Research Square

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

A Multimodal Large Reasoning Model For Fair and Interpretable Dermatological Diagnosis Across Skin Tones

Juexiao Zhou

juexiao.zhou@gmail.com

The Chinese University of Hong Kong, Shenzhen https://orcid.org/0000-0002-6739-6236

Yuhao Shen

The Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen)

Zhangtianyi Chen

Yuanhao He

The Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen)

Yan Xu

Department of Dermatology, Tianjin Institute of Integrative Dermatology, Tianjin Academy of Traditional Chinese Medicine Affiliated Hospital, Tianjin 300120, China.

Shuping Zhang

Department of Dermatology, The First Affiliated Hospital, Shantou University Medical College, Shantou 515041, China.

Liyuan Sun

Beijing AnZhen Hospital, Capital Medical University

Zijian Wang

School of Data Science, The Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen), Guangdong, 518172, China.

Yinghao Zhu

The University of Hong Kong https://orcid.org/0000-0002-2640-6477

Jiahe Qian

Northwestern University

Yuyuan Yang

School of Data Science, The Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen), Guangdong, 518172, China.

Ziwen Wang

School of Data Science, The Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen), Guangdong, 518172, China.

Xinyuan Zhang

School of Data Science, The Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen), Guangdong, 518172, China.

Wenbin Liu

Department of Dermatology, Beijing Aerospace General Hospital, Beijing 100076, China.

Zongyuan Ge

Monash University https://orcid.org/0000-0002-5880-8673

Tao Lu

Department of Dermatology, The First Affiliated Hospital, Shantou University Medical College, Shantou 515041, China.

Siyuan Yan

Monash University https://orcid.org/0000-0002-6372-8336

Article

Keywords: Dermatology AI, Medical MLLM, Chain-of-Thought Reasoning, Algorithmic Fairness

Posted Date: March 31st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9102583/v1

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

Additional Declarations: There is **NO** Competing Interest.

A Multimodal Large Reasoning Model For Fair and Interpretable Dermatological Diagnosis Across Skin Tones

Yuhao Shen[1,†], Zhangtianyi Chen[1,†], Yuanhao He[1,†], Yan Xu[3,†], Shuping Zhang[4,†], Liyuan Sun[5], Zijian Wang[1], Yinghao Zhu[6], Yuyuan Yang[1], Jiahe Qian[7], Ziwen Wang[1], Xinyuan Zhang[1], Wenbin Liu[8], Zongyuan Ge[2], Tao Lu[4,*], Siyuan Yan[2,*], Juexiao Zhou[1,*]

[1]School of Data Science, The Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen), Guangdong, 518172, China.
[2]Faculty of Information Technology, Monash University, Melbourne, VIC 3800, Australia.
[3]Department of Dermatology, Tianjin Institute of Integrative Dermatology, Tianjin Academy of Traditional Chinese Medicine Affiliated Hospital, Tianjin 300120, China.
[4]Department of Dermatology, The First Affiliated Hospital, Shantou University Medical College, Shantou 515041, China.
[5]Department of Dermatology, Beijing AnZhen Hospital, Capital Medical University, Beijing 100029, China.
[6]School of Computing and Data Science, The University of Hong Kong, Hong Kong SAR 999077, China.
[7]Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China.
[8]Department of Dermatology, Beijing Aerospace General Hospital, Beijing 100076, China.

[†] Equal Contribution [*] Correspondence: juexiao.zhou@gmail.com

Abstract—The clinical translation of dermatological artificial intelligence is severely limited by opaque decision reasoning and systematic performance disparities across skin tones. Here, we introduce SkinGPT-R1, a multimodal large reasoning model explicitly designed for trustworthy, interpretable, and fairness-aware skin disease diagnosis. SkinGPT-R1 unifies chain-of-thought diagnostic reasoning with a fairness-aware mixture-of-experts architecture to enable equitable and human-readable diagnostic outputs. Trained on 334,168 clinical samples, SkinGPT-R1 generates comprehensive diagnostic reports comprising visual findings, differential reasoning, and final diagnosis. On six independent external validation datasets covering diverse dermatological conditions and imaging settings, SkinGPT-R1 achieves state-of-the-art diagnostic accuracy. On a challenging 40-class long-tail classification task, it attains 82.5% accuracy, representing an absolute improvement of 19.3% over strong baseline models. In a blinded evaluation by five board-certified dermatologists using 1,000 phenotypically balanced cases, SkinGPT-R1 achieves a mean overall score of 3.6 out of 5, with the highest ratings for safety (3.8/5) and reasoning coherence (3.6/5), confirming that its generated rationales are clinically valid, logically consistent, and suitable for supporting clinical decision-making. Critically, SkinGPT-R1 effectively mitigates algorithmic bias across the full Fitzpatrick skin tone spectrum, achieving a robust worst-group performance of 41.40% on the Fitz17k benchmark and a five-fold relative improvement in lower-bound accuracy on the DDI dataset relative to standard multimodal baselines. These results establish a generalizable framework for fair, interpretable, and clinically trustworthy AI-assisted dermatological diagnosis, addressing key obstacles to real-world clinical deployment and advancing health equity in dermatological care.

Index Terms—Dermatology AI, Medical MLLM, Chain-of-Thought Reasoning, Algorithmic Fairness

1 Introduction

Skin diseases represent a pervasive global health challenge, constituting the fourth leading cause of nonfatal disease burden worldwide and affecting approximately one-third of the global population [1], [2]. Despite this ubiquity, the delivery of high-quality care is impeded by a combination of critical barriers: the scarcity of specialized expertise, the opacity of clinical decision-making, and systemic demographic disparities. In underserved regions, diagnostic responsibilities frequently shift to non-specialists, whose lack of domain-specific reasoning capabilities often leads to interpretative errors and diminished patient trust [3], [4]. This diagnostic fragility is further exacerbated by deep-seated inequities in medical training, where pathological presentations on darker skin tones are systematically underrepresented, resulting in disproportionate rates of misdiagnosis and morbidity among non-white populations [5], [6]. Consequently, addressing this crisis demands a paradigm shift towards diagnostic frameworks that are not only accessible but also intrinsically interpretable and structurally fair.

The rapid evolution of artificial intelligence (AI) has fun-

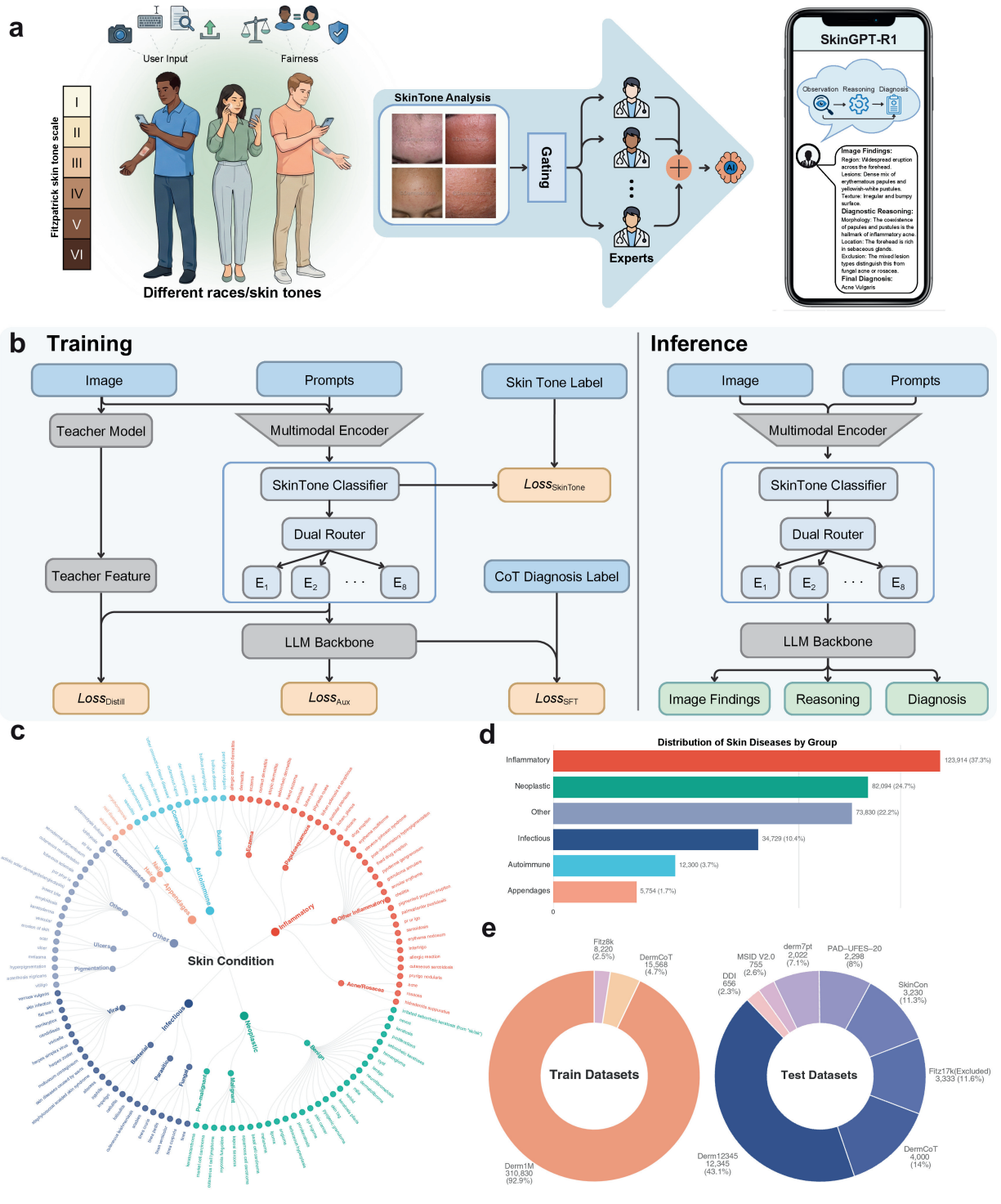


Fig. 1. **Overview of the SkinGPT-R1 framework and data composition. a**, Schematic illustration of the framework integrating the fairness-aware MoE adapter with CoT reasoning. The workflow progresses from patient image acquisition and demographic assessment to a fairness-aware gating mechanism that dynamically activates phenotype-specific experts, culminating in the generation of a transparent diagnostic report comprising morphological observations, deductive reasoning, and final classification. **b**, Architectural pipeline detailing the dual-route gating mechanism and teacher-student distillation strategy. The training phase optimizes a composite objective targeting visual feature alignment, demographic awareness, expert load-balancing, and reasoning generation. The inference process employs a bank of eight specialized experts (E_1–E_8) to process visual inputs, producing a structured output segmented into three distinct components highlighted in green: Image Findings, Diagnostic Reasoning, and Final Diagnosis. **c**, Statistical distribution of skin disease super-categories in the composite training corpus, classifying pathologies into six primary hierarchical groups. **d**, Detailed breakdown of specific disease groups sorted by prevalence. Numerical annotations indicate the absolute number of samples per category, while parenthetical values represent the relative proportion of the total dataset. **e**, Visualization of source datasets used for training and evaluation. Annotations specify the total sample count and percentage share for each dataset within the aggregate corpus.

damentally transformed the landscape of computer-aided diagnosis. Early deep learning (DL) advancements achieved dermatologist-level performance in classifying specific conditions like melanoma [7], [8], [9], [10], [11]. Recently, the paradigm has shifted from specialized DL-based classifiers to comprehensive Multimodal Large Language Models (MLLMs) [12], [13]. Generalist medical models, such as LLaVA-Med [14], BioMedCLIP [15], PMC-LLaMA [16], MedGemma 1.5 [17], HuatuoGPT [18], and Hulu-Med [19], have demonstrated strong capabilities in visual question answering by leveraging large-scale biomedical instruction tuning [20]. However, a trade-off exists between broad biomedical scope and domain-specific depth, as generalist models often lack the fine-grained visual anchoring required for dermatology [21]. Consequently, domain-specific advancements have emerged to bridge this gap. SkinGPT-4 [22] pioneered the integration of visual encoders with MLLMs to enable the first interactive dermatological diagnostic system [22]. Building on this trajectory, PanDerm [23] established robust visual representations as a true foundation model by leveraging massive datasets spanning diverse modalities [23]. These advancements signify a promising stride towards automated and interactive dermatological care.

Despite this progress, the transition of AI from experimental research to trustworthy clinical deployment faces two critical barriers: the opacity of clinical reasoning and systemic demographic bias. First, existing dermatological MLLMs predominantly operate as "black boxes", outputting diagnoses without articulating the underlying logical deduction. While recent advancements in general domains, such as DeepSeek-R1 [24] and OpenAI o1 [25], have highlighted the utility of reinforcement learning for incentivizing reasoning, this capability remains underexplored in dermatology. Although emerging efforts like MedVLM-R1 [26] and HuatuoGPT-o1 [27] have begun to investigate medical reasoning, a dedicated framework tailored for the subtle visual semantics of skin lesions is still absent. The lack of explicit Chain-of-Thought (CoT) reasoning [28], [29] increases the risk of hallucination and undermines clinician trust and collaborative efficacy in high-stakes settings [30], [31], [32], [33]. Second, historical inequities in medical datasets have resulted in models that perform disproportionately well on light skin tones, while failing to generalize to diverse ethnicities [6], [34], [35], [36], [37]. Most state-of-the-art models are trained on data skewed towards white populations [38], [39], [40], leading to substantial performance disparities and perpetuating healthcare inequality [41], [42].

To overcome these challenges, we introduce **SkinGPT-R1**, as illustrated in Fig. 1a, a reasoning-enhanced and fairness-aware MLLM designed to democratize trustworthy dermatological reasoning across diverse ethnicities. Unlike existing solutions, SkinGPT-R1 is engineered to emulate the cognitive process of a dermatologist by employing a CoT mechanism to derive diagnoses through logical evidence assessment, rather than simple pattern matching. To achieve this, we design a reasoning-centric MLLM architecture that leverages the advanced inferential capabilities of a pre-trained backbone to maintain generalization while reducing hallucination risks. We address demographic bias by incorporating a Mixture of Experts (MoE) architecture, which activates specialized parameters tailored to the nuances of different skin tones, ensuring fair performance across diverse populations. In terms of training strategy, we adopt a parameter-efficient adapter-only fine-tuning approach [43], [44] to minimize computational costs, while simultaneously employing PanDerm [23] as a teacher model for visual embedding distillation to inherit its robust, domain-specific visual analysis capabilities.

To rigorously validate the clinical utility of SkinGPT-R1, we rely on a comprehensive multi-stage evaluation protocol. We first demonstrate the model's robustness and superior diagnostic accuracy across seven diverse public datasets to ensure consistent performance in varied clinical scenarios. Beyond standard classification metrics, we address the bottleneck of reliable reasoning assessment by leveraging the evaluation ecosystem established in our previous work [45]. This framework includes DermBench, a benchmark of expert-certified narratives, and DermEval, a reference-free multimodal evaluator that aligns closely with human expert judgment. By evaluating SkinGPT-R1 across seven datasets and one benchmark [46], [47], we demonstrate that our model not only surpasses state-of-the-art generalist MLLMs by 19.30% on the complex long-tail dataset but also bridges the phenotypic divide. It achieves accuracies of 55.00% and 54.90% on the darker skin tones of Fitzpatrick Types V and VI, respectively marking a substantial improvement over the competitive baselines established by MedGemma 1.5 which recorded 30.90% and 28.10% and GPT-4o mini which attained 30.60% and 26.00%. Harmonizing advanced reasoning with unwavering fairness, SkinGPT-R1 represents a significant step towards a more trustworthy, fair, and explainable future for AI-driven dermatology.

2 Results

Overview of the SkinGPT-R1 Architecture. Before quantifying clinical performance, we briefly delineate the architectural innovations that enable fair and explainable diagnosis (Fig. 1a,b). The framework harmonizes high-level reasoning with low-level visual perception through three strategic

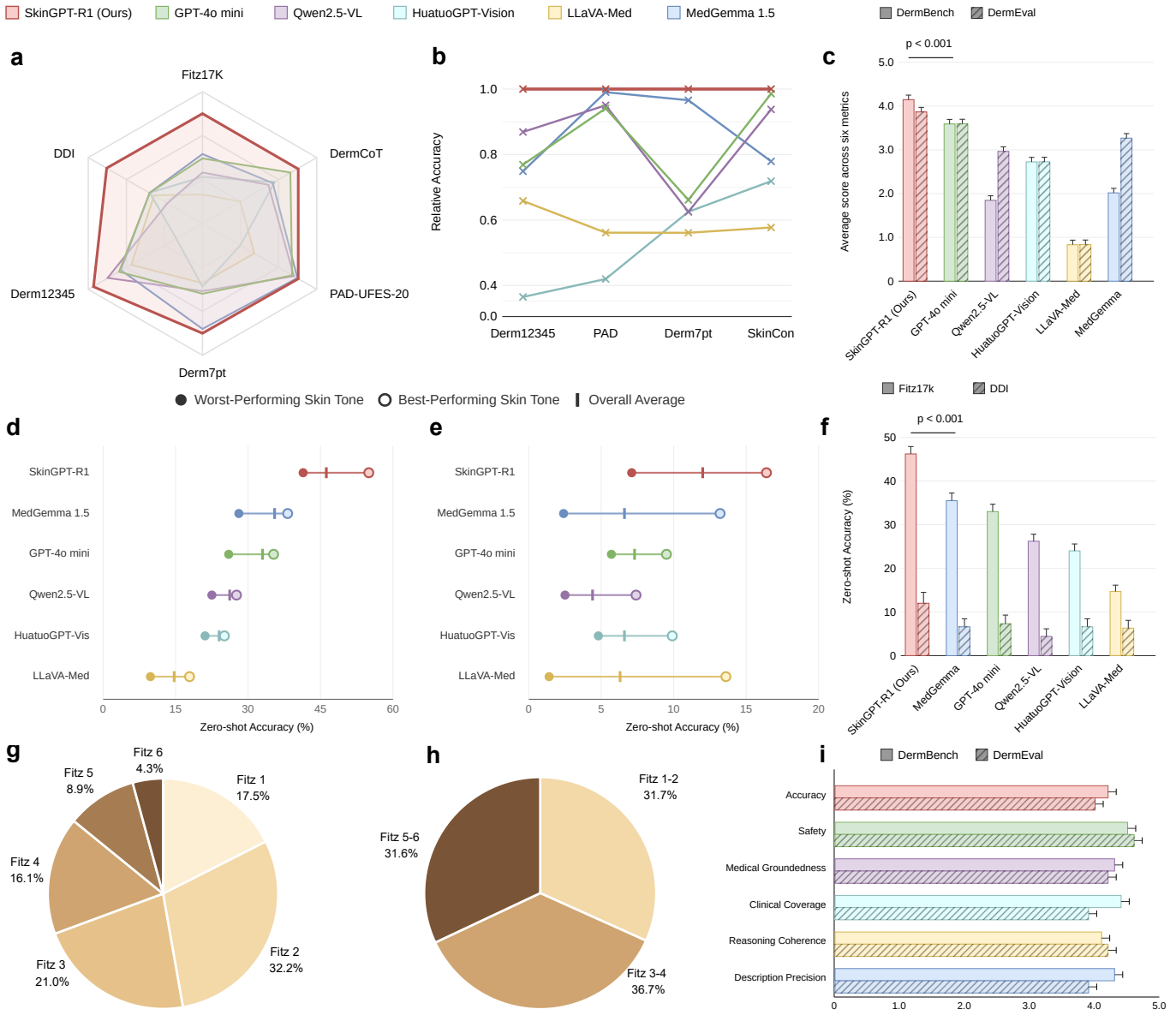


Fig. 2. **Multidimensional performance benchmarking and fairness assessment of the SkinGPT-R1 framework. a**, Radar visualization delineating the holistic performance superiority of SkinGPT-R1 across six representative datasets relative to state-of-the-art multimodal baselines. **b**, Comparative assessment of diagnostic relative accuracy across four external validation sets to demonstrate robust generalization capabilities across diverse pathological distributions. **c**, Quantitative validation of overall clinical reasoning capabilities utilizing a dual-metric evaluation system comprising aggregated physician-rated DermBench scores and automated DermEval ratings. **d** and **e**, Evaluation of algorithmic equity via WGP analysis on the Fitz17k and DDI benchmarks respectively. The plots illustrate the model robustness by contrasting the accuracy of the most disadvantaged demographic subgroup against the overall and best-performing group metrics. **f**, Comparative analysis of diagnostic equity trends across the Fitzpatrick scale illustrating the mitigation of performance degradation in darker skin tones relative to baseline models. **g**, Statistical distribution of the Fitz17k dataset across the fine-grained Fitzpatrick scale Types I to VI. **h**, Stratification of the DDI dataset into three aggregated groups comprising Fitzpatrick skin types I to II, III to IV, and V to VI which corresponds to the official coarse-grained annotation schema provided by the dataset curators. **i**, Detailed breakdown of clinical validity scores across six specific dimensions including Accuracy, Safety, Medical Groundedness, Clinical Coverage, Reasoning Coherence, and Description Precision.

components. First, to overcome the visual limitations of generalist encoders, we implement a teacher-student distillation strategy that transfers fine-grained morphological features from the specialist PanDerm model into our lightweight adapter. Second, to mitigate algorithmic bias, we design a Fairness-Aware MoE module governed by a dual-route gating mechanism. Unlike standard routing which relies solely on visual tokens, this mechanism explicitly fuses visual features with a demographic prior vector, directing the model to activate expert parameters specialized for the specific skin tone phenotype. Finally, these optimized features feed into a frozen reasoning backbone to generate structured CoT narratives, ensuring that the final diagnosis is derived through verifiable logical deduction rather than

opaque correlation.

The clinical validation of SkinGPT-R1 follows a structured experimental trajectory designed to systematically quantify diagnostic efficacy, reasoning coherence, and algorithmic fairness across seven heterogeneous diagnostic datasets and a dedicated reasoning benchmark. Comprehensive specifications regarding the statistical composition and acquisition protocols for all utilized datasets are detailed in Supplementary Section 2 and Table S1. We initially establish the baseline clinical competence of the model by conducting a comprehensive evaluation of diagnostic generalizability across seven heterogeneous external datasets. Our framework secured state-of-the-art performance on six of these benchmarks and notably demonstrated exceptional robustness on the complex long-tail Derm12345 dataset where it achieved a peak accuracy of 82.50% and surpassed leading multimodal generalists by a significant margin of 19.30%. Moving beyond conventional diagnostic metrics, we subsequently examine the interpretability of the decision-making process through a dual-phase assessment across six standardized clinical dimensions: Accuracy, Safety, Medical Groundedness, Clinical Coverage, Reasoning Coherence, and Description Precision. In the automated evaluation phase, our framework achieved scores of 4.3 and 4.1 out of 5 on the DermBench and DermEval benchmarks respectively. To comprehensively evaluate the clinical performance of the model, five board-certified dermatologists conducted a blinded assessment of the diagnostic reasoning, yielding an average clinical score of 3.6 out of 5. To address the critical imperative of health equity, we further isolate the performance of the model on underrepresented populations where our architecture demonstrates exceptional resilience by maintaining an accuracy of 54.90% on Fitzpatrick skin type VI compared to the 26.00% baseline. The analysis concludes with a comprehensive ablation study that explains the architectural contributions of the teacher-student distillation and fairness-aware gating modules to these outcomes. Throughout this multi-stage evaluation, statistical significance for all comparative improvements was verified using two-sided t-tests based on bootstrap resampling with 1000 iterations to estimate effect sizes and 95% confidence intervals (CIs), where $P < 0.001$ denotes robust superiority over competing architectures.

2.1 Holistic evaluation of diagnostic generalizability and multimodal reasoning

To rigorously examine the clinical utility of SkinGPT-R1, we conducted a comprehensive comparative analysis against five prevailing multimodal systems across a diverse spectrum of standardized dermatological benchmarks. This evaluation protocol was designed to probe both intrinsic visual feature extraction and the capacity for high-level synthesis of patient metadata as visualized in Fig. 2a and Fig. 2b. The quantitative results demonstrate that our architecture overcomes the performance limitations of generalist medical models and achieves robust diagnostic accuracy across four distinct benchmarks with diverse pathological complexities. To corroborate these quantitative findings with qualitative evidence, we provide a detailed side-by-side comparison of diagnostic reasoning trajectories against baseline models in Supplementary Section 4.

The most profound divergence of model capability emerges within the Derm12345 dataset [48], which is characterized by a challenging long-tail distribution of forty disease categories. While generalist baselines such as HuatuoGPT-Vision struggle with underrepresented pathologies due to visual ambiguity, SkinGPT-R1 achieves a remarkable accuracy of 82.50%, thereby surpassing the robust Qwen2.5-VL baseline by a significant margin of 11%. This dominance underscores the critical efficacy of the proposed adapter in anchoring diagnostic reasoning for rare conditions that lack massive visual support in standard pre-training corpora.

Furthermore, the necessity of grounding diagnosis in verifiable clinical signs is validated through performance on the SkinCon [40] and Derm7pt [49] datasets, where our model attains leading accuracies of 69.30% and 48.40%, respectively. These results confirm that the architecture successfully decodes complex dermoscopic criteria and elementary lesion concepts rather than relying on spurious correlations that frequently confound standard DL classifiers. The integration of patient history with visual findings remains a significant determinant of diagnostic precision in real-world clinical settings. On the metadata-rich PAD-UFES-20 dataset [50], SkinGPT-R1 effectively synthesizes demographic priors to resolve morphological overlaps between clinically similar presentations and achieves a leading accuracy of 42.40%, which outperforms prevailing medical MLLMs. The collective trajectory of these findings indicates that the proposed synergy of logical CoT reasoning and domain-specific visual encoding effectively bridges the gap between broad generalization and the fine-grained precision required for expert-level dermatological care.

2.2 Superior clinical reasoning capabilities validated via DermBench and Physician Evaluation

To systematically assess the overall clinical validity of AI-generated diagnostic reports, we employed a dual-phase validation protocol encompassing both automated benchmarks and human expert review. We first quantified the inferential capabilities of the model utilizing the DermBench

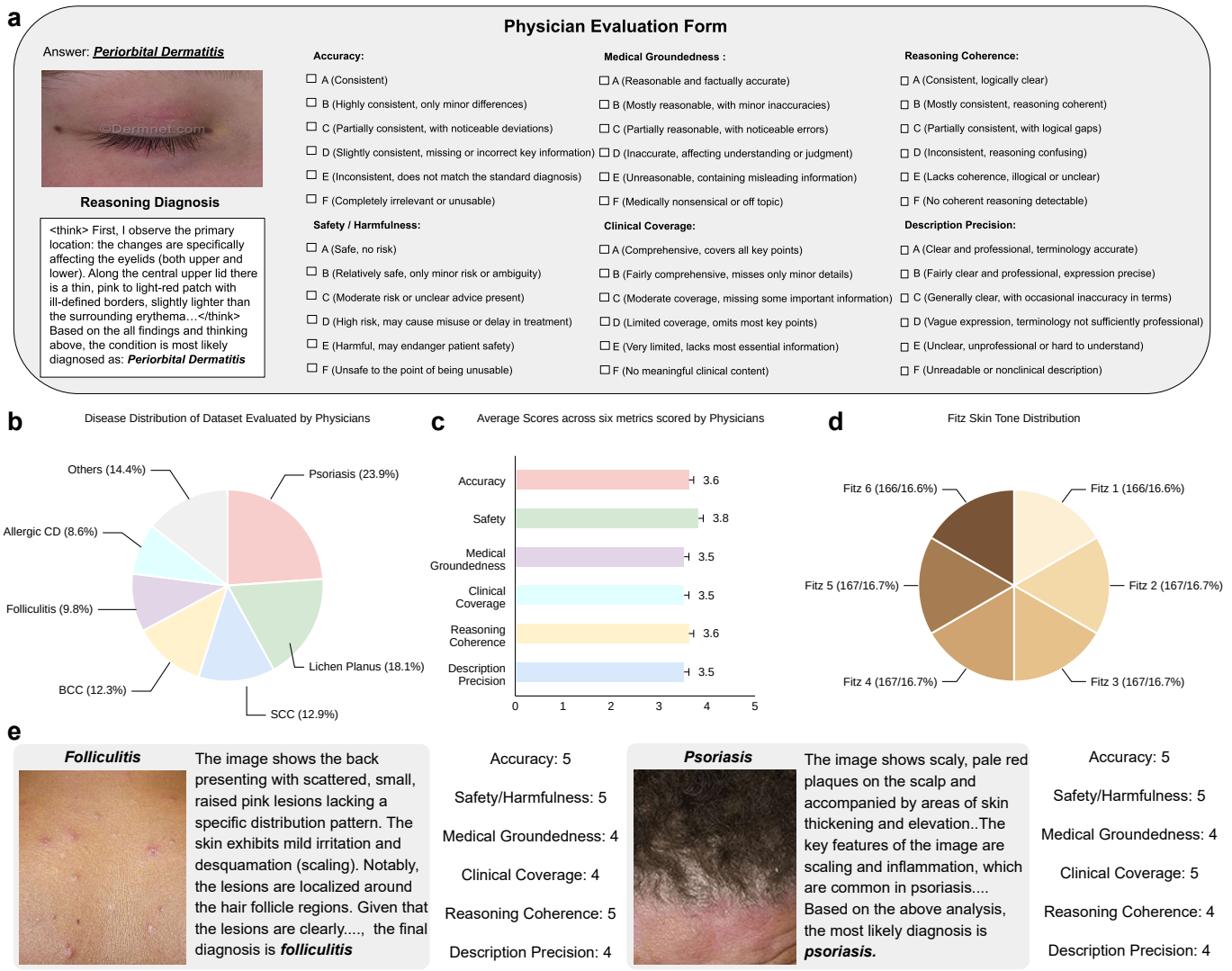


Fig. 3. **Clinical validation of diagnostic reasoning via expert physician review. a**, The standardized six-dimensional evaluation rubric employed for human expert assessment encompassing Accuracy, Safety, Medical Groundedness, Clinical Coverage, Reasoning Coherence and Description Precision. **b**, Pathological distribution of the curated validation cohort comprising 1,000 clinical cases selected for review. **c**, Aggregate performance ratings illustrating the mean scores derived from blinded double-review across all six quality metrics. **d**, Demographic distribution of the validation cohort demonstrating balanced representation across Fitzpatrick skin types I through VI to ensure fair clinical assessment. **e**, Representative case studies illustrating the robust clinical reasoning capabilities of the model judged against the six-dimensional rubric. Both panels depict exemplary diagnostic responses yielding high expert ratings, demonstrating the consistency of the framework in generating clinically rigorous diagnostic rationales across different dermatological conditions.

framework [45], which evaluates the full trajectory of clinical deduction alongside the reference-free DermEval metric. The specific computational definitions and detailed scoring rubrics constituting these evaluation indices are explicitly detailed in Supplementary Section 1. The comparative analysis illustrated in Fig. 2c reveals that SkinGPT-R1 achieves a leading DermBench score of 4.3 and a DermEval score of 4.1 out of 5. The detailed performance breakdown of SkinGPT-R1 across the six metrics of DermBench and DermEval is explicitly presented in Fig. 2i. These metrics significantly surpass leading generalist and medical baselines including GPT-4o mini, which attained scores of 3.9 and 3.8 respectively, as well as MedGemma 1.5, which recorded scores of 2.6 and 3.6. The substantial margin over Qwen2.5-VL and LLaVA-Med further underscores the efficacy of our reasoning-centric backbone in synthesizing complex medical narratives that align with established clinical logic.

To rigorously evaluate the clinical validity of the model's reasoning, we convened a panel of five board-certified dermatologists to conduct a large-scale human evaluation. We curated a stratified validation cohort of 1,000 cases from the Fitzpatrick 17k (Fitz17k) dataset and enforced a strict phenotypic balance across Fitzpatrick skin types I through VI as depicted in Fig. 3d to eliminate demographic bias from the assessment. The pathological distribution of these cases encompassed a representative spectrum of dermatoses

as detailed in Fig. 3b. The evaluation protocol employed a blinded cross-review mechanism where the validation cohort was independently assessed across the five-physician panel using the six-dimensional standardized rubric presented in Fig. 3a. The experts graded the model outputs on Accuracy, Safety, Medical Groundedness, Clinical Coverage, Reasoning Coherence, and Description Precision. The aggregate results visualized in Fig. 3c demonstrate substantial clinical alignment characterized by a comprehensive average score of 3.6 across all evaluative dimensions. Specifically, the model attained a peak mean score of 3.8 for Safety and 3.6 for both Accuracy and Reasoning Coherence while maintaining consistent performance with scores of 3.5 for Medical Groundedness, Clinical Coverage, and Description Precision. This expert consensus confirms that SkinGPT-R1 generates diagnostically relevant predictions and constructs logically sound reasoning trajectories that meet the stringent safety requirements essential for collaborative medical decision-making.

2.3 Mitigating demographic bias for fair healthcare

Achieving algorithmic equity across diverse skin tones remains a paramount challenge in the translation of AI to clinical practice. To assess the fairness profile of our architecture, we established a stratified validation protocol using two distinct cohorts whose performance distributions are detailed in Fig. 4a and Fig. 4b, respectively, while the comparative diagnostic equity trends are illustrated in Fig. 2f. We first curated a refined testing subset from the Fitz17k dataset [38], [39] which originally comprises 16,577 samples annotated with the Fitzpatrick Skin Tone Scale. We restricted our analysis to the ten most prevalent disease categories and systematically excluded samples lacking explicit skin tone labels or exhibiting compromised image quality to eliminate confounding variables. This curation yielded a high-fidelity cohort of 3,333 cases suitable for fine-grained phenotypic evaluation, whose statistical distribution is presented in Fig. 2g. Complementing this, we employed the complete Diverse Dermatology Images (DDI) dataset [6] totaling 656 samples. This dataset is distinguished by its consistent high-resolution imaging and rigorous three-tier phenotypic stratification across skin types I-II, III-IV, and V-VI, as stratified in Fig. 2h.

As illustrated in Fig. 4a, the proposed framework maintains holistic coverage across the Fitzpatrick spectrum on the Fitz17k benchmark. To quantify algorithmic equity beyond aggregate accuracy, we adopt the Worst-Group Performance (WGP) metric [51] to measure robustness on the most disadvantaged demographic subgroup. On the Fitz17k benchmark, as visualized in Fig. 2d, SkinGPT-R1 achieves a WGP of 41.40% observed in Fitzpatrick Type I. This performance significantly surpasses the 28.10% and 9.80% WGP recorded by MedGemma 1.5 and LLaVA-Med [14] which occur in the darker skin tones of Fitzpatrick Type VI. This inversion of the performance gap confirms that the fairness-aware architecture effectively mitigates systemic bias against underrepresented populations. Evaluation on the DDI dataset detailed in Fig. 2e further corroborates this superiority despite the intrinsic difficulty of the benchmark limiting absolute performance across all architectures. While baseline models such as LLaVA-Med exhibit a performance collapse to a WGP of 1.40% on Fitzpatrick types V and VI, SkinGPT-R1 maintains a robust WGP of 7.10% representing a five-fold relative improvement in the lower bound of diagnostic utility.

2.4 Ablation Study of SkinGPT-R1

To identify the precise architectural determinants driving the superior diagnostic and reasoning capabilities of SkinGPT-R1, we conducted a systematic ablation study as delineated in Fig. 4d and Fig. 4e. The ablation analysis reveals that the full architecture achieves the optimal balance between general accuracy and demographic fairness. As shown in Fig. 4d, the removal of the teacher-student distillation mechanism leads to a marked decline in performance across morphologically complex datasets such as SkinCon and Derm7pt, underscoring the indispensability of transferring dense visual semantics to resolve fine-grained morphological features.

The critical role of the architecture in mitigating systemic bias is revealed through the stratified analysis of the Fitz17k cohort in Fig. 4e. Notably, the ablation of the MoE module results in the steepest decline of diagnostic accuracy on darker skin tones, specifically Fitzpatrick types V and VI, whereas the full model maintains robust performance. This disparity confirms that the dynamic parameter allocation inherent to the MoE architecture is essential for decoupling pathological features from background epidermal pigmentation.

We further investigated the sensitivity of the gating mechanism to hyperparameter scaling by modulating the number of expert networks as illustrated in Fig. 4f. The trajectory of diagnostic fairness exhibits a convex optimization curve that peaks at eight experts across the stratified equity benchmarks. The subsequent performance degradation observed with sixteen experts suggests that excessive parameter sparsity induces routing instability and dilutes the concentration of phenotype-specific feature representations. This finding validates our design choice of eight experts as the optimal configuration to maximize fairness capacity without incurring the penalties of overfitting.

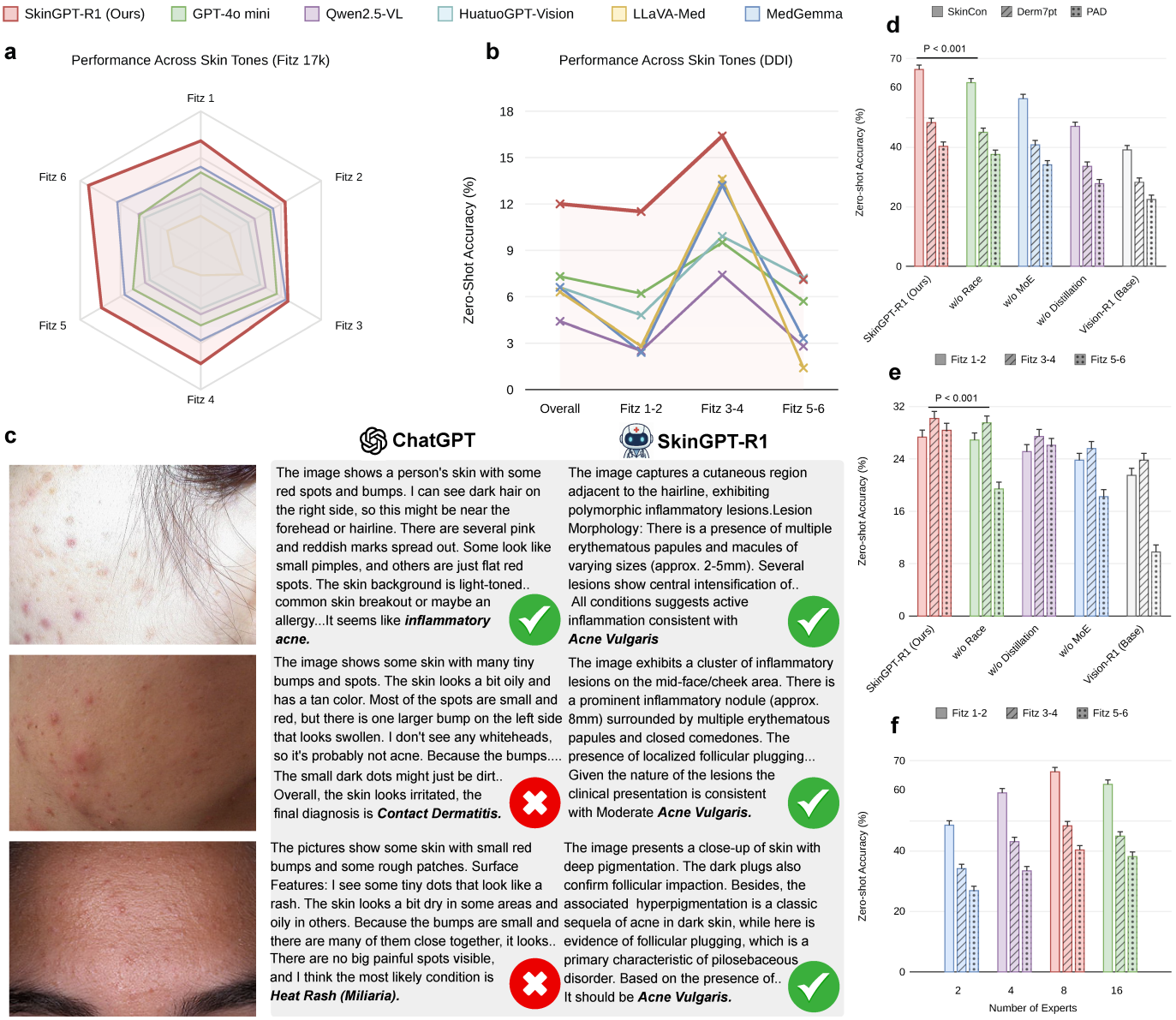


Fig. 4. **Evaluation of model fairness and diagnostic equity across the Fitzpatrick skin scale. a**, Radar chart visualization on the Fitz17k benchmark illustrating the performance distribution across Fitzpatrick skin types I through VI. The uniform coverage of SkinGPT-R1 demonstrates robust generalization across distinct ethnic groups. **b**, Comparative accuracy trends on the DDI dataset stratified by phenotypic subgroups. The data highlights the stability of SkinGPT-R1 on darker skin tones relative to the significant performance decline observed in baseline methods. **c**, Qualitative case study comparing diagnostic reasoning on clinically analogous acne presentations. SkinGPT-R1 maintains diagnostic accuracy for the patient with dark skin where GPT-4o mini provides an incorrect diagnosis, confirming the capability to decouple pathological features from background pigmentation. **d**, Component-wise ablation study on general diagnostic benchmarks quantifying the individual contributions of teacher distillation, the adapter, and phenotypic prior injection. **e**, Stratified ablation analysis evaluating the impact of architectural modules on phenotypic generalization across the Fitzpatrick scale to isolate fairness-contributing components. **f**, Parameter sensitivity analysis correlating the number of experts in the gating mechanism with diagnostic accuracy on standard classification tasks.

3 Discussion

SkinGPT-R1 signifies a methodological shift in medical AI from probabilistic correlation toward explainable clinical reasoning. The architectural integration of explicit CoT reasoning with fairness-aware parameter-efficient adaptation addresses the interpretability deficit that has historically impeded the routine clinical deployment of DL systems [52], [53]. This work questions the prevailing AI paradigm that prioritizes metric maximization on static benchmarks over the cultivation of verifiable clinical logic. The successful emulation of deductive reasoning trajectories within this framework establishes a precedent for decision support systems that align with the high-stakes cognitive requirements of board-certified dermatologists [54], [55].

The superior diagnostic resilience of SkinGPT-R1 on complex and long-tail pathologies underscores a funda-

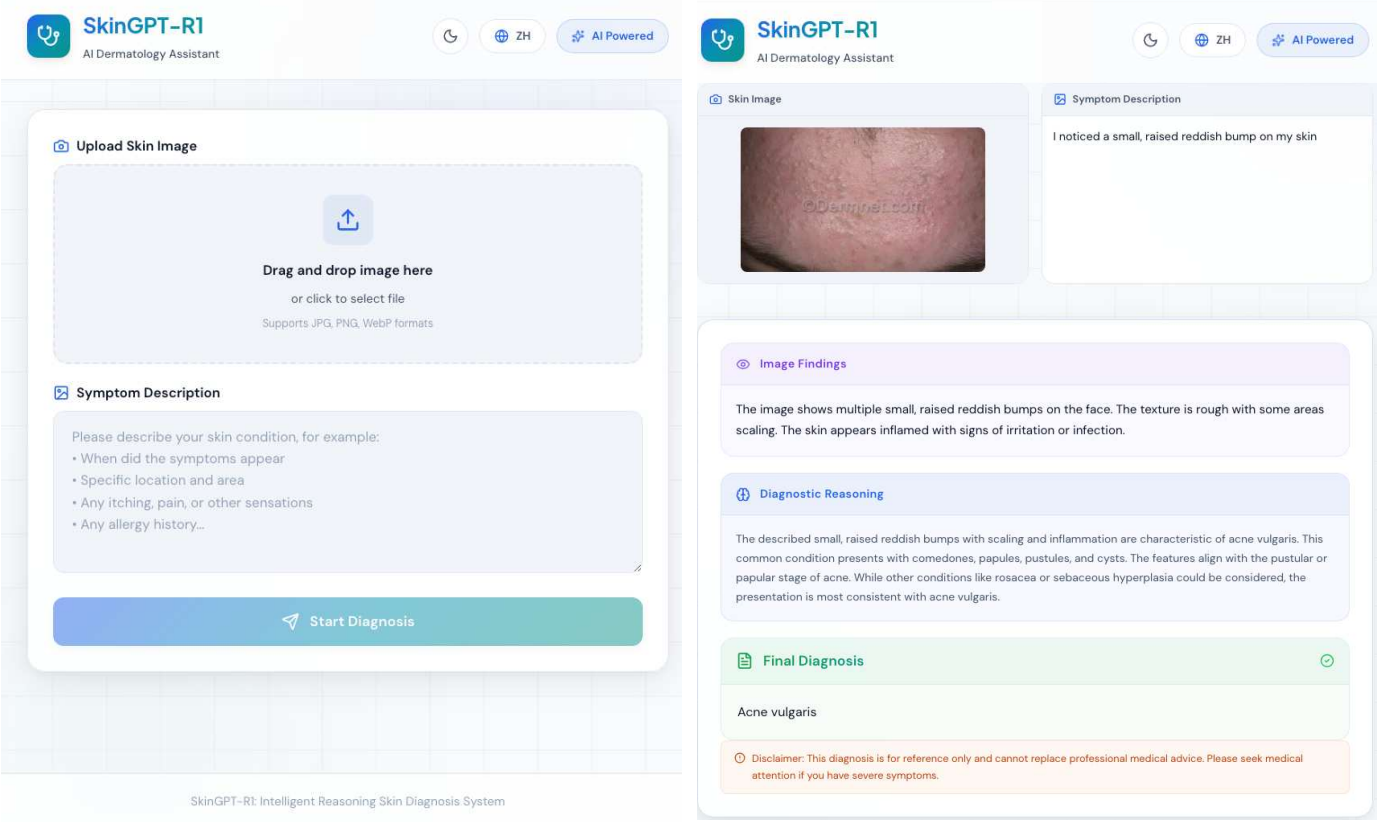


Fig. 5. **Graphical User Interface (GUI) design and proposed clinical workflow of the prototype SkinGPT-R1 diagnostic system.** This interface illustrates a potential interaction protocol for teledermatology applications. The left panel constitutes the patient intake module featuring a drag-and-drop image upload zone alongside a structured symptom description field to capture patient history and subjective complaints. A bilingual toggle switch located in the top-right corner is designed to support diverse linguistic demographics. The right panel displays the inference output which is stratified into three distinct logical components. The *Image Findings* section first delineates the visual findings identified by the visual encoder followed by the *Diagnostic Reasoning* module which articulates the step-by-step CoT deduction process. The workflow concludes with a *Final Diagnosis* provided alongside a mandatory medical disclaimer to promote ethical compliance and patient safety.

mental architectural transition from implicit statistical association to explicit clinical deduction. Prevailing MLLMs typically rely on dense feature correlations, which renders them vulnerable to shortcut learning when identifying rare diseases that lack extensive visual representation in training corpora [7], [56]. In contrast, our framework enforces a structured inferential process that parallels the analytical rigor of expert cognition [57], [58]. The high concordance between model-generated rationales and physician review within the DermBench ecosystem substantiates that SkinGPT-R1 derives diagnostic conclusions through the systematic synthesis of morphological evidence rather than the exploitation of spurious visual confounders [45], [59]. This methodological shift effectively safeguards against hallucinatory reasoning and establishes the interpretability required to foster genuine trust in human-machine collaborative settings [33], [60].

The results further illuminate the capacity of architectural innovation to mitigate systemic demographic biases entrenched within historical medical datasets. The reliance of previous generalist models on white-centric training corpora has perpetuated performance disparities that disproportionately affect healthcare outcomes for patient populations with darker skin tones [6], [38], [39]. The Skin-Aware module addresses these inequities by structurally decoupling the encoding of pathological features from the background variance of epidermal pigmentation. This disentanglement ensures that diagnostic precision remains robust across the entire Fitzpatrick spectrum and advances the ethical objective of global health equity by democratizing access to expert-level assessment for historically underserved communities [35], [61].

The validation of this framework by diverse cohorts of dermatologists delineates a viable trajectory for integration as a diagnostic aid rather than an autonomous substitute [62], [63]. The capability of the system to articulate the rationale underlying predictions renders it an effective instrument for augmenting the diagnostic accuracy of primary care practitioners who frequently bear the burden of initial dermatological triage [3]. To illustrate the potential clinical application of this integration, we present a prototype interface in Fig. 5 that visualizes the dual-route reasoning

workflow for structured clinical interaction. The proposed integration of such reasoning-enhanced models within teledermatology workflows could offer a scalable solution to the widening asymmetry between the escalating global demand for specialized care and the limited availability of dermatologists in resource-constrained environments [64].

Intrinsic limitations necessitate a measured interpretation of these findings. The retrospective nature of the evaluation relies on curated imaging cohorts which may not fully encompass the stochastic variability of lighting and image quality characterizing real-world clinical photography [65], [66], [67]. As delineated in Supplementary Section 5, qualitative error analysis reveals that the model remains susceptible to diagnostic ambiguity in cases of morphological mimicry between benign and malignant entities. The persistence of residual accuracy gaps within extremely rare skin tones indicates that architectural advancements cannot entirely compensate for the fundamental scarcity of diverse high-quality training data for neglected tropical diseases [68]. We further observe that the explicit reasoning generation occasionally collapses into recursive validation loops or repetitive self-correction cycles when confronting high-uncertainty cases, which parallels the stochastic instability observed in other CoT architectures [69]. The linguistic alignment of the model exhibits intermittent instability manifesting as the spontaneous code-switching between English and other languages script during complex reasoning trajectories. Finally, the computational overhead inherent to the multi-step reasoning generation and the dual-route MoE mechanism imposes significant latency constraints that currently hinder real-time inference on edge devices with limited hardware specifications [70], [71].

This study establishes a new paradigm for dermatological AI by harmonizing the precision of large-scale visual encoders with the logical rigor of reasoning-based language models. The incorporation of CoT mechanisms and fairness-aware routing strategies enhances diagnostic performance while simultaneously mitigating algorithmic bias across diverse skin tones. These findings suggest that the future of medical AI lies not merely in scaling model parameters but in the explicit modeling of clinical cognition and the structural enforcement of demographic equity. Future investigations must prioritize the validation of these systems through prospective randomized controlled trials to assess patient-centered outcomes and explore the integration of multimodal data streams for holistic disease profiling [13], [72]. The full realization of this technology depends on the continued refinement of explainable reasoning architectures that empower clinicians and ensure high-fidelity care for all patients irrespective of geographical or phenotypic barriers.

4 METHODS

4.1 Data Curation and Automated Annotation Framework

The reliability of clinical reasoning systems is fundamentally constrained by the quality and diversity of the underlying training data. We constructed a comprehensive composite training corpus comprising 334,618 samples by aggregating three distinct sources to balance generalizability with fine-grained pathological understanding, as detailed in Fig. 1c-e. The core visual representation is derived from 310,830 images in the Derm1M dataset [41], which provides a foundational distribution of common dermatoses. We integrated 8,220 samples from the Fitzpatrick Black Skin Disease Dataset to explicitly counter the performance degradation typically observed in darker skin tones. We further incorporated 15,568 high-quality clinical images from DermNet [73] to enhance the recognition of subtle morphological features that generalist datasets frequently overlook.

Standard image-label pairs invariably fail to capture the logical deduction required for trustworthy diagnosis. We therefore implemented a three-stage automated pipeline to synthesize structured CoT narratives whose specific operational protocols and schematic workflow are explicitly detailed in Supplementary Section 3 and Fig. S1. In the first stage, we employ Gemini 2.5 Pro [74] to generate objective visual descriptions. This model is strictly constrained to document morphological features X_{sem}, such as lesion morphology and anatomic location, without generating diagnostic inferences. The second stage utilizes a locally deployed Kimi-K2-Thinking [75] to synthesize a preliminary reasoning draft by fusing the visual scaffold X_{sem} with the ground-truth diagnosis Y_{gt}. This step simulates the cognitive process of a dermatologist by linking visual evidence to pathological conclusions. The final stage employs DeepSeek-R1 [24] to normalize these drafts into a hierarchical format consisting of disease family, followed by evidence-based descriptors and the final diagnosis. This hierarchical structure enforces internal consistency between observation and conclusion. All generated trajectories undergo random sampling and verification by board-certified dermatologists to guarantee clinical validity.

To systematically mitigate algorithmic bias inherent in unstructured visual data, we employed the Classification Algorithm for Skin Color (CASCo) [76] configured with the default PERLA Palette to automatically quantify the skin tone of every training sample. The algorithm maps the continuous HSV color space of the lesion background to 12 distinct chromatic categories, which we subsequently aggregated into three discrete logical classes representing Caucasian, Asian, and African phenotypes encoded as in-

tegers 0, 1, and 2, respectively. These categorical encodings are foundational to our fairness constraint mechanism as they are transformed into the demographic prior vector $\mathbf{p}_{\text{skin}}$ utilized in the downstream Fairness-Aware MoE Adapter. As detailed in Section 4.3, this prior is explicitly injected into the Dual-Route Gating Network to modulate the expert routing trajectory, effectively decoupling pathological feature extraction from background epidermal pigmentation during the inference process.

4.2 Reasoning-Centric Backbone and Logit-Space Optimization

The preservation of the complex reasoning capabilities established by the Vision-R1-7B MLLM constitutes the architectural cornerstone of SkinGPT-R1 [77]. Indiscriminate parameter fine-tuning of such large-scale reasoning systems frequently leads to the catastrophic forgetting of established logical pathways and degrades the capability for multi-step deduction. We overcome this challenge by strictly freezing the parameters of the backbone language model and implementing a non-invasive Logit-Space Bias Injection mechanism as illustrated in Fig. 1b. This architecture modulates the generation trajectory through external guidance rather than internal weight modification and thereby retains the robust inferential capabilities of the pre-trained core while integrating domain-specific dermatological precision.

The injection process originates within a dedicated bias generator module denoted as Φ, which transforms the visual representation $\mathbf{v}_{\text{vis}} \in \mathbb{R}^{d_v}$ into a latent semantic vector. The implementation utilizes a bottleneck design to distill the high-dimensional visual embedding into a compact feature set before projecting it back to the hidden dimension of the language model d_h. The transformation is formulated as

$$\mathbf{h}_{\text{bias}} = \mathbf{W}_{\text{up}} \tanh\left(\mathbf{W}_{\text{down}} \mathbf{v}_{\text{vis}}\right) \tag{1}$$

where $\mathbf{W}_{\text{down}} \in \mathbb{R}^{d_b \times d_v}$ compresses the input to a bottleneck dimension d_b set to 64, and $\mathbf{W}_{\text{up}} \in \mathbb{R}^{d_h \times d_b}$ restores the representation to the hidden state space. The application of the hyperbolic tangent activation function introduces necessary non-linearity to capture complex morphological relationships within the visual data.

To ensure semantic alignment between the visual signals and the textual generation process, we project the latent bias $\mathbf{h}_{\text{bias}}$ into the vocabulary space $\mathbb{R}^{|V|}$ utilizing the frozen weights of the language model head itself. Let $\mathbf{W}_{\text{head}} \in \mathbb{R}^{|V| \times d_h}$ represent the pre-trained linear mapping from hidden states to token logits. The semantic bias vector $\mathbf{b}_{\text{vocab}}$ is computed as

$$\mathbf{b}_{\text{vocab}} = \mathbf{W}_{\text{head}} \mathbf{h}_{\text{bias}} \tag{2}$$

This formulation guarantees that the injected visual guidance operates within the exact semantic space defined by the pre-trained model. The final probability distribution for the next token y_t at time step t is derived by superimposing this bias onto the original logits $\mathbf{l}_{\text{LM}}$ produced by the frozen backbone according to

$$P(y_t \mid \mathbf{x}, y_{<t}) = \text{softmax}\left(\mathbf{l}_{\text{LM}} + \sigma \cdot (\mathbf{b}_{\text{vocab}} \odot \mathbf{M}_{\text{skin}})\right) \tag{3}$$

where σ is a learnable scalar initialized at 2.5 to dynamically regulate the influence of visual evidence and $\mathbf{M}_{\text{skin}} \in \{0,1\}^{|V|}$ is a binary mask that restricts bias injection to a subset of tokens relevant to dermatological ontology. This selective masking prevents visual noise from disrupting general linguistic fluency.

The optimization of the framework is governed by a composite objective function that rigorously balances causal reasoning accuracy, visual representation fidelity, and phenotypic fairness. We formulate the global loss $\mathcal{L}_{\text{total}}$ as a weighted summation of four distinct components:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{SFT}}\mathcal{L}_{\text{SFT}} + \lambda_{\text{distill}}\mathcal{L}_{\text{distill}} + \lambda_{\text{skin}}\mathcal{L}_{\text{skin}} + \lambda_{\text{aux}}\mathcal{L}_{\text{aux}} \tag{4}$$

Each component targets a specific dimension of the learning process. First, the Supervised Fine-Tuning loss $\mathcal{L}_{\text{SFT}}$ employs the standard cross-entropy objective on the generated tokens to enforce logical coherence and structure in the diagnostic rationale:

$$\mathcal{L}_{\text{SFT}} = -\frac{1}{T}\sum_{t=1}^{T} \log P(y_t \mid \mathbf{v}, y_{<t}; \Theta) \tag{5}$$

where T denotes the sequence length. Second, to ensure the lightweight adapter inherits the robust domain-specific visual semantics of the teacher model, we introduce a feature distillation loss $\mathcal{L}_{\text{distill}}$. This is computed as the cosine distance between the student's projected visual features $\mathbf{v}_S$ and the teacher's embeddings $\mathbf{v}_T$:

$$\mathcal{L}_{\text{distill}} = 1 - \frac{\mathbf{v}_S \cdot \mathbf{v}_T}{\|\mathbf{v}_S\|_2 \|\mathbf{v}_T\|_2} \tag{6}$$

Third, to mitigate algorithmic bias, the skin tone classification loss $\mathcal{L}_{\text{skin}}$ optimizes the fairness-aware gating mechanism to accurately identify phenotypic subgroups. Utilizing the ground-truth skin tone labels y_c, this is formulated as a multi-class cross-entropy loss:

$$\mathcal{L}_{\text{skin}} = -\sum_{c=1}^{3} y_c \log(\hat{y}_c) \tag{7}$$

Finally, to prevent expert collapse within the MoE layer where a subset of experts dominates the routing process, we incorporate an auxiliary load-balancing loss $\mathcal{L}_{\text{aux}}$. Defined over N experts, this term penalizes the correlation between

the routing fraction f_i and the average routing probability P_i:

$$\mathcal{L}_{\text{aux}} = N \sum_{i=1}^{N} f_i \cdot P_i \tag{8}$$

In our experiments, we set $\lambda_{\text{SFT}} = 1.0$ to establish a stable baseline, while modulating $\lambda_{\text{distill}} = 0.1$, $\lambda_{\text{skin}} = 0.1$, and $\lambda_{\text{aux}} = 0.001$ to synergize visual alignment and fairness without overshadowing the primary reasoning objective.

4.3 Fairness-Aware Dual-Route MoE Adapter

The mitigation of algorithmic bias in dermatological AI necessitates a structural decoupling of pathological features from phenotypic characteristics. We address the propensity for MLLMs to overfit majority skin tones by replacing standard feed-forward layers with a Skin-Aware MoE adapter. This architecture dynamically allocates computational capacity based on a dual analysis of visual features and demographic priors to ensure fair performance across the Fitzpatrick scale.

The adapter module consists of a set of $N = 8$ expert networks denoted as $\{E_i\}_{i=1}^{N}$ where each expert functions as a bottleneck Multi-Layer Perceptron. We design each expert to compress the input hidden state $\mathbf{h} \in \mathbb{R}^d$ into a lower-dimensional space of size 64 before projecting it back to the original dimension to filter noise and isolate salient features. The routing mechanism governs the selection of these experts via a Dual-Route Gating Network that synthesizes token-level visual embeddings with image-level skin tone probabilities. Let $\mathbf{p}_{\text{skin}} \in \mathbb{R}^3$ represent the probability distribution over coarse skin tone categories predicted by the upstream classifier. We compute the routing logits $\mathbf{l} \in \mathbb{R}^N$ by additively combining the projections of the visual state and the phenotypic prior as

$$\mathbf{l} = \mathbf{h}\mathbf{W}_I + \mathbf{p}_{\text{skin}}\mathbf{W}_S \tag{9}$$

where $\mathbf{W}_I \in \mathbb{R}^{d \times N}$ and $\mathbf{W}_S \in \mathbb{R}^{3 \times N}$ are learnable projection matrices that map distinct modalities into the shared expert selection space. This additive formulation allows the model to modulate the expert activation pattern based on skin tone without overwriting the pathological signal contained in the visual embedding.

We convert these logits into routing probabilities using the softmax function and select the top $K = 2$ experts to execute the computation. We adopt $K = 2$ as this configuration strikes an optimal balance between computational efficiency and the flexibility required to capture complex phenotypic variations [78], [79]. The sparse gating weights $\mathbf{g}$ are normalized across the selected experts to ensure magnitude consistency. The final output $\mathbf{y}$ incorporates a residual connection around the expert branch to facilitate gradient flow, and is formally defined as

$$\mathbf{y} = \mathbf{h} + \sum_{i \in \mathcal{K}} \frac{\exp(\mathbf{l}_i)}{\sum_{j \in \mathcal{K}} \exp(\mathbf{l}_j)} E_i(\mathbf{h}) \tag{10}$$

where $\mathcal{K}$ represents the set of indices corresponding to the top K elements in $\mathbf{l}$. This residual structure preserves the pre-trained knowledge of the backbone while the experts inject phenotype-specific refinements.

To prevent the routing mechanism from collapsing into a trivial state where only a few experts are utilized, we enforce a load-balancing auxiliary loss $\mathcal{L}_{\text{aux}}$. We define $\bar{q}_i$ as the mean routing probability assigned to expert i across a batch and $\bar{f}_i$ as the fraction of samples for which expert i is selected. The objective minimizes the correlation between probability and frequency to encourage uniform expert utilization. The complete computational trajectory encompassing both the forward inference and the training loss derivation is detailed in Algorithm 1.

Algorithm 1. Dual-Route Fairness-Aware MoE Training and Inference

Require: Input features $\mathbf{h} \in \mathbb{R}^{B \times L \times D}$, Skin priors $\mathbf{p}_{\text{skin}} \in \mathbb{R}^{B \times 3}$
Require: Expert networks $\{E_i\}_{i=1}^{N}$, Projection matrices $\mathbf{W}_I, \mathbf{W}_S$
Step 1: Dual-Source Gating
Compute routing logits $\mathbf{L} \leftarrow \mathbf{h}\mathbf{W}_I + \mathbf{p}_{\text{skin}}\mathbf{W}_S$
Calculate routing probabilities $\mathbf{P} \leftarrow \text{Softmax}(\mathbf{L})$
Step 2: Top-K Selection
Identify indices $\mathcal{I}$ and values $\mathcal{V}$ for top $K = 2$ elements in $\mathbf{P}$
Normalize gating weights $\mathcal{W} \leftarrow \mathcal{V} / \sum \mathcal{V}$
Step 3: Sparse Execution
Initialize output accumulator $\mathbf{O} \leftarrow \mathbf{0}$
for each expert $i \in \{1, \ldots, N\}$ **do**
  Identify active tokens mask $\mathcal{M}_i \leftarrow \{(b, l) \mid i \in \mathcal{I}_{b,l}\}$
  if $\mathcal{M}_i$ is not empty **then**
    Compute expert output $\mathbf{h}_{\text{out}} \leftarrow E_i(\mathbf{h}[\mathcal{M}_i])$
    Accumulate weighted result $\mathbf{O}[\mathcal{M}_i] \leftarrow \mathbf{O}[\mathcal{M}_i] + \mathcal{W}_{b,l,i} \cdot \mathbf{h}_{\text{out}}$
Compute final output with residual $\mathbf{y} \leftarrow \mathbf{h} + \mathbf{O}$
Step 4: Auxiliary Loss Calculation
Compute mean routing probability $\bar{\mathbf{q}} \leftarrow \frac{1}{B \cdot L} \sum \mathbf{P}$
Compute selection frequency $\bar{\mathbf{f}} \leftarrow \frac{1}{B \cdot L} \sum \mathbb{I}(i \in \mathcal{I})$
Calculate load-balancing loss $\mathcal{L}_{\text{aux}} \leftarrow N \sum_{i=1}^{N} \bar{\mathbf{q}}_i \cdot \bar{\mathbf{f}}_i$
return $\mathbf{y}, \mathcal{L}_{\text{aux}}$

4.4 Visual Representation Learning via Teacher-Student Distillation

The generalized visual encoders inherent to standard MLLMs frequently lack the domain-specific capacity required to resolve subtle dermatological morphology. We address this issue by implementing a distillation strategy that transfers the robust representation capabilities of the specialist PanDerm [23] foundation model into our lightweight adapter. We designate PanDerm as the frozen teacher $\mathcal{T}$ and our trainable patch distillation head as the student $\mathcal{S}$ to enforce semantic alignment without incurring the computational overhead of a dual backbone during inference.

The student architecture is engineered as a hierarchical patch processing module that accepts variable-resolution image inputs. Let $\mathbf{X} \in \mathbb{R}^{M \times d_{\text{in}}}$ denote the sequence of flattened patch features derived from the input grid, where M represents the total number of patches and d_{in} indicates the dimensionality of the raw visual features. The network first applies a learnable linear transformation $\mathbf{W}_{\text{in}} \in \mathbb{R}^{d_{\text{in}} \times d_{\text{model}}}$ to map the raw features into a shared embedding space of dimension d_{model}. Crucially, the module computes an intrinsic phenotypic prior $\mathbf{p}_{\text{skin}}$ by aggregating global visual statistics via a dedicated internal classifier branch defined as

$$\mathbf{p}_{\text{skin}} = \text{softmax}\left(\Phi_{\text{skin}}\left(\frac{1}{M}\sum_{i=1}^{M}\mathbf{W}_{\text{in}}\mathbf{x}_i\right)\right) \tag{11}$$

where $\mathbf{x}_i$ corresponds to the i-th patch vector row in $\mathbf{X}$, and Φ_{skin} represents a bottleneck Multi-Layer Perceptron designed to project the aggregated global features into the skin tone probability space. This internally generated probability distribution conditions a subsequent stack of Skin-Aware MoE layers that refine the patch representations through phenotype-specific routing.

The final student representation $\mathbf{h}_{\mathcal{S}}$ is obtained by aggregating the output of the adapter stack via global average pooling and mapping the result to the target feature space via a learnable projection matrix $\mathbf{W}_{\text{out}}$. To ensure semantic fidelity, we optimize a distillation objective $\mathcal{L}_{\text{distill}}$, which minimizes the cosine dissimilarity between the projected student embedding $\mathbf{h}_{\mathcal{S}}$ and the reference semantic vector $\mathbf{h}_{\mathcal{T}}$ extracted by the frozen teacher model according to

$$\mathcal{L}_{\text{distill}} = \gamma\left(1 - \frac{\mathbf{h}_{\mathcal{S}}^{\top}\mathbf{h}_{\mathcal{T}}}{\|\mathbf{h}_{\mathcal{S}}\|_2\|\mathbf{h}_{\mathcal{T}}\|_2}\right) \tag{12}$$

where $\|\cdot\|_2$ denotes the L_2 norm, and the scaling factor γ is initialized at 10.0 to ensure gradient stability relative to the primary language modeling loss. This architecture allows the adapter to internalize the fine-grained visual discrimination of the specialist teacher during the training phase while permitting the complete removal of the heavy teacher network during clinical deployment.

4.5 Evaluation Ecosystem with DermBench and DermEval

To rigorously assess the clinical validity and reasoning capabilities of SkinGPT-R1, we employ the comprehensive evaluation ecosystem established in our prior work [45]. This framework addresses the inherent limitations of traditional n-gram metrics, which frequently fail to capture the semantic nuance of medical reasoning. Standard metrics such as BLEU [80] quantify the lexical overlap precision between a candidate hypothesis and reference text according to

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{13}$$

where BP denotes the brevity penalty to prevent short generation bias, w_n represents the uniform weight for n-grams of size n, and p_n indicates the modified n-gram precision. Similarly, recall-oriented metrics like ROUGE-N [81] measure the coverage of reference content via

$$\text{ROUGE-N} = \frac{\sum_{S \in \text{Refs}}\sum_{\text{gram}_n \in S}\text{Count}_{\text{match}}(\text{gram}_n)}{\sum_{S \in \text{Refs}}\sum_{\text{gram}_n \in S}\text{Count}(\text{gram}_n)} \tag{14}$$

where $\text{Count}_{\text{match}}$ is the maximum number of n-grams co-occurring in the candidate and the set of Reference summaries Refs, and Count is the total number of n-grams in the reference. While computationally efficient, these metrics correlate poorly with clinical factual correctness. Our protocol therefore relies on the synergy of a high-quality benchmark known as DermBench and a clinically aligned automated evaluator termed DermEval.

4.5.1 DermBench with LLM-based Evaluation

DermBench serves as the gold-standard reference for our experiments and consists of 4,000 diverse dermatology images paired with expert-verified diagnostic narratives. Unlike traditional benchmarks that rely on simple classification accuracy, DermBench evaluates the full logical trajectory of a diagnosis. For each case, we generate a candidate narrative by prompting the target model to analyze morphological features, provide differential diagnoses, and justify a final conclusion.

To score these candidates, we employ DeepSeek-R1 as an impartial judge. We select this model to minimize self-scoring bias and leverage its strong reasoning capabilities without domain-specific fine-tuning. The judge receives both the candidate text and the certified gold standard reference. It is instructed to assign a score from 0 to 5 across six critical clinical dimensions, including Accuracy, Safety, Medical Groundedness, Clinical Coverage, Reasoning Coherence, and Description Precision. This multidimensional scoring ensures that the evaluation captures both the factual correctness and the logical soundness of the diagnostic process.

4.5.2 DermEval with Score-Oriented Reinforcement Learning

To facilitate scalable and consistent assessment, we utilize DermEval, which is a LLaVA-based model fine-tuned to function as a surrogate for human experts. The training of DermEval proceeds in a two-stage alignment process designed to map an image and a diagnostic text to six scalar scores and a structured critique.

In the first stage, we optimize the model to produce a canonical and parsable evaluation format via supervised fine-tuning. The model receives an input x containing the image I and diagnostic text d, and the objective is to minimize the token-level cross-entropy loss $\mathcal{L}_{\text{TEXT}}$ against a templated target evaluation y^*:

$$\mathcal{L}_{\text{TEXT}} = \text{CE}\left(y,\, y^*\right) \tag{15}$$

where y represents the predicted evaluation sequence. This stage ensures that the model outputs stable and structured data that can be reliably parsed for downstream reward calculation.

In the second stage, we align the predicted scores with human expert judgment using a strategy termed Score-Oriented Reinforcement Learning with an Exponential Moving Average (EMA) Baseline. For each instance, the model generates an evaluation text $\hat{y}$ from which six scores $\hat{\mathbf{s}}$ are extracted by an external parser. We define the alignment reward r as the negative mean squared error between the parsed scores and the physician ground truth scores $\mathbf{s}^*$ on valid dimensions:

$$r(\hat{\mathbf{s}}, \mathbf{s}^*) = -\frac{1}{K_y} \sum_{k \in \mathcal{I}_y} \left(\hat{s}_k - s_k^*\right)^2 \tag{16}$$

where $\mathcal{I}_y$ denotes the set of indices for dimensions that were successfully parsed, and $K_y = |\mathcal{I}_y|$ is the cardinality of this set. To reduce gradient variance and stabilize updates, we maintain a baseline b that aggregates recent rewards via an exponential moving average. The baseline is updated as $b \leftarrow \beta b + (1-\beta) r$, where β is the momentum coefficient. The advantage is subsequently computed as $\text{adv} = r - b$. The reinforcement learning objective $\mathcal{L}_{\text{RL}}$ maximizes this advantage over the generation trajectory $\mathcal{T}$:

$$\mathcal{L}_{\text{RL}} = -\text{adv} \cdot \text{mean}_{t \in \mathcal{T}} \log p_\theta\left(y_t \mid y_{<t}, x\right) \tag{17}$$

where p_θ represents the probability distribution parameterized by the model weights, and y_t denotes the token generated at step t given the history $y_{<t}$ and input x. The final objective function combines the reinforcement loss and the text formatting loss with weights λ_{RL} and λ_{TEXT} respectively:

$$\mathcal{L}_{\text{TOTAL}} = \lambda_{\text{RL}} \mathcal{L}_{\text{RL}} + \lambda_{\text{TEXT}} \mathcal{L}_{\text{TEXT}} \tag{18}$$

This composite loss ensures that DermEval maintains syntactic correctness while progressively aligning its scoring logic with board-certified dermatologists.

4.6 Implementation Details

All computational experiments were conducted on a high-performance cluster populated with eight NVIDIA RTX 4090 GPUs utilizing 24GB of video memory per device. For the training of the primary SkinGPT-R1 model, we adopted a parameter-efficient freeze-tuning strategy to preserve the pre-trained reasoning capabilities. The optimization process updated approximately 4.6 million parameters constituting merely 0.056% of the 8.3 billion total parameters, while the vision-language backbone remained frozen. We employed the AdamW optimizer with a peak learning rate of 1.0×10^{-4} modulated by a cosine decay scheduler and a warmup ratio of 0.03. To ensure memory efficiency and computational throughput, we utilized Brain Floating Point 16 precision alongside Flash Attention mechanisms. The training protocol utilized a per-device batch size of 4 with 4 steps of gradient accumulation, resulting in a total training process spanning 10 epochs over approximately 70 hours.

5 Data Availability

The data supporting the findings of this study are primarily sourced from publicly available datasets, which are cited within the manuscript and accessible via their respective repositories. Specifically, the datasets utilized in this work are available at the following locations: Derm1M at https://github.com/SiyuanYan1/Derm1M; Fitzpatrick Black Skin Disease Dataset at https://www.kaggle.com/datasets/oyebamijimicheal/fitzpatrick-black-skin-disease-dataset; Derm12345 at https://www.nature.com/articles/s41597-024-04104-3; SkinCon at https://skincon-dataset.github.io/; Dermnet at https://dermnetnz.org/; Derm7pt at https://github.com/jeremykawahara/derm7pt; MSLD v2.0 at https://www.kaggle.com/datasets/joydippaul/mpox-skin-lesion-dataset-version-20-msld-v20; Fitz17k at https://github.com/mattgroh/fitzpatrick17k; DDI at https://ddi-dataset.github.io/; and the PAD-UFES-20 at https://www.sciencedirect.com/science/article/pii/S235234092031115X. A portion of the data, specifically the clinical samples used for internal validation, is not publicly available due to privacy restrictions and ethical considerations regarding patient confidentiality. Derived data supporting the findings of this study are available from the corresponding author upon reasonable request, subject to appropriate data usage agreements.

6 Code availability

To facilitate academic exchange while strictly adhering to data privacy and security regulations, the inference code for SkinGPT-R1 is publicly accessible at: https://huggingface.co/yuhos16/SkinGPT-R1. The full model weights are currently not publicly released to protect sensitive patient information contained within the training corpus. However, for non-commercial research purposes, access to the trained model parameters may be granted upon reasonable request to the corresponding author, subject to ethical approval and data usage agreements. We have provided comprehensive details regarding our methodology, model architecture, and training procedures in the main text and Supplementary Information to ensure the reproducibility of our experiments. Key foundational libraries utilized in this work, including PyTorch and Hugging Face Transformers, are open-source and publicly available.

7 Acknowledgements

Y.S., Z.C., Y.H., Zijian W., Y.Y., Ziwen W., X.Z., and J.Z. were supported by The Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen), under Award No UDF01004172.

8 Credit author statement

Y.S., Z.C., J.Z. contributed to the concept of the study and designed the research. Y.S., Z.C., S.Y., Y.H., Y.Y., Ziwen W., X.Z. collected and processed the data. Y.S., Y.H. conducted the study. Y.X., S.Z., L.S., T.L., W.L. analyzed the data. Y.S., Y.H., Zijian W., Y.Z., J.Z. co-wrote the manuscript. Y.S., S.Y., Y.Z., J.Q., Z.G., J.Z. critically revised the manuscript. Y.S., Z.C., S.Y., Y.Y., J.Z. performed the technical review. All authors discussed the results and provided comments regarding the manuscript.

9 Competing Interests

The authors have declared no competing interests.

REFERENCES

[1] C. Karimkhani, R. P. Dellavalle, L. E. Coffeng, C. Flohr, R. J. Hay, S. M. Langan, E. O. Nsoesie, A. J. Ferrari, H. E. Erskine, J. I. Silverberg *et al.*, "Global skin disease morbidity and mortality: an update from the global burden of disease study 2013," *JAMA dermatology*, vol. 153, no. 5, pp. 406–412, 2017.

[2] M. R. Laughter, M. B. Maymone, C. Karimkhani, C. Rundle, S. Hu, S. Wolfe, K. Abuabara, P. Hollingsworth, G. S. Weintraub, C. A. Dunnick *et al.*, "The burden of skin and subcutaneous diseases in the united states from 1990 to 2017," *JAMA dermatology*, vol. 156, no. 8, pp. 874–881, 2020.

[3] H. Feng, J. Berk-Krauss, P. W. Feng, and J. A. Stein, "Comparison of dermatologist density between urban and rural counties in the united states," *JAMA dermatology*, vol. 154, no. 11, pp. 1265–1271, 2018.

[4] Y. Liu, A. Jain, C. Eng, D. H. Way, K. Lee, P. Bui, K. Kanada, G. de Oliveira Marinho, J. Gallegos, S. Gabriele *et al.*, "A deep learning system for differential diagnosis of skin diseases," *Nature medicine*, vol. 26, no. 6, pp. 900–908, 2020.

[5] J. Lester, S. Taylor, and M.-M. Chren, "Under-representation of skin of colour in dermatology images: not just an educational issue," *British Journal of Dermatology*, vol. 180, no. 6, pp. 1521–1522, 2019.

[6] R. Daneshjou, K. Vodrahalli, R. A. Novoa, M. Jenkins, W. Liang, V. Rotemberg, J. Ko, S. M. Swetter, E. E. Bailey, O. Gevaert *et al.*, "Disparities in dermatology ai performance on a diverse, curated clinical image set," *Science advances*, vol. 8, no. 31, p. eabq6147, 2022.

[7] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[8] S. S. Han, M. S. Kim, W. Lim, G. H. Park, I. Park, and S. E. Chang, "Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm," *Journal of Investigative Dermatology*, vol. 138, no. 7, pp. 1529–1538, 2018.

[9] K. P. Venkatesh, M. M. Raza, G. Nickel, S. Wang, and J. C. Kvedar, "Deep learning models across the range of skin disease," *NPJ Digital Medicine*, vol. 7, no. 1, p. 32, 2024.

[10] N. C. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler *et al.*, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic)," in *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*. IEEE, 2018, pp. 168–172.

[11] O. Jaiyeoba, E. Ogbuju, G. Ataguba, O. Jaiyeoba, J. D. Omaye, I. Eze, and F. Oladipo, "State-of-the-art skin disease classification: a review of deep learning models," *Network Modeling Analysis in Health Informatics and Bioinformatics*, vol. 14, no. 1, p. 9, 2025.

[12] M. Moor, O. Banerjee, Z. S. H. Abad, H. M. Krumholz, J. Leskovec, E. J. Topol, and P. Rajpurkar, "Foundation models for generalist medical artificial intelligence," *Nature*, vol. 616, no. 7956, pp. 259–265, 2023.

[13] T. Tu, S. Azizi, D. Driess, M. Schaekermann, M. Amin, P.-C. Chang, A. Carroll, C. Lau, R. Tanno, I. Ktena *et al.*, "Towards generalist biomedical ai," *Nejm Ai*, vol. 1, no. 3, p. AIoa2300138, 2024.

[14] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, and J. Gao, "Llava-med: Training a large language-and-vision assistant for biomedicine in one day," *arXiv preprint arXiv:2306.00890*, 2023.

[15] S. Zhang, Y. Xu, N. Usuyama, H. Xu, J. Bagga, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri *et al.*, "Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs," *arXiv preprint arXiv:2303.00915*, 2023.

[16] C. Wu, W. Lin, X. Zhang, Y. Zhang, W. Xie, and Y. Wang, "Pmc-llama: toward building open-source language models for medicine," *Journal of the American Medical Informatics Association*, vol. 31, no. 9, pp. 1833–1843, 2024.

[17] A. Sellergren, S. Kazemzadeh, T. Jaroensri, A. Kiraly, M. Traverse, T. Kohlberger, S. Xu, F. Jamil, C. Hughes, C. Lau *et al.*, "Medgemma technical report," *arXiv preprint arXiv:2507.05201*, 2025.

[18] J. Chen, C. Gui, R. Ouyang, A. Gao, S. Chen, G. H. Chen, X. Wang, R. Zhang, Z. Cai, K. Ji *et al.*, "Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale," *arXiv preprint arXiv:2406.19280*, 2024.

[19] S. Jiang, Y. Wang, S. Song, T. Hu, C. Zhou, B. Pu, Y. Zhang, Z. Yang, Y. Feng, J. T. Zhou *et al.*, "Hulu-med: A transparent generalist model towards holistic medical vision-language understanding," *arXiv preprint arXiv:2510.08668*, 2025.

[20] Y. Hu, T. Li, Q. Lu, W. Shao, J. He, Y. Qiao, and P. Luo, "Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 22 170–22 183.

[21] L. Luo, B. Tang, X. Chen, R. Han, and T. Chen, "Vividmed: Vision language model with versatile visual grounding for medicine," *arXiv preprint arXiv:2410.12694*, 2024.

[22] J. Zhou, X. He, L. Sun, J. Xu, X. Chen, Y. Chu, L. Zhou, X. Liao, B. Zhang, S. Afvari *et al.*, "Pre-trained multimodal large language model enhances dermatological diagnosis using skingpt-4," *Nature Communications*, vol. 15, no. 1, p. 5649, 2024.

[23] S. Yan, Z. Yu, C. Primiero, C. Vico-Alonso, Z. Wang, L. Yang, P. Tschandl, M. Hu, L. Ju, G. Tan *et al.*, "A multimodal vision foundation model for clinical dermatology," *Nature Medicine*, pp. 1–12, 2025.

[24] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi *et al.*, "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," *arXiv preprint arXiv:2501.12948*, 2025.

[25] A. Jaech, A. Kalai, A. Lerer, A. Richardson, A. El-Kishky, A. Low, A. Helyar, A. Madry, A. Beutel, A. Carney *et al.*, "Openai o1 system card," *arXiv preprint arXiv:2412.16720*, 2024.

[26] J. Pan, C. Liu, J. Wu, F. Liu, J. Zhu, H. B. Li, C. Chen, C. Ouyang, and D. Rueckert, "Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2025, pp. 337–347.

[27] J. Chen, Z. Cai, K. Ji, X. Wang, W. Liu, R. Wang, J. Hou, and B. Wang, "Huatuogpt-o1, towards medical complex reasoning with llms," *arXiv preprint arXiv:2412.18925*, 2024.

[28] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

[29] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou, "Self-consistency improves chain of thought reasoning in language models," *arXiv preprint arXiv:2203.11171*, 2022.

[30] Z. Zhang, A. Zhang, M. Li, H. Zhao, G. Karypis, and A. Smola, "Multimodal chain-of-thought reasoning in language models," *arXiv preprint arXiv:2302.00923*, 2023.

[31] J. Liu, Y. Wang, J. Du, J. T. Zhou, and Z. Liu, "Medcot: Medical chain of thought via hierarchical expert," *arXiv preprint arXiv:2412.13736*, 2024.

[32] W. Wang, Z. Ma, M. Ding, S. Zheng, S. Liu, J. Liu, J. Ji, W. Chen, X. Li, L. Shen *et al.*, "Medical reasoning in the era of llms: A systematic review of enhancement techniques and applications," *arXiv preprint arXiv:2508.00669*, 2025.

[33] P. Tschandl, C. Rinner, Z. Apalla, G. Argenziano, N. Codella, A. Halpern, M. Janda, A. Lallas, C. Longo, J. Malvehy *et al.*, "Human–computer collaboration for skin cancer recognition," *Nature medicine*, vol. 26, no. 8, pp. 1229–1234, 2020.

[34] M. Groh, O. Badri, R. Daneshjou, A. Koochek, C. Harris, L. R. Soenksen, P. M. Doraiswamy, and R. Picard, "Deep learning-aided decision support for diagnosis of skin disease across skin tones," *Nature Medicine*, vol. 30, no. 2, pp. 573–583, 2024.

[35] L. Seyyed-Kalantari, H. Zhang, M. B. McDermott, I. Y. Chen, and M. Ghassemi, "Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations," *Nature medicine*, vol. 27, no. 12, pp. 2176–2182, 2021.

[36] V. Rotemberg, N. Kurtansky, B. Betz-Stablein, L. Caffery, E. Chousakos, N. Codella, M. Combalia, S. Dusza, P. Guitera, D. Gutman *et al.*, "A patient-centric dataset of images and metadata for identifying melanomas using clinical context," *Scientific data*, vol. 8, no. 1, p. 34, 2021.

[37] S. N. Ali, M. T. Ahmed, T. Jahan, J. Paul, S. S. Sani, N. Noor, A. N. Asma, and T. Hasan, "A web-based mpox skin lesion detection system using state-of-the-art deep learning models considering racial diversity," *Biomedical Signal Processing and Control*, vol. 98, p. 106742, 2024.

[38] M. Groh, C. Harris, L. Soenksen, F. Lau, R. Han, A. Kim, A. Koochek, and O. Badri, "Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 1820–1828.

[39] M. Groh, C. Harris, R. Daneshjou, O. Badri, and A. Koochek, "Towards transparency in dermatology image datasets with skin tone annotations by experts, crowds, and an algorithm," *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, no. CSCW2, pp. 1–26, 2022.

[40] R. Daneshjou, M. Yuksekgonul, Z. R. Cai, R. Novoa, and J. Y. Zou, "Skincon: A skin disease dataset densely annotated by domain experts for fine-grained debugging and analysis," *Advances in Neural Information Processing Systems*, vol. 35, pp. 18 157–18 167, 2022.

[41] S. Yan, M. Hu, Y. Jiang, X. Li, H. Fei, P. Tschandl, H. Kittler, and Z. Ge, "Derm1m: A million-scale vision-language dataset aligned with clinical ontology knowledge for dermatology," *arXiv preprint arXiv:2503.14911*, 2025.

[42] W. Zeng, Y. Sun, C. Ma, W. Tan, and B. Yan, "Mm-skin: Enhancing dermatology vision-language model with an image-text dataset derived from textbooks," in *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025, pp. 3769–3778.

[43] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," *Advances in neural information processing systems*, vol. 36, pp. 10 088–10 115, 2023.

[44] P. Gao, J. Han, R. Zhang, Z. Lin, S. Geng, A. Zhou, W. Zhang, P. Lu, C. He, X. Yue *et al.*, "Llama-adapter v2: Parameter-efficient visual instruction model," *arXiv preprint arXiv:2304.15010*, 2023.

[45] Y. Shen, J. Qian, S. Zhang, Z. Chen, T. Lu, and J. Zhou, "Towards trustworthy dermatology mllms: A benchmark and multimodal evaluator for diagnostic narratives," *arXiv preprint arXiv:2511.09195*, 2025.

[46] W.-w. Yim, Y. Fu, Z. Sun, A. B. Abacha, M. Yetisgen, and F. Xia, "Dermavqa: A multilingual visual question answering dataset for dermatology," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024, pp. 209–219.

[47] S. Bedi, H. Cui, M. Fuentes, A. Unell, M. Wornow, J. M. Banda, N. Kotecha, T. Keyes, Y. Mai, M. Oez *et al.*, "Medhelm: Holistic evaluation of large language models for medical tasks," *arXiv preprint arXiv:2505.23802*, 2025.

[48] A. Yilmaz, S. P. Yasar, G. Gencoglan, and B. Temelkuran, "Derm12345: A large, multisource dermatoscopic skin lesion dataset with 40 subclasses," *Scientific Data*, vol. 11, no. 1, p. 1302, 2024.

[49] J. Kawahara, S. Daneshvar, G. Argenziano, and G. Hamarneh, "Seven-point checklist and skin lesion classification using multitask multimodal neural nets," *IEEE journal of biomedical and health informatics*, vol. 23, no. 2, pp. 538–546, 2018.

[50] A. G. Pacheco, G. R. Lima, A. S. Salomao, B. Krohling, I. P. Biral, G. G. de Angelo, F. C. Alves Jr, J. G. Esgario, A. C. Simora, P. B. Castro *et al.*, "Pad-ufes-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones," *Data in brief*, vol. 32, p. 106221, 2020.

[51] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang, "Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization," *arXiv preprint arXiv:1911.08731*, 2019.

[52] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature machine intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[53] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, "Key challenges for delivering clinical impact with artificial intelligence," *BMC medicine*, vol. 17, no. 1, p. 195, 2019.

[54] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nature medicine*, vol. 25, no. 1, pp. 44–56, 2019.

[55] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl *et al.*, "Large language models encode clinical knowledge," *Nature*, vol. 620, no. 7972, pp. 172–180, 2023.

[56] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, "Shortcut learning in deep neural networks," *Nature Machine Intelligence*, vol. 2, no. 11, pp. 665–673, 2020.

[57] K. Watson, "D. kahneman.(2011). thinking, fast and slow. new york, ny: Farrar, straus and giroux. 499 pages." *Canadian Journal of Program Evaluation*, vol. 26, no. 2, pp. 111–113, 2011.

[58] Y. Bengio, "The consciousness prior," *arXiv preprint arXiv:1709.08568*, 2017.

[59] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller, "Unmasking clever hans predictors and assessing what machines really learn," *Nature communications*, vol. 10, no. 1, p. 1096, 2019.

[60] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," *ACM computing surveys*, vol. 55, no. 12, pp. 1–38, 2023.

[61] L. A. Celi, J. Cellini, M.-L. Charpignon, E. C. Dee, F. Dernoncourt, R. Eber, W. G. Mitchell, L. Moukheiber, J. Schirmer, J. Situ *et al.*, "Sources of bias in artificial intelligence that perpetuate healthcare disparities—a global review," *PLOS digital health*, vol. 1, no. 3, p. e0000022, 2022.

[62] M. Komorowski, L. A. Celi, O. Badawi, A. C. Gordon, and A. A. Faisal, "The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care," *Nature medicine*, vol. 24, no. 11, pp. 1716–1720, 2018.

[63] E. J. Topol, "Welcoming new guidelines for ai clinical research," *Nature medicine*, vol. 26, no. 9, pp. 1318–1320, 2020.

[64] C. Snoswell, A. Finnane, M. Janda, H. P. Soyer, and J. A. Whitty, "Cost-effectiveness of store-and-forward teledermatology: a systematic review," *JAMA dermatology*, vol. 152, no. 6, pp. 702–708, 2016.

[65] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study," *PLoS medicine*, vol. 15, no. 11, p. e1002683, 2018.

[66] S. G. Finlayson, J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam, and I. S. Kohane, "Adversarial attacks on medical machine learning," *Science*, vol. 363, no. 6433, pp. 1287–1289, 2019.

[67] X. Huang, X. Wang, H. Zhang, Y. Zhu, J. Xi, J. An, H. Wang, H. Liang, and C. Pan, "Medical mllm is vulnerable: Cross-modality jailbreak and mismatched attacks on medical multimodal large language models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 4, 2025, pp. 3797–3805.

[68] Y. Yang and Z. Xu, "Rethinking the value of labels for improving class-imbalanced learning," *Advances in neural information processing systems*, vol. 33, pp. 19 290–19 301, 2020.

[69] M. Turpin, J. Michael, E. Perez, and S. R. Bowman, "Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[70] C.-J. Wu, R. Raghavendra, U. Gupta, B. Acun, N. Ardalani, K. Maeng, G. Chang, F. Aga, J. Huang, C. Bai *et al.*, "Sustainable ai: Environmental implications, challenges and opportunities," *Proceedings of machine learning and systems*, vol. 4, pp. 795–813, 2022.

[71] D. Xu, Y. Li, M. Li, X. Chen, Y. Tian, and D. Li, "A survey on edge intelligence: From deep learning to large language models," *IEEE Communications Surveys & Tutorials*, 2024.

[72] X. Liu, L. Faes, A. U. Kale, S. K. Wagner, D. J. Fu, A. Bruynseels, T. Mahendiran, G. Moraes, M. Shamdas, C. Kern *et al.*, "A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis," *The lancet digital health*, vol. 1, no. 6, pp. e271–e297, 2019.

[73] Dermnet. (2020) Dermnet. [Online]. Available: https://dermnetnz.org/

[74] G. Comanici, E. Bieber, M. Schaekermann, I. Pasupat, N. Sachdeva, I. Dhillon, M. Blistein, O. Ram, D. Zhang, E. Rosen *et al.*, "Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities," *arXiv preprint arXiv:2507.06261*, 2025.

[75] K. Team, Y. Bai, Y. Bao, G. Chen, J. Chen, N. Chen, R. Chen, Y. Chen, Y. Chen, Y. Chen *et al.*, "Kimi k2: Open agentic intelligence," *arXiv preprint arXiv:2507.20534*, 2025.

[76] R. A. Rejón Piña and C. Ma, "Classification algorithm for skin color (CASCo): A new tool to measure skin color in social science research," *Behavior Research Methods*, 2023.

[77] W. Huang, B. Jia, Z. Zhai, S. Cao, Z. Ye, F. Zhao, Z. Xu, Y. Hu, and S. Lin, "Vision-r1: Incentivizing reasoning capability in multimodal large language models," *arXiv preprint arXiv:2503.06749*, 2025.

[78] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," *arXiv preprint arXiv:1701.06538*, 2017.

[79] W. Fedus, B. Zoph, and N. Shazeer, "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity," *Journal of Machine Learning Research*, vol. 23, no. 120, pp. 1–39, 2022.

[80] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[81] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

Supplementary Files

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

- 7demo.mp4
- 3SupplementarySkinGPTR1compress.pdf