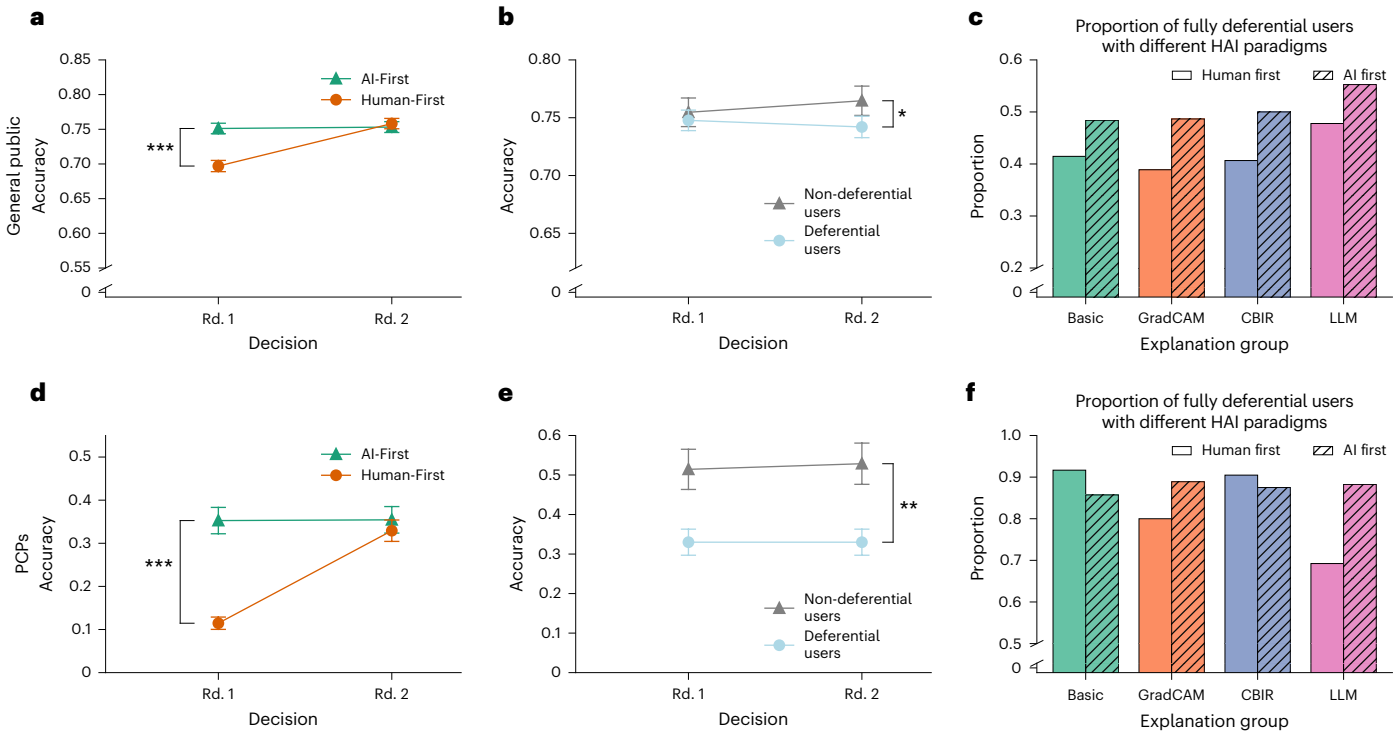


Fig. 5 | Influence of human–AI collaboration paradigm on diagnostic performance. a–d, Analysis for the general public. **a**, Comparison between AI-First ($n = 288$) and Human-First ($n = 335$) human–AI collaboration (HAI) systems on diagnostic accuracy across two rounds. **b**, Accuracy trajectories across two decision rounds for deferential and non-deferential people in the AI-First paradigm. **c**, Comparison of deferential user (participants who defer to the output of each type of XAI) proportion between two HAI paradigms across four XAI groups (AI-First: Basic ($n = 62$), GradCAM ($n = 74$), CBIR ($n = 76$) and LLM ($n = 76$); Human-First: Basic ($n = 82$), GradCAM ($n = 72$), CBIR ($n = 91$) and LLM ($n = 90$)). **d–f**, Equivalent results for PCP participants (AI-First PCPs $n = 57$, among which Basic ($n = 14$), GradCAM ($n = 18$), CBIR ($n = 8$) and LLM ($n = 17$); Human-First PCPs $n = 96$, among which Basic ($n = 24$), GradCAM ($n = 25$), CBIR ($n = 21$) and LLM ($n = 26$)) as **a–c**. All presented bars or data points of a plot in this figure indicate the mean values of metrics measured. All error bars indicate the standard error of the mean. *P* values were calculated with post hoc marginal comparison after an initial analysis with a linear mixed model. $^{***}P < 0.001$, $^{**}P < 0.01$, $^{*}P < 0.05$. NS, not significant; Rd., round.

(+13.4%) when its suggestions were right, it also led to the largest performance decrease (−21.1%) when wrong. The general public also tended to trust LLM explanations regardless of their quality, suggesting that readable, semantic explanations are more convincing for those without medical backgrounds. This heightened deference to LLMs likely stems from the cognitive ease of narrative persuasion[38,39]. Unlike structured concept-based explanations such as GradCAM, which requires a user to interpret an abstract heatmap, or CBIR, which demands a high-effort visual comparison, an LLM provides unstructured narrative rationales that create a cohesive and seemingly complete diagnostic story. It connects visual features to a conclusion with authoritative, human-like prose (for example, 'The lesion is suspicious because of its irregular border…'). These free-form narrative explanations can create an illusion of understanding, making the reasoning feel more intuitive and plausible to a nonexpert. Moreover, when the LLM highlights specific features, such as 'irregular border', it can anchor the user's perception, causing them to reinterpret the ambiguous visual data in a way that confirms the AI's plausible-sounding rationale. Our work extends previous research on nonexpert overreliance and automation bias on AI[40–42] and finds greater risk of LLM-based explanations, which has been supported in recent human–LLM trust work outside healthcare[43]. With numerous AI-assisted medical decision support tools available to the general public[44,45], and the advancement of LLM in health AI research[46–48], our findings highlight the importance of strong scrutiny and careful design for AI safety.

By contrast, PCPs applied LLM assistance more carefully and maintained their diagnostic accuracy even when AI suggestions were wrong. This was true regardless of explanation quality. To isolate the impact of task difficulty levels, we compared medical students to PCPs and found that students are more easily misled by LLM than PCPs. This suggests that the expertise of PCPs could be a mediator to help participants override or defer incorrect suggestions[49,50]. Although this contrasts with previous findings that non-dermatology experts are also misleading AI outcomes[30,51], our results are aligned with previous studies in dermatology[10] that experts are better at adopting AI than nonexperts. Our further investigation revealed that AI deference is less common when medical students and the PCP groups were equipped with more medical expertise. This indicates that medical knowledge and experience could be one of the key reasons for appropriate AI adoption for medical experts. Compared to the general public, experts' medical expertise acts as a crucial 'cognitive firewall'[30,52]. Although a lay person uses the explanation to 'form' a belief, a PCP uses it to 'validate' an existing hypothesis against their own structured clinical knowledge.

These findings have major design implications for how to implement AI suggestions for both the general public and PCPs. First, our work strongly suggests that a 'one-size-fits-all' approach to medical XAI is not only suboptimal but also potentially hazardous. Instead, systems must be carefully tailored to the user's expertise and the specific collaborative goal. For the general public, where the risk of overreliance on persuasive LLM narratives is high, designs should prioritize safety over persuasiveness. This may involve explicitly showing the AI's uncertainty and fallibility on a case-by-case basis to calibrate user trust, framing suggestions not as diagnoses but as preliminary information that requires professional validation. This is also supported by the recent call for medical safety disclaimer in generative AI model outcomes[53]. For clinicians, the goal shifts from simple guidance to expert augmentation. Systems should support adaptive explanations,

https://doi.org/10.1038/s41591-026-04553-w

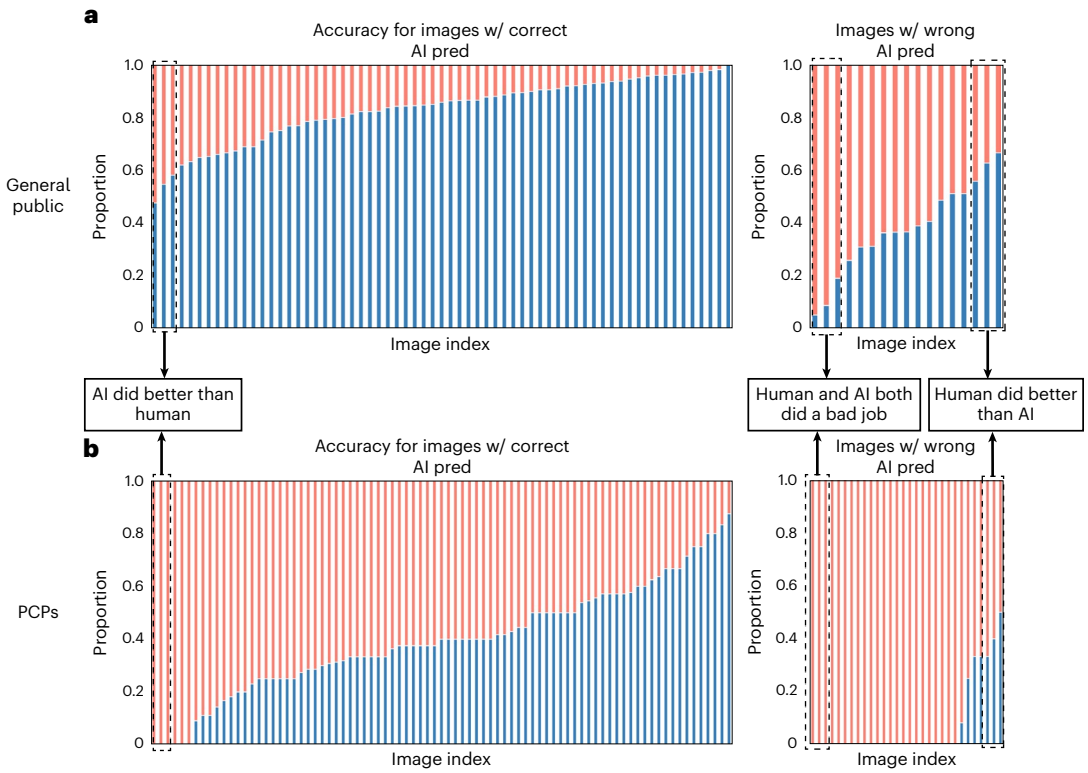


Fig. 6 | Case study characterizing when AI or humans did better. a, Accuracy proportion of images in study 1 for the general public, categorized by whether the AI prediction was correct or incorrect. The blue bars represent the proportion of correct human participants' decisions; the red overlays indicate their incorrect decisions. Representative images are labeled as 'AI did better than human' (AI-right images with the highest human error rate), and 'Human did better than AI' (AI-wrong images with the highest human accuracy) is shown to the right. **b**, Same as **a** for study 2 for PCPs. Detailed examples can be found in Supplementary Table 40. pred, prediction.

potentially offering a concise rationale for routine cases but revealing more in-depth evidence when the clinician's initial diagnosis differs from the AI's suggestion.

Furthermore, the structure of the interaction itself is a powerful design tool for mitigating cognitive bias. Our finding shows that AI suggestions first lead to stronger anchoring bias[54] and higher deference, which enriches the current understanding of the effect of human–AI decision orders[31,55]. This is clear evidence for a 'Human-First' workflow to preserve independent human reasoning and minimize the powerful anchoring bias that an upfront AI suggestion can create.

Finally, we need to design systems that account for inevitable disagreements and individual differences in AI deference. Our finding that participants prone to AI deference often have lower baseline performance reveals a crucial challenge: the users most in need of help are also the most vulnerable to being misled. This extends recent work[30] on AI deference among humans and points toward the future of an adaptive system. Instead of a static information handoff, the system should dynamically tailor its interaction style to individual users. A highly deferential user might require an iterative explanation[56] or explicit prompts for the user's confidence before finalizing a diagnosis[57] or automatically trigger a workflow for a second human opinion when the AI's suggestion meaningfully alters an initial assessment[58]. Designing these feedback loops is essential to better align AI reliance with actual user skill and minimize the risks of misjudgment in high-stakes medical contexts.

Our work has several key limitations. First, although we designed our study to be close to real-world scenarios, the ecological validity of the study is still limited. Participants did not have access to contextual data (for example, age, gender and socioeconomic status), which are crucial factors in real-world diagnosis, and the task designs were also simplified for both the general public and PCPs compared to real-world diagnosis. Second, our online studies include only 12 images per participant. This number may be insufficient to fully capture enough data to analyze individual patterns, potentially limiting the generalizability of our findings. Third, our results are specific to the AI models and XAI methods that we implemented. The characteristics of the post hoc explanations are inherently tied to the underlying model's training algorithm (that is, our model necessarily shapes the internal representations from which explanations are derived). This is a potential confounding factor common in studies of post hoc XAI. Although XAI explanations were evaluated by experts to ensure their utility, the performance of the basic model is still at a moderate level, which could bring potential influence on the distilled information. Moreover, to balance the samples of different diseases, we intentionally sampled test images, which did not reflect the realistic model behavior. Similarly, the stochasticity in LLM explanation generation, together with the specific prompt that may elevate a confident tone, introduces a confounder that is difficult to control. In our present study, LLM explanations were generated in a single pass without selection, refinement or post hoc optimization. Future work should address these limitations by developing more advanced predicting models, incorporating richer patient context in more authentic clinical workflows and evaluating a larger number of cases.

In conclusion, our work shows both the potential and challenges of using XAI, especially LLMs, in dermatology diagnosis. AI collaboration can improve diagnostic accuracy and fairness, but the substantial differences in how PCPs and the general public use AI assistance demonstrate the need for carefully designed systems. Our findings should be replicated in other medical domains and verified with additional patient information (for example, medical history and environment) in AI systems. More work should focus on making explanations helpful for different users while encouraging critical thinking.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41591-026-04553-w.

References

1. Malvehy, J. et al. Clinical performance of the Nevisense system in cutaneous melanoma detection: an international, multicentre, prospective and blinded clinical trial on efficacy and safety. *Br. J. Dermatol.* **171**, 1099–1107 (2014).
2. Venkatesh, K. P., Kadakia, K. T. & Gilbert, S. Learnings from the first AI-enabled skin cancer device for primary care authorized by FDA. *NPJ Digit. Med.* **7**, 156 (2024).
3. Resneck, J. Jr. Too few or too many dermatologists? Difficulties in assessing optimal workforce size. *Arch. Dermatol.* **137**, 1295–1301 (2001).
4. Uscher-Pines, L., Malsberger, R., Burgette, L., Mulcahy, A. & Mehrotra, A. Effect of teledermatology on access to dermatology care among medicaid enrollees. *JAMA Dermatol.* **152**, 905–912 (2016).
5. Krogue, J. D. et al. Searching for dermatology information online using images vs text: a randomized study. Preprint at *medRxiv* https://doi.org/10.1101/2024.10.25.24316155 (2024).
6. van der Velden, B. H. M., Kuijf, H. J., Gilhuijs, K. G. A. & Viergever, M. A. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Med. Image Anal.* **79**, 102470 (2022).
7. Esteva, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
8. Selvaraju, R. R. et al. Grad-CAM: visual explanations from deep networks via gradient-based localization. In *Proc. IEEE International Conference on Computer Vision* 618–626 https://doi.org/10.1109/ICCV.2017.74 (IEEE, 2017).
9. Gudivada, V. N. & Raghavan, V. V. Content based image retrieval systems. *Computer* **28**, 18–22 (1995).
10. Tschandl, P. et al. Human–computer collaboration for skin cancer recognition. *Nat. Med.* **26**, 1229–1234 (2020).
11. Hauser, K. et al. Explainable artificial intelligence in skin cancer recognition: a systematic review. *Eur. J. Cancer* **167**, 54–69 (2022).
12. AlSaad, R. et al. Multimodal large language models in health care: applications, challenges, and future outlook. *J. Med. Internet Res.* **26**, e59505 (2024).
13. Panagoulias, D. P., Virvou, M. & Tsihrintzis, G. A. Evaluating LLM—generated multimodal diagnosis from medical images and symptom analysis. Preprint at https://doi.org/10.48550/arXiv.2402.01730 (2024).
14. Goktas, P. & Grzybowski, A. Assessing the impact of ChatGPT in dermatology: a comprehensive rapid review. *J. Clin. Med.* **13**, 5909 (2024).
15. Pillai, A. et al. Evaluating the diagnostic and treatment capabilities of GPT-4 Vision in dermatology: a pilot study. *J. Cutan. Med. Surg.* **29**, 570–576 (2025).
16. Robinette, P., Li, W., Allen, R., Howard, A. M. & Wagner, A. R. Overtrust of robots in emergency evacuation scenarios. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* 101–108 https://doi.org/10.1109/HRI.2016.7451740 (IEEE, 2016).
17. Ghassemi, M., Oakden-Rayner, L. & Beam, A. L. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit. Health* **3**, e745–e750 (2021).
18. Fragiadakis, G., Diou, C., Kousiouris, G. & Nikolaidou, M. Evaluating human-AI collaboration: a review and methodological framework. Preprint at https://doi.org/10.48550/arXiv.2407.19098 (2024).
19. Tu, T. et al. Towards conversational diagnostic artificial intelligence. *Nature* **642**, 442–450 (2025).
20. Quinn, T. P., Senadeera, M., Jacobs, S., Coghlan, S. & Le, V. Trust and medical AI: the challenges we face and the expertise needed to overcome them. *J. Am. Med. Inform. Assoc.* **28**, 890–894 (2021).
21. Rosenbacke, R., Melhus, Å, McKee, M. & Stuckler, D. How explainable artificial intelligence can increase or decrease clinicians' trust in AI applications in health care: systematic review. *JMIR AI* **3**, e53207 (2024).
22. Groh, M. et al. Deep learning-aided decision support for diagnosis of skin disease across skin tones. *Nat. Med.* **30**, 573–583 (2024).
23. Alowais, S. A. et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC Med. Educ.* **23**, 689 (2023).
24. Han, S. S. et al. Augmented intelligence dermatology: deep neural networks empower medical professionals in diagnosing skin cancer and predicting treatment options for 134 skin disorders. *J. Invest. Dermatol.* **140**, 1753–1761 (2020).
25. Sharma, A., Lin, I. W., Miner, A. S., Atkins, D. C. & Althoff, T. Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nat. Mach. Intell.* **5**, 46–57 (2023).
26. Lin, H., Werner, K. M. & Inzlicht, M. Promises and perils of experimentation: the mutual-internal-validity problem. *Perspect. Psychol. Sci.* **16**, 854–863 (2021).
27. Mayanja, J., Asanda, E. H., Mwesigwa, J., Tumwebaze, P. & Marvin, G. Explainable artificial intelligence and deep transfer learning for skin disease diagnosis. In *Fourth International Conference on Image Processing and Capsule Networks* (eds Shakya, S., Tavares, J. M. R. S., Fernández-Caballero, A. & Papakostas, G.) 711–724 https://doi.org/10.1007/978-981-99-7093-3_47 (Springer, 2023).
28. Wang, S., Yin, Y., Wang, D., Wang, Y. & Jin, Y. Interpretability-based multimodal convolutional neural networks for skin lesion diagnosis. *IEEE Trans. Cybern.* **52**, 12623–12637 (2022).
29. Ueda, D. et al. Fairness of artificial intelligence in healthcare: review and recommendations. *Jpn. J. Radiol.* **42**, 3–15 (2024).
30. Gaube, S. et al. Do as AI say: susceptibility in deployment of clinical decision-aids. *NPJ Digit. Med.* **4**, 31 (2021).
31. Fogliato, R. et al. Who goes first? Influences of human-AI workflow on decision making in clinical imaging. In *Proc. 2022 ACM Conference on Fairness, Accountability, and Transparency* 1362–1374 https://doi.org/10.1145/3531146.3533193 (Association for Computing Machinery, 2022).
32. Jiang, L., Qin, X., Dong, X., Chen, C. & Liao, W. How and when AI-human order influences procedural justice in a multistage decision-making process. *Acad. Manag. Proc.* **2022**, 10419 (2022).
33. Daneshjou, R. et al. Disparities in dermatology AI performance on a diverse, curated clinical image set. *Sci. Adv.* **8**, eabq6147 (2022).
34. Li, J., Yang, Y., Liao, Q. V., Zhang, J. & Lee, Y.-C. As confidence aligns: understanding the effect of AI confidence on human self-confidence in human-AI decision making. In *Proc. 2025 CHI Conference on Human Factors in Computing Systems* (eds Evers, V. et al.) 1–16 https://doi.org/10.1145/3706598.3713336 (Association for Computing Machinery, 2025).
35. Said, A. On explaining recommendations with Large Language Models: a review. *Front. Big Data* **7**, 1505284 (2025).
36. Frederick, S. Cognitive reflection and decision making. *J. Econ. Perspect.* **19**, 25–42 (2005).
37. Chanda, T. et al. Dermatologist-like explainable AI enhances trust and confidence in diagnosing melanoma. *Nat. Commun.* **15**, 524 (2024).
38. Carrasco-Farre, C. Large Language Models are as persuasive as humans, but how? About the cognitive effort and moral-emotional language of LLM arguments. Preprint at https://doi.org/10.48550/arXiv.2404.09329 (2024).

39. Bansal, G. et al. Does the whole exceed its parts? The effect of ai explanations on complementary team performance. In *Proc. 2021 CHI Conference on Human Factors in Computing Systems* (eds Kitamura, Y. et al.) 1–16 https://doi.org/10.1145/3411764.3445717 (Association for Computing Machinery, 2021).
40. Lyell, D. & Coiera, E. Automation bias and verification complexity: a systematic review. *J. Am. Med. Inform. Assoc.* **24**, 423–431 (2017).
41. Goddard, K., Roudsari, A. & Wyatt, J. C. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *J. Am. Med. Inform. Assoc.* **19**, 121–127 (2012).
42. Larasati, R. Trust and explanation in artificial intelligence systems: a healthcare application in disease detection and preliminary diagnosis. https://doi.org/10.21954/ou.ro.00015aca (The Open University, 2023).
43. Kim, S. S. Y., Vaughan, J. W., Liao, Q. V., Lombrozo, T. & Russakovsky, O. Fostering appropriate reliance on large language models: the role of explanations, sources, and inconsistencies. In *Proc. 2025 CHI Conference on Human Factors in Computing Systems* (eds Evers, V. et al.) 1–19 https://doi.org/10.1145/3706598.3714020 (Association for Computing Machinery, 2025).
44. Abouelid, S., Liu, R. H., Desta, B. N., Chaurasia, A. & Ebrahim, S. The use of artificially intelligent self-diagnosing digital platforms by the general public: scoping review. *JMIR Med. Inform.* **7**, e13445 (2019).
45. Fraser, H., Coiera, E. & Wong, D. Safety of patient-facing digital symptom checkers. *Lancet* **392**, 2263–2264 (2018).
46. Kim, Y., Xu, X., McDuff, D., Breazeal, C. & Park, H. W. Health-LLM: large language models for health prediction via wearable sensor data. In *Proc. Fifth Conference on Health, Inference, and Learning* 522–539 https://proceedings.mlr.press/v248/kim24b.html (PMLR, 2024).
47. Senoner, J., Schallmoser, S., Kratzwald, B., Feuerriegel, S. & Netland, T. Explainable AI improves task performance in human–AI collaboration. *Sci. Rep.* **14**, 31150 (2024).
48. Strong, J., Men, Q. & Noble, J. A. Trustworthy and practical AI for healthcare: a guided deferral system with large language models. In *Proc. AAAI Conference on Artificial Intelligence* Vol. 39, 28413–28421 https://doi.org/10.1609/aaai.v39i27.35063 (AAAI, 2025).
49. Bayer, S., Gimpel, H. & Markgraf, M. The role of domain expertise in trusting and following explainable AI decision support systems. *J. Decis. Syst.* **32**, 110–138 (2022).
50. Nourani, M., King, J. & Ragan, E. The role of domain expertise in user trust and the impact of first impressions with intelligent systems. In *Proc. AAAI Conference on Human Computation and Crowdsourcing* (eds Aroyo, L. & Simperl, E.) 112–121 https://doi.org/10.1609/hcomp.v8i1.7469 (PKP Publishing Services, 2020).
51. Ayorinde, A. et al. Health care professionals' experience of using AI: systematic review with narrative synthesis. *J. Med. Internet Res.* **26**, e55766 (2024).
52. Bussone, A., Stumpf, S. & O'Sullivan, D. The role of explanations on trust and reliance in clinical decision support systems. In *2015 International Conference on Healthcare Informatics* 160–169 https://doi.org/10.1109/ICHI.2015.26 (IEEE, 2015).
53. Sharma, S., Alaa, A. M. & Daneshjou, R. A longitudinal analysis of declining medical safety messaging in generative AI models. *NPJ Digit. Med.* **8**, 592 (2025).
54. Gomez, C., Cho, S. M., Ke, S., Huang, C.-M. & Unberath, M. Human-AI collaboration is not very collaborative yet: a taxonomy of interaction patterns in AI-assisted decision making from a systematic review. *Front. Comput. Sci.* https://doi.org/10.3389/fcomp.2024.1521066 (2025).
55. Cabitza, F. et al. Rams, hounds and white boxes: investigating human–AI collaboration protocols in medical diagnosis. *Artif. Intell. Med.* **138**, 102506 (2023).
56. Kim, C. et al. Transparent medical image AI via an image–text foundation model grounded in medical literature. *Nat. Med.* **30**, 1154–1165 (2024).
57. Yan, S. et al. A multimodal vision foundation model for clinical dermatology. *Nat. Med.* **31**, 2691–2702 (2025).
58. Kolkur, S., Kalbande, D., Shimpi, P., Bapat, C. & Jatakia, J. Human skin detection using RGB, HSV and YCbCr color models. In *Proc. International Conference on Communication and Signal Processing 2016 (ICCASP 2016)* https://doi.org/10.2991/iccasp-16.2017.51 (Atlantis Press, 2016).

Xuhai 'Orson' Xu[1] ✉, Haoyu Hu[2], Haoran Zhang[3], Will Ke Wang[1], Reina Wang[3], Luis R. Soenksen[3,4], Omar Badri[5], Sheharbano Jafry[6], Elise Burger[7], Lotanna Nwandu[8], Apoorva Mehta[1], Erik P. Duhaime[9], Asif Qasim[10], Hause Lin[3], Janis Karleen Pereira[10], Jonathan Hershon[11], Paulius Mui[12], Alejandro A. Gru[1], Noémie Elhadad[1], Lena Mamykina[1], Matthew Groh[13], Philipp Tschandl[14], Roxana Daneshjou[6] & Marzyeh Ghassemi[3] ✉

[1]Columbia University, New York, NY, USA. [2]Cornell University, Ithaca, NY, USA. [3]Massachusetts Institute of Technology, Cambridge, MA, USA. [4]Johns Hopkins University, Baltimore, MD, USA. [5]Northeast Dermatology Associates, Beverly, MA, USA. [6]Stanford University, Stanford, CA, USA. [7]University of Utah, Salt Lake City, UT, USA. [8]OSF HealthCare, Peoria, IL, USA. [9]Centaur.AI, Boston, MA, USA. [10]MedShr, London, UK. [11]Pathway, Montréal, Québec, Canada. [12]X=Primary Care, Cambridge, MA, USA. [13]Northwestern University, Evanston, IL, USA. [14]Medical University of Vienna, Vienna, Austria.
✉e-mail: xx2489@cumc.columbia.edu; mghassem@mit.edu

Methods

User study design

Study populations. We designed two complementary large-scale digital studies to evaluate human–AI collaborative diagnostic performance (Fig. 1a). Both studies were based on clinical images and designed to resemble real-world practices: a regular lay person may take a photo of their skin and resort to a search engine or AI tools[59,60], and an expert often needs to make a differential diagnosis based on a clinical image (for example, patient communication through electronic health record messaging systems). To ensure data quality, we provided comprehensive tutorial materials for each participant group.

Study 1 engaged 623 participants (314 females, 305 males, four nonbinary or others, aged 33 ± 9 years) from the general public without medical backgrounds, in which participants reviewed clinical images of skin and performed a binary classification task to distinguish melanoma versus nevus (Fig. 1b). The general public assessed 12 images that were randomly sampled from a pool of 82 images. These images were evenly distributed between light-skinned (Fitzpatrick labels 1–4) and dark-skinned (Fitzpatrick labels 5–6) patients[61] and also evenly distributed between nevus and melanoma. To ensure consistent assessment of human–AI collaboration while maintaining ecological validity, we preserved AI model outputs while strategically sampling images to maintain a consistent accuracy of 83.3% (always 10 correct and two incorrect predictions per participant), which is close to the model's actual performance (see AI model details below). Each participant went through 12 clinical images and made two rounds of decisions per image (with and without AI).

Study 2 involved 153 PCPs (66 females, 84 males, three nonbinary or others, aged 34 ± 10 years, years of practice experience range 1–25 years, average experience 6 years). They performed a more complex open-ended differential diagnosis with free-text entry. Participants were asked to enter their top-3 diagnoses (Fig. 1c). In study 2, we focused on four main conditions previously identified as having potential diagnostic disparities across skin tones (better performance on patients with light skin than on those with dark skin)[22,33]: atopic dermatitis, pityriasis rosea, Lyme disease and CTCL. To simulate real-life applications, these four conditions were mixed with the other 30 common skin conditions. Text entry was assisted by auto-completion based on string matching from a comprehensive list of 445 skin diseases. Each participant evaluated 12 cases, including two images for each of the four main conditions with correct AI predictions, two images sampled from the four main conditions with incorrect AI predictions and two images sampled from other conditions with a mix of AI correctness (expected 1.5 correct and 0.5 incorrect), resulting in expected AI predictions to be 9.5 correct and 2.5 incorrect (that is, expected accuracy of 79.2%). Similar to study 1, skin tone was balanced in each condition.

We note that study 1 and study 2 are not directly comparable as they emphasize different tasks appropriate for groups with different medical expertise. To measure the impact of medical training experience within the same task, we additionally recruited 320 medical students (171 females, 141 males, eight nonbinary or others, aged 28 ± 9 years) with less expertise in dermatology to compare to the cohort reported in study 2 (Extended Data Figs. 5 and 6).

We further employed validated questionnaires to collect additional data on human–AI collaboration experience[62], XAI quality[63] and cognitive traits of critical thinking[36] and openmindness[64]. More details can be found in the experiment design sections in the Methods. Overall, we collected 14,952 diagnostic decisions from the general public (study 1), 3,672 diagnoses from PCPs and 7,680 diagnoses from medical students (study 2).

Human–AI decision paradigm. For both studies, we adopted a between-subject factorial design with 4 × 2 conditions, including four XAI methods: basic outcomes with model prediction and confidence; GradCAM, highlighting which areas contribute to the model prediction; CBIR, presenting similar images with the same label as the prediction; and multimodal LLM, semantic explanations of the reason for the model prediction (Fig. 1d–g). For the rest of the paper, we omit 'multimodal' and use LLM for simplicity, as our studies mainly focus on the language capability of the multimodal LLM, but do note that we used a vision–language model, GPT-4V (see Methods for more details), as well as two human–AI decision paradigms (Human-First, where users make a decision first before reviewing AI suggestions, and AI-First, where users review both images and AI suggestions before making the final decision). Each participant was randomly assigned to one condition and went through 12 clinical images. Because the Human-First condition naturally required participants to make two rounds of decisions (the first round without AI and the second round with AI), to control the effect of cognitive effort, participants also made two decision rounds in AI-First: after making the first round of decisions with AI, they were asked to review and reconsider their final decisions in the second round when AI results were hidden. Besides decisions, participants also rated their confidence in their decisions at each round.

Model training. We used Python (specifically PyTorch 1.13) and trained a deep learning model, ViT-B/32 (ref. 65), for the binary classification task in study 1. We applied the training strategy of CDANN[66] to achieve the fairness constraint and ensure performance balance between skin tones (weighted AUROC of 0.930, accuracy Δ = 2.1% between skin tones). In study 2, we trained a DenseNet-121 (ref. 67) with an AUROC of 0.772 and accuracy Δ = 5.6% between skin tones. Post hoc explanations (GradCAM, CBIR and LLM) were generated using the same trained models' outputs (see the Methods section for more details).

Evaluation metrics. Responses for study 1 were evaluated using standard binary outcome measures. For the free-text task in study 2, two authors with medical expertise manually evaluated all text entered in the 'Diagnosis' section to determine the top-1 accuracy (based on the first entry) and the top-3 accuracy (based on all three entries) compared to the ground truth.

All AI-generated explanations for GradCAM, CBIR and LLM were evaluated by three dermatologists with at least 2 years of experience and independently scored for correctness (that is, how accurate the explanation is based on the skin image) and informativeness (that is, how much this explanation can support diagnosis) with access to the ground truth on a 5-point Likert scale, achieving fair to moderate interrater reliability (Fleiss' κ = 0.42 and 0.37). We calculated the average and divided explanation quality into high (informativeness score ≥3 and correctness score ≥3) and low (informativeness score <3 or correctness score <3). As a validation of the AI explanation, the correctness score was higher when AI made correct predictions (general public: AI correct 3.63 ± 1.05 versus AI incorrect 2.74 ± 1.13, t = 5.279, P < 0.001, Cohen's d = 0.830, two-sided unpaired t-test; PCPs: 3.32 ± 1.43 versus 2.19 ± 0.95, t = 11.319, P < 0.001, Cohen's d = 0.957) as well as the informativeness score (general public: 3.25 ± 1.05 versus 2.42 ± 0.96, t = 5.087, P < 0.001, Cohen's d = 0.800; PCPs: 3.14 ± 1.44 versus 1.91 ± 0.84. t = 12.643, P < 0.001, Cohen's d = 1.069). These ensured that AI had learned useful information in diagnosing.

Image data collection

We collected a set of diverse datasets of clinical images spanning multiple skin tones for both studies with the general public and experts. Specifically, we combined 16 public and private datasets (see Supplementary Table 4 for details of each dataset), resulting in a total of 108,585 images after removing duplicates. These images contain human expert-annotated ground truth labels and cover a wide range of skin conditions, including nevus (n = 24,746, 22.8%), melanoma (n = 8,686, 8.0%), atopic dermatitis (n = 1,912, 1.8%), pityriasis rosea (n = 609, 0.6%), Lyme disease (n = 248, 0.2%), CTCL (n = 330, 0.3%) and other skin diseases. In the study 2 differential diagnosis task, we

mainly focused on four diseases—atopic dermatitis, pityriasis rosea, Lyme disease and CTCL—as these four were found to have the greatest performance disparities across skin tones in a recent study[22]. In total, 18,265 images (16.8%) also contain expert-annotated Fitzpatrick labels (Fitzpatrick labels 1–4: 15,845 of 18,265 images, 86.8%; Fitzpatrick labels 5–6: 2,420 images, 13.2%). For the rest, we used individual typology angle (ITA) calculated from the segmented skin pixels with the YCbCr algorithm[58,61]. Using the thresholds from Kinyanjui et al.[68], we assigned Fitzpatrick labels to these images. Note that these images with weak Fitzpatrick labels were only used for fair model training. For the actual user study, we randomly sampled 82 high-quality, expert-annotated images from the image pool for study 1 with the general public (equally distributed between nevus and melanoma and between Fitzpatrick labels 1–4 and 5–6) and 182 images for study 2 with medical experts (82 atopic dermatitis, 54 pityriasis rosea, 27 Lyme disease, 27 CTCL and 92 other diseases, equally distributed between skin tones within each category). The sample sizes balanced the variance of the images and adequate number of decisions per image.

AI models and explanations

We trained fair AI models for skin disease diagnosis. For study 1, we trained a binary classification model (nevus versus melanoma). For study 2, we implemented a hierarchical structured multiclass model that distinguished a total of 34 most common skin diseases (see the total list in Supplementary Table 4). We trained two models, including a primary five-class classification model (four main diseases and an 'other' class) and a secondary 30-class classification model (30 other common diseases). The secondary model was only used when the first model predicted 'other'.

All three models were trained following the protocol in Yang et al.[69] to achieve fair medical AI models. Specifically, for each task, we trained a set of models using five well-established fairness algorithms, including ERM[70], Domain Adversarial Neural Networks (DANN)[71], Conditional DANN (CDANN)[72], group distributionally robust optimization (GroupDRO)[73] and exponential moving average[74]. We performed a grid search on three pretrained models (DenseNet-121 (ref. 67) pretrained on ImageNet 1k[75], ViT-B/32 (ref. 65) pretrained on ImageNet 1k and ViT-B/32 pretrained on ImageNet 21k[76]) and model hyperparameters via a random search of 25 trials. Each model was fine-tuned for up to 30,000 steps with a batch size of 64 using the Adam optimizer, with checkpointing every 1,000 steps and early stopping if the overall validation macro AUROC does not improve for five checkpoints. For each algorithm, we first selected the architecture and hyperparameter combination that maximizes the worst-group validation AUROC. We then selected algorithms based on the tradeoff between overall accuracy and fairness. The final selected configurations were as follows. For study 1, the binary classification model used CDANN with a ViT-B/32 backbone pretrained on ImageNet-21k ($\lambda = 0.218$, learning rate = 0.0005). For study 2, the primary five-class model used CDANN with a DenseNet-121 backbone pretrained on ImageNet-1k ($\lambda = 8.388$, learning rate = 0.0029), and the secondary 30-class model used CDANN with a ViT-B/32 backbone pretrained on ImageNet-1k ($\lambda = 0.377$, learning rate = 0.0003). All three configurations were selected against the ERM baseline using the tradeoff procedure described above, and their performance is reported in the Results section. Supplementary Tables 5 and 6 list out the specific performance details of each disease. Finally, we intentionally sampled images from our testing set with high-quality disease and skin tone labels (that is, diagnosis verified through human experts such as biopsy and image resolution is higher than 400 × 400 pixels) to achieve the expected accuracy and fair results presented to participants, as described in the Results section.

Building upon the model, we then generated three types of post hoc explanations (GradCAM, CBIR and multimodal LLM). For GradCAM, we adopted the algorithm in Selvaraju et al.[8] and picked the last layer in the model to generate a heatmap and overlaid it on the original image. For CBIR, we computed the cosine distance of the embedding of the target image against the embeddings of images in the training set and selected the top-2 images with the same skin tone and the top-1 image with the opposite skin tone, leading to three retrieved images for each target image. As for multimodal LLM-based explanations, we used multimodal GPT-4V (a vision–language model) and constructed the prompt to generate explanations for the target reader (the general public or medical experts; see more prompt details in Supplementary Table 3). Note that the diagnoses were determined by our own model (including cases when predictions were wrong) and sent to GPT-4V using the OpenAI API, and GPT-4V was only prompted to generate text-based explanations rather than perform diagnosis.

Ethics approval for human subject studies

This research complies with all relevant ethical regulations. The Massachusetts Institute of Technology's Committee on the Use of Humans as Experimental Subjects approved this study as Exempt Category 3—Benign Behavioral Intervention (COUHES no. E-5365). Participants were informed about the potential sensitive visual content and provided consent before joining the study. They had the option to leave the experiment at any time. Images used in the study were publicly released by previous work. All clinical images displayed in this paper were obtained from publicly released datasets, for which informed consent for publication and research use was obtained by the original dataset curators in accordance with each dataset's release terms.

Experimental interface

We developed a reactive experiment interface based on Qualtrics for online digital experiments. For study 1, the experiment was designed to mimic real-life scenarios where users can self-diagnose using a photo captured by smartphones/webcams. Prior to the study, the general public participants were provided with educational content and common recognition guidelines about melanoma (ABCDE rules[77]; Supplementary Fig. 1), together with practice questions (Supplementary Fig. 2) and experiment instructions of AI and its explanations (Supplementary Fig. 3). They then went through 12 images and made a diagnosis (distinguishing melanoma versus nevus as well as their confidence of the decision). Participants could hover on each image (including GradCAM or CBIR images) to use a magnifier for closer inspection. The interface design of study 2 was similar to study 1. Expert participants were asked to do open-ended differential diagnosis as free-text entry tasks, in which they were asked to enter top-3 diagnoses, confidence and whether they had opinions on referral (including no referral, a dermatologist, biopsy and both). Each text entry box had auto-completion based on the matched string they entered, and the corpus included 445 (Supplementary Fig. 5a). These setups were designed to resemble real-world differential diagnosis scenarios, such as telehealth, or patients sending images through electronic health record messaging systems. They received task-specific instructions with examples (Supplementary Fig. 4) and practice cases (Supplementary Fig. 5), before going through 12 images.

Human subject study design and protocol

We recruited general public participants from Prolific (https://www.prolific.com/) and Facebook Groups for study 1 (20 March to 1 May 2024). In study 2 (10 June to 3 September 2024), we leveraged several clinical networks and platforms to recruit medical experts, including local networks at the Boston Medical Center and Stanford University, Centaur Lab (https://www.centaurlabs.com/), MedShr (https://en.medshr.net/), Pathway (https://www.pathway.md/) and XPC (https://www.xprimarycare.com/). For both studies, we embedded attention check questions in the middle of the study (see an example in Supplementary Fig. 5c) for quality control. We also filtered answers with abnormal durations (that is, fewer than 10 seconds or more than 5 minutes per image on average). After filtering, we collected 623 lay

people from the general public for study 1 and 153 PCPs for study 2 (320 medical students in addition).

We adopted a between-subjects design. Participants were randomly assigned to one of the eight experimental conditions varying in the four XAI types (basic confidence explanations, GradCAM, CBIR or LLM) and two human–AI decision paradigms (Human-First versus AI-First). In both studies, each participant evaluated 12 images and provided two rounds of diagnoses per image. In the Human-First paradigm, participants made unassisted initial diagnoses and reported their confidence before receiving AI input at round 1. They then reviewed the AI suggestions and explanations (depending on their experiment group) and reported round 2 decision and confidence. In the AI-First paradigm, AI suggestions and explanations were presented alongside the initial image at round 1. In round 2, AI outcomes were hidden, and participants were asked to review their decisions again. By introducing round 2 in the AI-First condition, we controlled the number of decisions and cognitive engagement.

Assessment of individual characteristics

In addition to the diagnostic tasks of 12 images, participants completed a series of questionnaires, including demographics and established assessments to measure human–AI collaboration experience[62] and AI explanation quality[63] (Supplementary Table. 2). For medical experts, we further collected additional cognitive traits, including openmindness[64] and critical thinking[36] (Supplementary Table 1). Although it is not the main focus of our work, the relationship between these subjective aspects and the final decision performance is presented in Extended Data Fig. 5.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

For machine learning model training, our datasets are organized and available in the figshare subfolder 'ml-model' (https://figshare.com/s/eaf2110da509abbe9a6a). For the two human subject studies, the images used in the human experiment and the study results are organized and available in the figshare subfolder 'human-study-results-analysis' (https://figshare.com/s/eaf2110da509abbe9a6a). Source data are provided with this paper.

Code availability

The data and code are organized together. For machine learning model training, our model training code and explanation generation code are organized and available in the figshare subfolder 'ml-model' (https://figshare.com/s/eaf2110da509abbe9a6a). For the two human subject studies, the code for data analysis visualization is organized and available in the figshare subfolder 'human-study-results-analysis' (https://figshare.com/s/eaf2110da509abbe9a6a).

References

59. Jain, A. et al. Development and assessment of an artificial intelligence–based tool for skin condition diagnosis by primary care physicians and nurse practitioners in teledermatology practices. *JAMA Netw. Open* **4**, e217249 (2021).
60. Lee, P., Bubeck, S. & Petro, J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N. Engl. J. Med.* **388**, 1233–1239 (2023).
61. Groh, M. et al. Evaluating deep neural networks trained on clinical images in dermatology with the Fitzpatrick 17k dataset. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* 1820–1828 https://doi.org/10.1109/CVPRW53098.2021.00201 (IEEE, 2021).
62. Jian, J.-Y., Bisantz, A. M. & Drury, C. G. Foundations for an empirically determined scale of trust in automated systems. *Int. J. Cogn. Ergon.* https://doi.org/10.1207/S15327566IJCE0401_04 (2000).
63. Hoffman, R. R., Mueller, S. T., Klein, G. & Litman, J. Measures for explainable AI: explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. *Front. Comput. Sci.* https://doi.org/10.3389/fcomp.2023.1096257 (2023).
64. Haran, U., Ritov, I. & Mellers, B. A. The role of actively open-minded thinking in information acquisition, accuracy, and calibration. *Judgm. Decis. Mak.* **8**, 188–201 (2013).
65. Dosovitskiy, A. et al. An image is worth 16x16 words: transformers for image recognition at scale. In *9th International Conference on Learning Representations* https://openreview.net/forum?id=YicbFdNTTy (ICLR, 2021).
66. Long, M., Cao, Z., Wang, J. & Jordan, M. I. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems* 1640–1650 (NeurIPS, 2018).
67. Huang, G., Liu, Z., van der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 2261–2269 https://doi.org/10.1109/CVPR.2017.243 (IEEE, 2017).
68. Kinyanjui, N. M. et al. Fairness of classifiers across skin tones in dermatology. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020* (eds Martel, A. L. et al.) 320–329 https://doi.org/10.1007/978-3-030-59725-2_31 (Springer, 2020).
69. Yang, Y., Zhang, H., Gichoya, J. W., Katabi, D. & Ghassemi, M. The limits of fair medical imaging AI in real-world generalization. *Nat. Med.* **30**, 2838–2848 (2024).
70. Vapnik, V. N. An overview of statistical learning theory. *IEEE Trans. Neural Netw.* **10**, 988–999 (1999).
71. Ganin, Y. et al. Domain-adversarial training of neural networks. In *Domain Adaptation in Computer Vision Applications* (ed Csurka, G.) 189–209 https://doi.org/10.1007/978-3-319-58347-1_10 (Springer, 2017).
72. Li, Y. et al. Deep domain generalization via conditional invariant adversarial networks. In *Computer Vision – ECCV 2018: 15th European Conference* (eds Ferrari, V., Hebert, M., Sminchisescu, C. & Weiss, Y.) 647–663 https://doi.org/10.1007/978-3-030-01267-0_38 (Springer, 2018).
73. Sagawa, S., Koh, P. W., Hashimoto, T. B. & Liang, P. Distributionally robust neural networks for group shifts: on the importance of regularization for worst-case generalization. In *8th International Conference on Learning Representations* https://openreview.net/forum?id=ryxGuJrFvS (ICLR, 2020).
74. Polyak, B. T. & Juditsky, A. B. Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim.* **30**, 838–855 (1992).
75. Deng, J. et al. ImageNet: a large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* 248–255 https://doi.org/10.1109/CVPR.2009.5206848 (IEEE, 2009).
76. Ridnik, T., Ben-Baruch, E., Noy, A. & Zelnik-Manor, L. ImageNet-21K pretraining for the masses. In *Proc. Neural Information Processing Systems Track on Datasets and Benchmarks 1* (eds Vanschoren, J. & Yeung, S.) https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/98f13708210194c475687be6106a3b84-Abstract-round1.html (NeurIPS Datasets and Benchmarks, 2021).
77. Tsao, H. et al. Early detection of melanoma: reviewing the ABCDEs. *J. Am. Acad. Dermatol.* **72**, 717–723 (2015).

Acknowledgements

We thank all participants for taking part in our study.

Author contributions

X.O.X. led the project and was responsible for conceptualization, study design, experimental protocol development, institutional review board and data collection infrastructure, participant recruitment coordination, data analysis planning, manuscript writing and overall project management. H.H. conducted statistical analyses and figure generation and contributed to manuscript writing and revision. H.Z.

conducted AI model training and development. M. Ghassemi, M. Groh, R.D. and P.T. supported study design, results analysis strategy and interpretation of findings. S.J., E.B. and L.N. provided clinical expertise on XAI quality evaluation. A.M., E.P.D., A.Q., H.L., J.K.P., J.H., P.M. and A.A.G. provided network support for participant recruitment. L.R.S. and O.B. provided access to private datasets for model training. W.K.W., R.W., N.E. and L.M. provided manuscript writing and review support. All authors reviewed and approved the final manuscript.

Funding

This work was supported, in part, by a National Science Foundation CAREER Award (2339381, to M. Ghassemi), an AI2050 Early Career Fellowship (G-25-68042, to M. Ghassemi) and NBER Center for Aging and Health Research grant P30AG012810 (to M. Ghassemi). We also gratefully acknowledge support from the Columbia University Research Stabilization Grant (to X.O.X.). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Competing interests

X.O.X. holds a part-time visiting faculty researcher position at Google. The work presented in this paper was conducted independently of, and is unrelated to, his responsibilities at Google. The remaining authors declare no competing interests.

Additional information

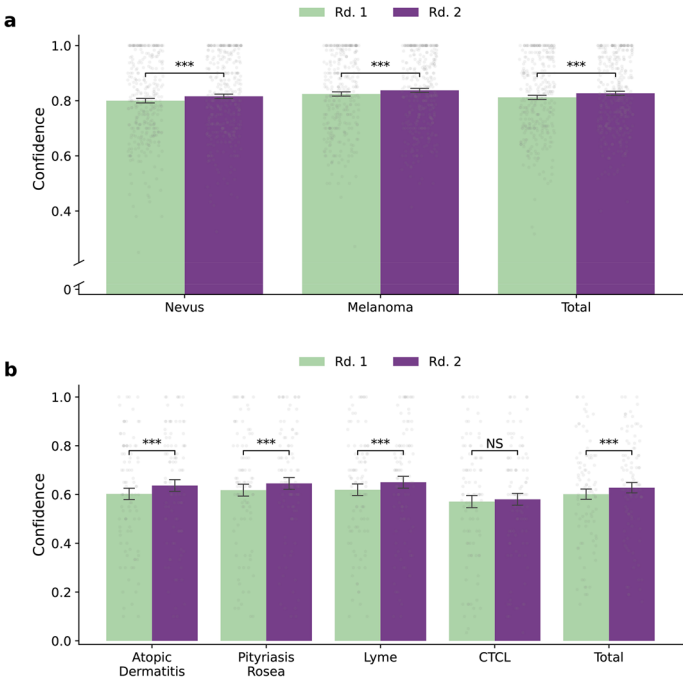
Extended data is available for this paper at https://doi.org/10.1038/s41591-026-04553-w.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41591-026-04553-w.

Correspondence and requests for materials should be addressed to Xuhai 'Orson' Xu or Marzyeh Ghassemi.

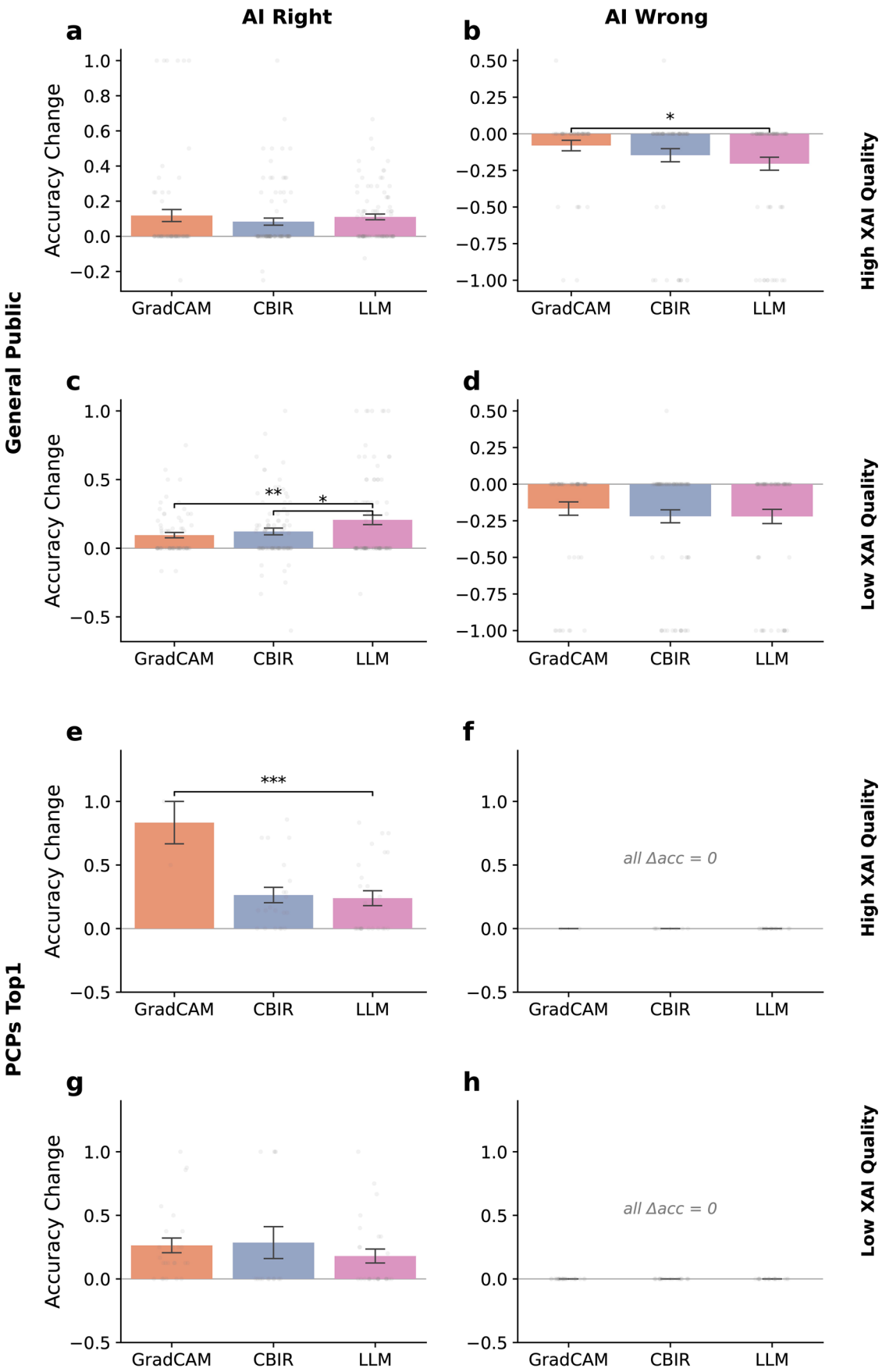
Peer review information *Nature Medicine* thanks Titus Brinker, Avishek Choudhury and Zongyuan Ge for their contribution to the peer review of this work. Peer reviewer reports are available. Primary Handling Editor: Michael Basson, in collaboration with the *Nature Medicine* team.

https://doi.org/10.1038/s41591-026-04553-w



Extended Data Fig. 1 | Diagnostic Confidence of the General Public and PCPs. **a**, Confidence levels of the general public (*N*=335) in detecting nevus, melanoma, and the overall diagnosis in two rounds: Rd. 1 (initial, human accuracy without AI assistance) and Rd. 2 (AI-assisted accuracy). **b**, Confidence levels of PCPs (*N*=96) in identifying atopic dermatitis, pityriasis rosea, Lyme disease, CTCL, and the overall performance on four main diseases, with results presented for two rounds. Each gray transparent scatter represents one measured sample. All presented bars or data points of a plot in this figure indicate the mean values of metrics measured. All error bars indicate the standard error of the mean. *P*-values are calculated with post-hoc marginal comparison after an initial analysis with linear mixed model. "***" means $p<0.001$, "**" means $p<0.01$, "*" means $p<0.05$, "NS" means no statistical significance.

https://doi.org/10.1038/s41591-026-04553-w

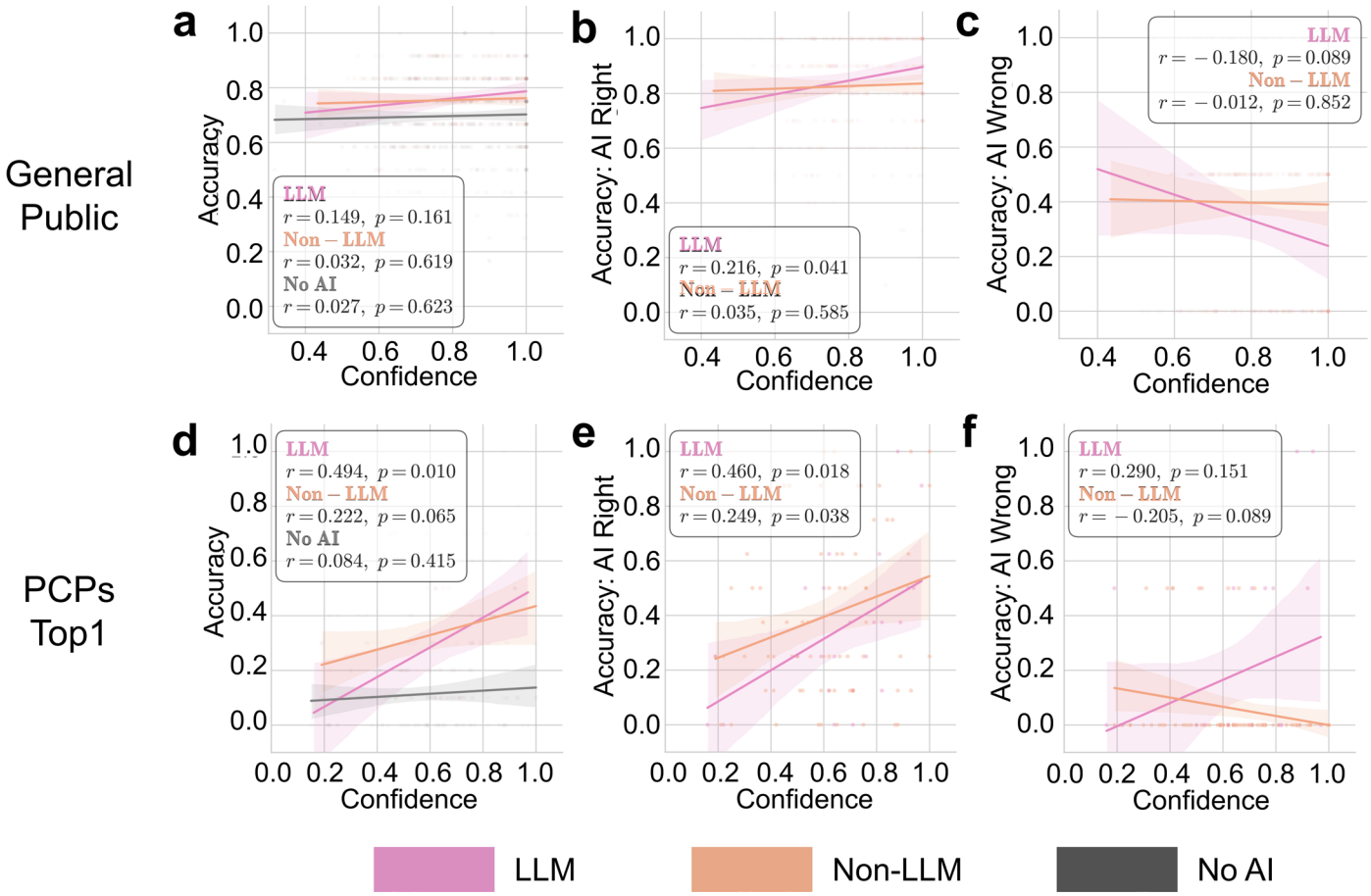


Extended Data Fig. 2 | See next page for caption.

https://doi.org/10.1038/s41591-026-04553-w

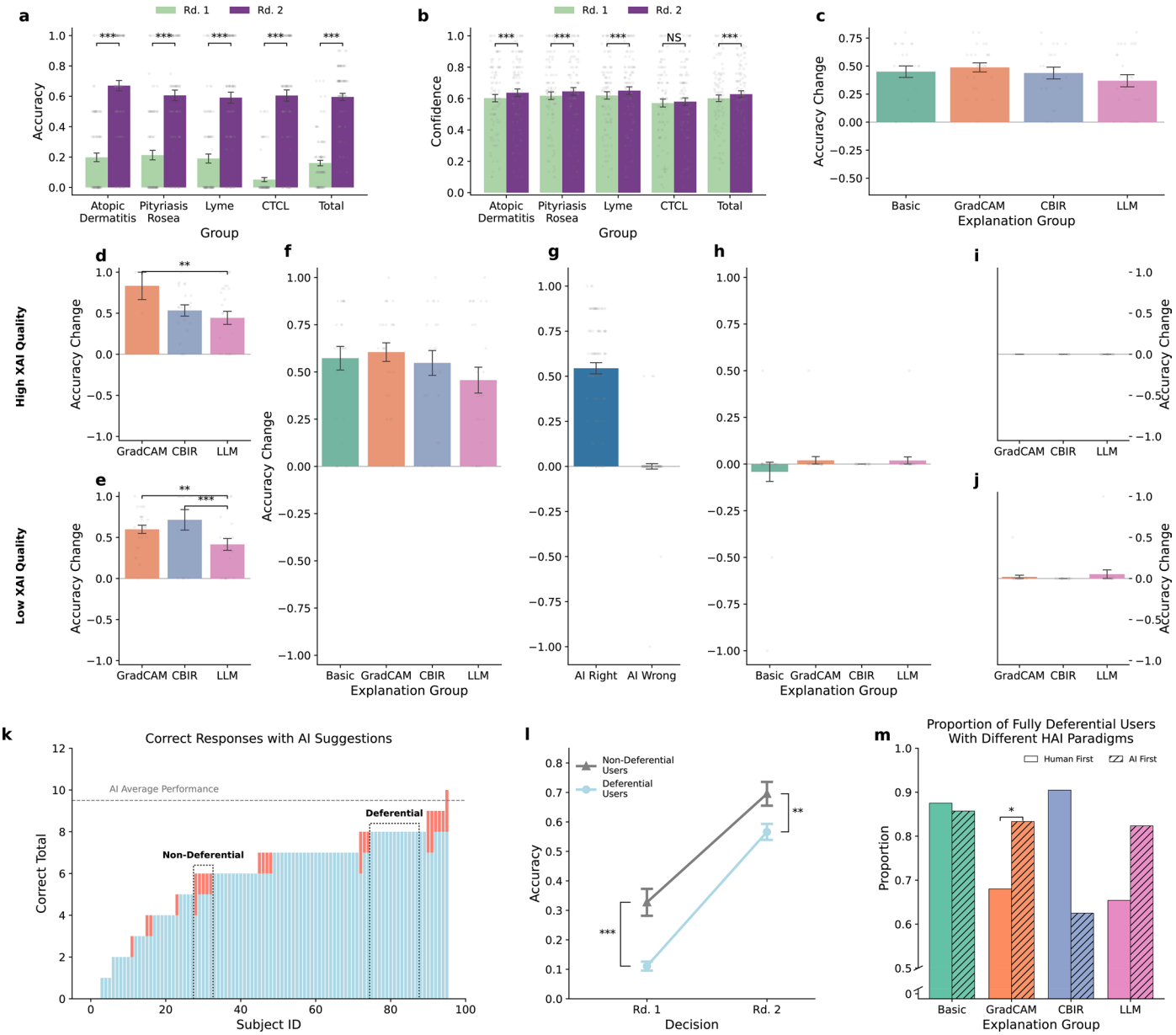
Extended Data Fig. 2 | Diagnostic Accuracy Under Different AI Explanation Qualities. a, **b**, Changes in diagnostic accuracy between the second diagnosis round with AI assistance and the first diagnosis round without AI assistance for the general public under high AI explanation (XAI) quality, comparing GradCAM, CBIR, and LLM explanations when AI is right (**a**) and AI is wrong (**b**) (AI right: GradCAM (N=72), CBIR (N=89), LLM (N=90); AI wrong: GradCAM (N=50), CBIR (N=65), LLM (N=71)). **c**, **d**, Visualization of diagnostic accuracy changes for the general public under low XAI quality conditions, highlighting the effects of right AI suggestions (**c**) and wrong suggestions (**d**) for each XAI approach (AI right: GradCAM (N=72), CBIR (N=91), LLM (N=89); AI wrong: GradCAM (N=60), CBIR (N=82), LLM (N=68)). **e**-**h**, Equivalent results as (**a**)-(**d**), shown for top1 diagnostic accuracy of PCP participants (High XAI quality, AI right: GradCAM (N=3), CBIR (N=21), LLM (N=26); AI wrong: GradCAM (N=2), CBIR (N=9), LLM (N=18). Low XAI quality, AI right: GradCAM (N=25), CBIR (N=14), LLM (N=26); AI wrong: GradCAM (N=25), CBIR (N=20), LLM (N=19)). Each gray transparent scatter represents one measured sample. All presented bars or data points of a plot in this figure indicate the mean values of metrics measured. All error bars indicate the standard error of the mean. *P*-values are calculated with post-hoc marginal comparison after an initial analysis with linear mixed model. "***" means $p<0.001$, "**" means $p<0.01$, "*" means $p<0.05$, "NS" means no statistical significance.

https://doi.org/10.1038/s41591-026-04553-w



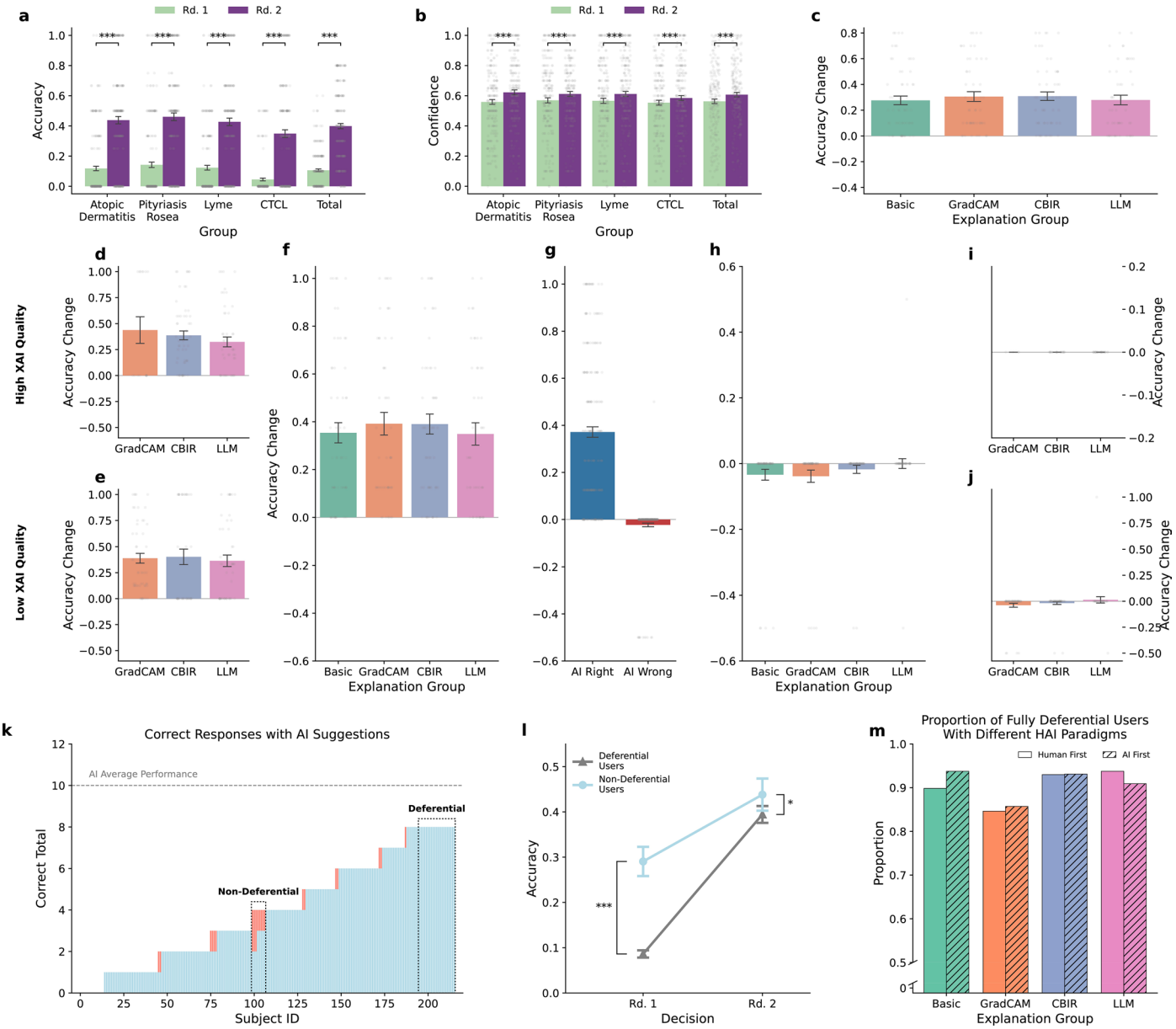
Extended Data Fig. 3 | Accuracy-Confidence Consistency of the General Public and Top1 Diagnosis of PCPs. a-c, General public's performance comparison between LLM and non-LLM assisted systems (LLM: N=90, Non-LLM: N=245, No AI: N=335), in (a) overall diagnostic accuracy-confidence consistency (A-C consistency) (**b**) A-C consistency with correct AI predictions, and (**c**) A-C consistency with wrong AI predictions. **d-f**, Equivalent results for top1 diagnosis of PCP participants (LLM: N=26, Non-LLM: N=70, No AI: N=96) as (**a**)-(**c**). Statistical significance was assessed using Pearson correlation coefficients (r) with associated p-values. Each transparent scatter represents one measured sample. Data is presented as predicted values by the linear regression model with accuracy as the dependent variable and confidence as the independent variable +/- 95% confidence intervals.

https://doi.org/10.1038/s41591-026-04553-w



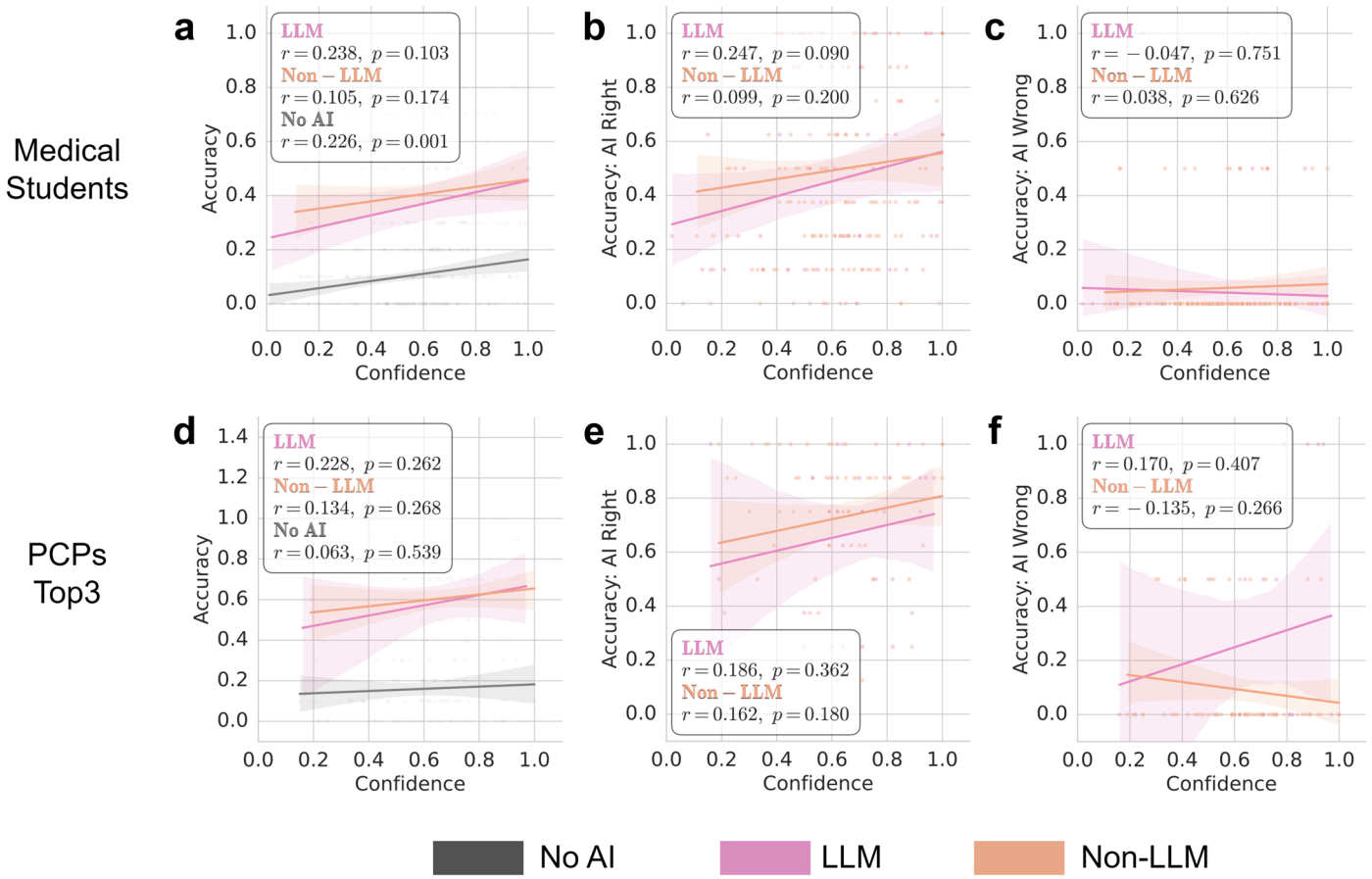
Extended Data Fig. 4 | PCP Performance - Top3 Accuracy. a, Accuracy in identifying atopic dermatitis, pityriasis rosea, Lyme disease, CTCL, and the overall performance on four main diseases, with results presented for two rounds (N=96). **b**, Confidence levels of participants in their diagnostic decisions for tasks in (**a**). **c**, Diagnostic accuracy improvement from the first diagnosis round without AI assistance to the second diagnosis round with AI assistance across different AI explanation (XAI) (Basic (*N*=24), GradCAM (*N*=25), CBIR (*N*=21), and LLM (*N*=26)). **d, e**, Changes in diagnostic accuracy between two diagnosis rounds under high XAI quality (**d**) and low XAI quality (**e**) when AI is right, comparing GradCAM, CBIR, and LLM explanations (High XAI quality, AI right: GradCAM (N=3), CBIR (N=21), LLM (N=26); AI wrong: GradCAM (N=2), CBIR (N=9), LLM (N=18). Low XAI quality, AI right: GradCAM (N=25), CBIR (N=14), LLM (N=26); AI wrong: GradCAM (N=25), CBIR (N=20), LLM (N=19)). **f, h** Diagnostic accuracy changes under AI-right (**f**) and AI-wrong (**h**) predictions for different XAI approaches. **g**, Overall accuracy changes under AI-right and AI-wrong conditions, integrating the effects of explanation quality and correctness. **i, j**, Visualization of diagnostic accuracy changes under AI-wrong conditions, highlighting the effects of good XAI quality (**i**) and low XAI quality (**j**). **k-l**, Performance comparison between deferential and non-deferential participants (Deferential N=74, Non-deferential N=22). (**k**) Distribution of correct responses with AI assistance, (**l**) accuracy trajectories across two decision rounds. **m**, Proportion of fully deferential participants across different XAI approaches and human-AI collaboration (HAI) paradigms ("AI-First": 57, "Human-First": 96). Each gray transparent scatter represents one measured sample. All presented bars or data points of a plot in this figure indicate the mean values of metrics measured. All error bars indicate the standard error of the mean. *P*-values are calculated with post-hoc marginal comparison after an initial analysis with linear mixed model. "***" means $p<0.001$, "**" means $p<0.01$, "*" means $p<0.05$, "NS" means no statistical significance.

https://doi.org/10.1038/s41591-026-04553-w



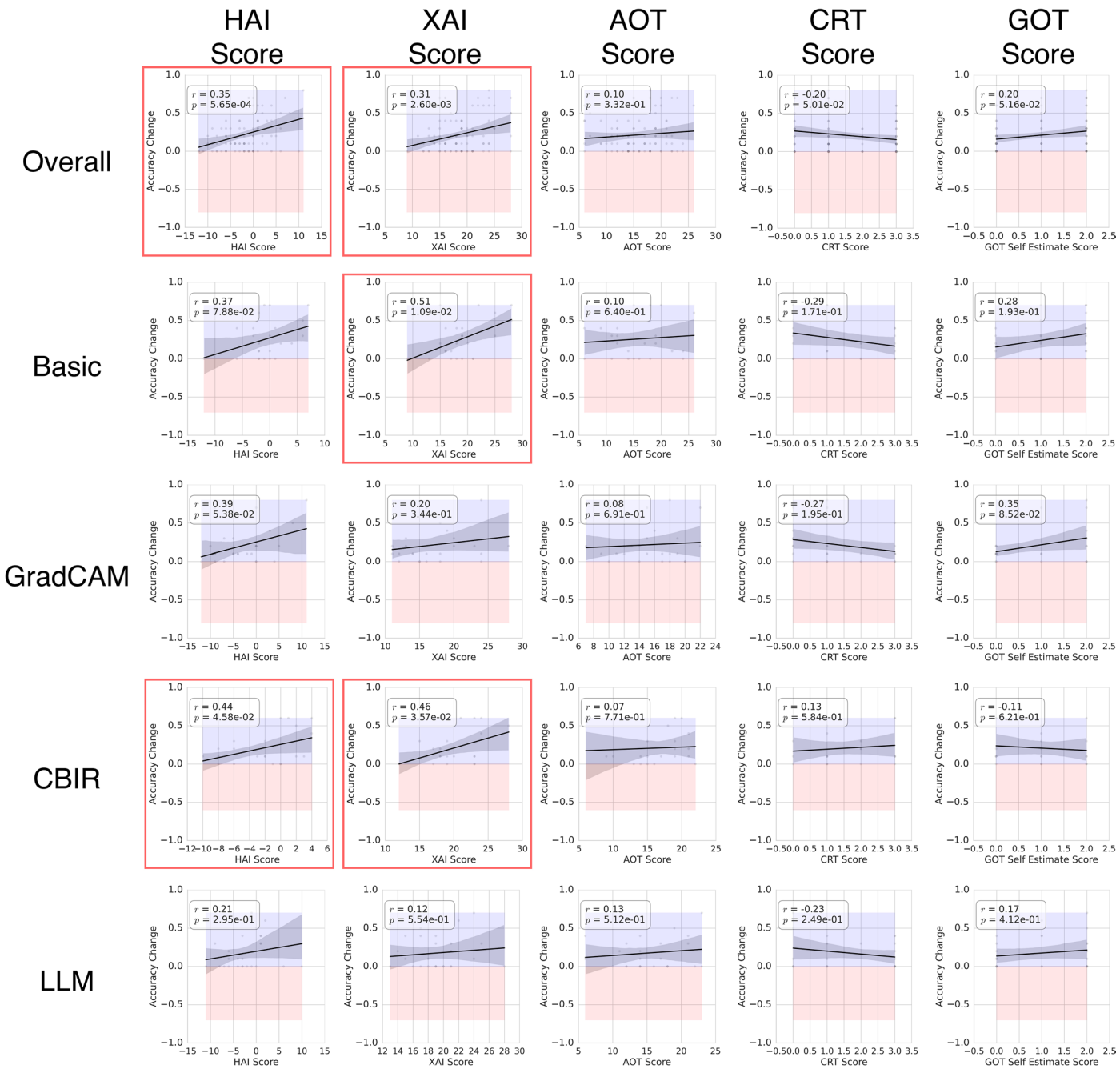
Extended Data Fig. 5 | Medical Student in Human-AI Collaboration. **a**, Accuracy in identifying atopic dermatitis, pityriasis rosea, Lyme disease, CTCL, and the overall performance on four main diseases, with results presented for two rounds (N=216). **b**, Confidence levels of participants in their diagnostic decisions for tasks in (**a**). **c**, Diagnostic accuracy improvement from the first diagnosis round without AI assistance to the second diagnosis round with AI assistance across different AI explanation (XAI) (Basic (*N*=59), GradCAM (*N*=52), CBIR (*N*=57), and LLM (*N*=48)). **d, e**, Changes in diagnostic accuracy between two diagnosis rounds under high XAI quality (**d**) and low XAI quality (**e**) when AI is right, comparing GradCAM, CBIR, and LLM explanations (High XAI quality, AI right: GradCAM (N=16), CBIR (N=57), LLM (N=48); AI wrong: GradCAM (N=6), CBIR (N=22), LLM (N=29). Low XAI quality, AI right: GradCAM (N=52), CBIR (N=41), LLM (N=48); AI wrong: GradCAM (N=52), CBIR (N=55), LLM (N=37)). **f, h** Diagnostic accuracy changes under AI-right (**f**) and AI-wrong (**h**) predictions for different XAI approaches. **g**, Overall accuracy changes under AI-right and AI-wrong conditions, integrating the effects of explanation quality and correctness. **i, j**, Visualization of diagnostic accuracy changes under AI-wrong conditions, highlighting the effects of good XAI quality (**i**) and low XAI quality (**j**). **k**-**l**, Performance comparison between deferential and non-deferential participants (Deferential N=195, Non-deferential N=21). (**k**) Distribution of correct responses with AI assistance, (**l**) accuracy trajectories across two decision rounds. **m**, Proportion of fully deferential participants across different XAI approaches and human-AI collaboration (HAI) paradigms ("AI-First": 104, "Human-First": 216). Each gray transparent scatter represents one measured sample. All presented bars or data points of a plot in this figure indicate the mean values of metrics measured. All error bars indicate the standard error of the mean. *P*-values are calculated with post-hoc marginal comparison after an initial analysis with linear mixed model. "***" means $p<0.001$, "**" means $p<0.01$, "*" means $p<0.05$, "NS" means no statistical significance.

https://doi.org/10.1038/s41591-026-04553-w



Extended Data Fig. 6 | Accuracy-Confidence Consistency of Medical Students and Top3 Diagnosis of PCPs. a-c, Medical students' performance comparison between LLM and non-LLM assisted systems (LLM: N=48, Non-LLM: N=168, No AI: N=216), in (a) overall diagnostic accuracy-confidence consistency (A-C consistency), (**b**) A-C consistency with correct AI predictions, and (**c**) A-C consistency with wrong AI predictions. **d-f**, Equivalent results for top3 diagnosis of PCP participants (LLM: N=26, Non-LLM: N=70, No AI: N=96) as (**a**)-(**c**). Each transparent scatter represents one measured sample. Statistical significance was assessed using Pearson correlation coefficients (r) with associated p-values. Data is presented as predicted values by the linear regression model with accuracy as the dependent variable and confidence as the independent variable +/- 95% confidence intervals.

https://doi.org/10.1038/s41591-026-04553-w



Extended Data Fig. 7 | Correlation Between Accuracy Change with AI Assistance And Personality Test Score. In each subfigure, the blue band (maximum value) and red band (minimum value) indicate the range of accuracy change. The red border around the subfigure represents a significant correlation found ($p<0.05$). Each gray transparent scatter represents one measured sample. Statistical significance was assessed using Pearson correlation coefficients (r) with associated p-values. Data is presented as predicted values by the linear regression model with accuracy as the dependent variable and confidence as the independent variable +/- 95% confidence intervals.

nature portfolio

Corresponding author(s): Xuhai Xu, Marzyeh Ghassemi

Last updated by author(s): May 12, 2026

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	For the data collection for ML model training, we merged 16 datasets (15 public access, 1 private access, see the list in Stable. 4). Using these datasets, we trained ML models as described in the Method section. As for the data collection for the human-AI collaboration study, we adopted Qualtrics as the platform for both human-subject studies.
Data analysis	We used python (specifically PyTorch 1.13) to write custom code to train ML models. For the human subject study results, we used python and Jupyter Notebook for data analysis.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

For ML model training, our datasets are organized and available in this link. Its description is available in the README file in the Figshare subfolder of ml-model (https://figshare.com/s/eaf2110da509abbe9a6a). For the two human-subject studies, the images used in the human experiment and the study results are organized and available in the Figshare subfolder of human-study-results-analysis (https://figshare.com/s/eaf2110da509abbe9a6a).

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Study 1 included 623 participants from the general public, consisting of 314 females, 305 males, and 4 who identified as non-binary or other. Study 2 involved 153 primary care physicians (PCPs), of whom 66 were female, 84 were male, and 3 identified as non-binary or other. An additional cohort of 320 medical students was recruited, which included 171 females, 141 males, and 8 who identified as non-binary or other. While these data were collected for demographic purposes, sex- and gender-based analyses were not the focus of this study, and therefore were not performed. The primary analyses centered on the effects of medical expertise, XAI methods, and human-AI decision paradigms on diagnostic performance.
Reporting on race, ethnicity, or other socially relevant groupings	Participant race and ethnicity were collected via self-report. The distribution of participants is as follows: - Study 1 (General Public, N=623): White (402), American Indian or Alaska Native (61), Black or African American (45), Asian (35), Hispanic or Latino or Spanish Origin (26), Native Hawaiian or Other Pacific Islander (14), Middle Eastern or North African (4), and other (10). 18 participants identified as multi-racial, and 8 preferred not to answer. - Study 2 (PCPs, N=153): Asian (50), White (46), Black or African American (19), Middle Eastern or North African (14), Hispanic or Latino or Spanish Origin (7), American Indian or Alaska Native (3), and other (5). 6 participants identified as multi-racial, and 3 preferred not to answer. - Study 2 (Medical Students, N=320): Asian (129), Black or African American (70), White (46), Middle Eastern or North African (25), Hispanic or Latino or Spanish Origin (9), American Indian or Alaska Native (1), and other (11). 17 participants identified as multi-racial, and 12 preferred not to answer. The primary socially relevant variable analyzed in this study was patient skin tone within the clinical images (not the human subjects), used to investigate diagnostic disparities and the fairness of AI assistance. Human subjects' info was collected solely to describe cohort diversity and to assess potential performance equity; they were not used as proxies for socioeconomic status or other constructs.
Population characteristics	See above.
Recruitment	Participants in Study 1 (general public) were recruited from the online platform Prolific. Participants in Study 2 (PCPs and medical students) were recruited from several clinical networks and platforms, including local networks at Boston Medical Center and Stanford University, as well as Centaur Labs, MedShr, Pathway, and XPC. This recruitment strategy may introduce a self-selection bias. Participants from online platforms like Prolific may be more digitally literate than the general population. Similarly, clinicians who opted into the study through professional networks may have a pre-existing interest in artificial intelligence or digital health technologies. This potential bias could mean that the observed effects of AI collaboration might be more pronounced than in a broader, less technologically inclined population.
Ethics oversight	The study protocol was reviewed and approved by the Massachusetts Institute of Technology's (MIT) Committee on the Use of Humans as Experimental Subjects as Exempt Category 3—Benign Behavioral Intervention. All participants were informed about the study's procedures and the nature of the visual content, and all provided informed consent before beginning the experiment. Participants were explicitly informed that they could withdraw from the study at any time.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	This was a quantitative, experimental study. It consisted of two large-scale, digital experiments that used a between-subject factorial design (4 XAI methods x 2 decision paradigms) to assess human-AI collaborative diagnostic performance. Data collected included quantitative measures of diagnostic accuracy (binary and free-text), decision confidence ratings, and responses to validated psychometric questionnaires.
Research sample	The study involved two primary research samples chosen to represent different levels of medical expertise: Study 1: 623 participants from the general public with no medical background; Study 2: 153 PCPs and an additional cohort of 320 medical students. The demographic details of these cohorts can be found in the materials above. The samples were collected across diverse populations from online platforms and clinical networks. The rationale for choosing these distinct groups was to systematically investigate how diagnostic performance with AI assistance varies across different levels of expertise.
Sampling strategy	Recruitment was convenience sampling: Prolific's pre-screened panels and professional e-mail/list-serv networks for clinicians and students. For Study 1, an a priori power analysis was conducted to determine a sufficient sample size for detecting differences in diagnostic accuracy across the four XAI explanation types (Basic, GradCAM, CBIR, LLM). The analysis was based on a one-way, between-subjects ANOVA with four groups. The significance level was set at $\alpha = 0.05$, with a target power of 0.80. Based on prior literature, we estimated a small to medium effect size, Cohen's $f = 0.15$, to reflect these modest expected differences in a general population. The analysis indicated that a total sample size of at least 492 participants would be required to detect an effect of this magnitude under the specified parameters. Therefore, our final sample of 623 participants was sufficient to detect the hypothesized differences between XAI conditions. We conducted a similar power analysis for Study 2. We estimated Cohen's $f = 0.30$ for PCPs and 0.20 for medical students -- we hypothesized that the utility of different explanation types could lead to more pronounced differences in performance compared to the general public. This leads to a total sample size of at least 128 for PCPs and 280 for students would be required. Our final samples were considered sufficient.
Data collection	All sessions ran in a custom Qualtrics interface. Participants reviewed 12 clinical images of human skin, entered diagnoses and confidence, viewed AI outputs/explanations according to assignment, and repeated the decision once. Responses and timings were recorded automatically; no experimenter was present, and researchers were blind to condition during data collection. Attention-check items and per-image time windows (10 s – 5 min) enforced data quality.
Timing	Study 1 was conducted during March 20 – May 01, 2024; Study 2 was conducted during June 10 - Sept 3, 2024.
Data exclusions	Data were excluded from the final analysis based on pre-established quality control criteria. The study embedded attention check questions, and participants who failed them were filtered out. Additionally, responses with abnormal durations (averaging less than 10 seconds or more than 5 minutes per image) were excluded. In Study 1, 237 participants were filtered and excluded. In Study 2, 48 PCPs and 89 medical students were excluded. Moreover, 289 participants were excluded due to their medical jobs (not PCPs or students).
Non-participation	Participants who didn't complete the Qualtrics survey were not recorded.
Randomization	Participants were randomly allocated to experimental groups. Upon entering the study, each participant was randomly assigned to one of eight conditions, representing a factorial combination of four different XAI methods (Basic, GradCAM, CBIR, or LLM) and two human-AI decision paradigms (Human-First or AI-First). Within each participant, the 12 images were presented in random order, and AI correctness was preset but balanced across skin tones to avoid systematic covariates.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks

Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.

Novel plant genotypes

Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.

Authentication

Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.