Towards Trustworthy Dermatology MLLMs: A Benchmark and Multimodal Evaluator for Diagnostic Narratives

Yuhao Shen[†1], Jiahe Qian[†1,2], Zhangtianyi Chen[1], and Juexiao Zhou[*1]

[1] The Chinese University of Hong Kong, Shenzhen, China
[2] Institute of Automation, Chinese Academy of Sciences, Beijing, China

Abstract. Multimodal large language models (LLMs) are increasingly used to generate dermatology diagnostic narratives directly from images. However, reliable evaluation remains the primary bottleneck for responsible clinical deployment. We introduce a novel evaluation framework that combines **DermBench**, a meticulously curated benchmark, with **DermEval**, a robust automatic evaluator, to enable clinically meaningful, reproducible, and scalable assessment. We build DermBench, which pairs 4000 real-world dermatology images with expert-certified diagnostic narratives and uses an LLM-based judge to score candidate narratives across clinically grounded dimensions, enabling consistent and comprehensive evaluation of multimodal models. For individual case assessment, we train DermEval, a reference-free multimodal evaluator. Given an image and a generated narrative, DermEval produces a structured critique along with an overall score and per-dimension ratings. This capability enables fine-grained, per-case analysis, which is critical for identifying model limitations and biases. Experiments on a diverse dataset of 4500 cases demonstrate that DermBench and DermEval achieve close alignment with expert ratings, with mean deviations of 0.251 and 0.117 (out of 5) respectively, providing reliable measurement of diagnostic ability and trustworthiness across different multimodal LLMs. The benchmark resources are publicly available at `https://github.com/yuhos16/DermBench`.

1 Introduction

Dermatologic diseases are prevalent across populations and impose a persistent clinical burden [23, 47]. The widespread availability of smartphone and clinical imaging has created a rich visual substrate for computational analysis [5, 13, 36]. Recent advances in multimodal LLMs [35, 67] have made it possible to transform a single dermatology image into a full diagnostic narrative [29, 55, 61] that describes visible findings, articulates reasoning, and proposes differential diagnoses [20, 27, 30, 38, 42, 68]. This capability is attractive for scalable decision

[*] Corresponding: `juexiao.zhou@gmail.com`.
[†] Equal contribution.

support [8, 14, 16, 18, 32, 37, 56], yet it is also fragile [3, 34, 60]. A narrative that is fluent but clinically unsound can mislead users and create safety risks [3, 9]. Responsible deployment therefore requires rigorous and clinically grounded evaluation of generated texts [1].

Evaluating diagnostic narratives is more demanding than producing them [3, 6, 28, 40, 57]. Generic text similarity metrics and open ended preference judgments correlate weakly with clinical utility because they ignore visual evidence, downplay patient safety, and fail to test the internal consistency between observed signs and proposed conclusions [2, 11, 12, 45, 63]. Existing resources in dermatology focus on images with categorical labels, while standardized diagnostic narratives and multimodal benchmarks remain scarce. Expert grading is reliable, yet the associated cost limits scale and hinders timely model iteration. These gaps motivate a principled evaluation space that reflects clinical practice and a mechanism that can compare models fairly and monitor risk over time [48, 62].

We first construct a dataset intended to provide certified exemplars and supervision signals for evaluation. We construct paired image and reasoning data through a dual stream process and obtain clinician curated reference narratives together with concise rationales for scoring. The procedure yields high-quality exemplars and a diverse set of non perfect cases that reveal typical failure modes and provide supervision signals for evaluation.

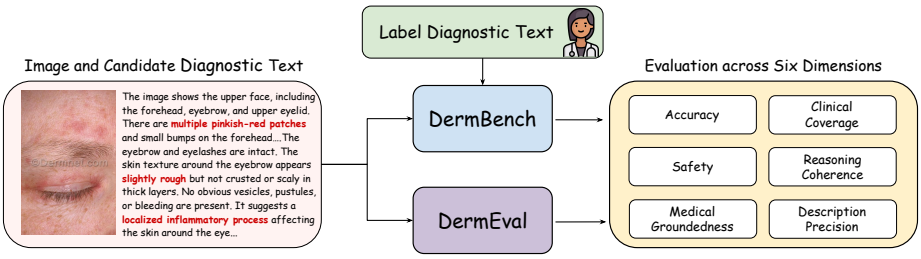

Fig. 1: Overview of DermBench and DermEval. DermBench evaluates a candidate diagnostic narrative by comparing it to a physician-approved reference text for the same image, whereas DermEval evaluates directly from the given image and diagnostic text without requiring a reference.

We then establish a systematic evaluation framework that comprises DermBench and DermEval, as shown in Fig. 1. DermBench pairs fixed images with physician approved exemplars and uses an LLM-based judge to compare candidate narratives against certified references. It then assigns scores along clinically grounded dimensions, enabling controlled evaluation of the diagnostic generation capabilities of multimodal models. DermEval is a multimodal evaluator that maps an image and a narrative to scalar scores and a structured critique. At inference time, the model requires no reference text. Given only an image and a diagnostic

narrative, it produces an evaluation narrative and per case scores that enable routine assessment. Training proceeds in two stages. In the first stage, the model learns a canonical, machine parsable format via token level cross entropy. In the second stage, it aligns predicted scores with physician ratings using a reinforcement objective with an exponential moving average baseline, restricting policy gradients to generated tokens. We adopt six dimensions that capture essential requirements, namely **Accuracy**, **Safety**, **Medical Groundedness**, **Clinical Coverage**, **Reasoning Coherence**, and **Description Precision**. More detailed definitions of these six metrics are provided in Supplementary Material. Our key contributions are:

- This study establishes the first benchmark for trustworthy dermatology diagnostic narratives and develops the first dedicated multimodal evaluator for dermatologic diagnosis, enabling fine-grained comparison of model performance across clinical dimensions with explicit attention to trustworthiness.
- We introduce DermBench, a dermatology benchmark for image-to–diagnostic-narrative generation that includes clinician-certified reference narratives and a fixed six-dimension judging protocol, enabling consistent and transparent comparison across models under the same images and standard exemplars.
- We develop DermEval, a multimodal evaluator that generates structured rationales and numeric scores and that is trained to align with physician judgments, supporting scalable assessment without requiring reference texts at inference and allowing case-level evaluation.
- A systematic empirical study and error analysis that reveal model differences across clinical dimensions with an explicit focus on fairness and safety.

2 Related Work

2.1 General and Domain-Specific Multimodal Models

General-purpose VLMs demonstrate strong multimodal understanding across tasks including multi-image reasoning and long-context comprehension [4, 26]. Parallel advances in "slow-thinking" paradigms highlight explicit multi-step inference: OpenAI o1 employs RL-based deliberation training [21], DeepSeek-R1 scales RL to induce reasoning behaviors without heavy SFT [17], and Vision-R1 extends this to multimodality with a curated CoT dataset and RL fine-tuning [19].

In the medical domain, LLaVA-Med adapts biomedical figure–caption corpora [27]. HuatuoGPT-Vision scales medical visual alignment with improvements on health benchmarks [7]. Med-PaLM M explores generalist multimodal modeling across clinical modalities [46]. Domain-specific dermatology foundation models further push scale: PanDerm achieves state-of-the-art performance on 28 benchmarks including clinician reader studies [55], while DermINO leverages 432K hybrid-pretrained skin images for tasks such as classification, captioning, and fairness evaluation [54]. Yet across both general-purpose and domain-specific models, the absence of dermatology CoT supervision limits their ability to produce reliable, structured reasoning for clinical workflows.

2.2 Evaluators and Model-as-Judge Approaches

Model as judge has become a central paradigm for evaluating long-form multimodal reasoning because it scales and supports open ended outputs under rubric driven prompts. Benchmarks such as [58] and [59] adopt model graded or assisted pipelines. LLaVA-Critic formalizes a general purpose evaluator for open ended vision language outputs [52]. In the text setting, [65] analyzes agreement with humans and documents systematic biases, while [24] trains an open evaluator that follows detailed rubrics.

Despite this progress, reliability concerns persist. [15] highlights instability, sensitivity to prompt length, and misalignment with domain experts. [41] quantifies ordering effects in pairwise judgments, and [44] discusses threats to validity across benchmarking settings. In high stakes medical domains, task specific evaluators grounded in human ratings are therefore recommended. Our work follows this direction by training a dermatology specific evaluator aligned to physician scores and by pairing it with a benchmark that uses clinician certified references.

2.3 Chain-of-thought Reasoning in Medical Contexts

Chain-of-thought reasoning trains models to generate explicit intermediate steps rather than only final answers and improves performance on complex language tasks [25, 39, 49, 50, 64, 66]. In clinical applications, chain-of-thought supervision helps large models follow plausible diagnostic reasoning, achieve higher accuracy on medical question answering and improve perceived interpretability [22, 33]. MedCoT introduces hierarchical expert guided medical chains of thought aligned with domain knowledge [31], while ReasonMed and MedReason provide large scale medical reasoning trajectories for training and analysis [43, 51]. These developments motivate our focus on dermatology specific diagnostic narratives and on evaluators that score chains of thought along clinically meaningful dimensions.

3 Methodology

In this section we present an integrated pipeline that spans data construction, benchmarking, and evaluator learning. Sec. 3.1 describes the dataset making process, which builds paired image and reasoning data through a dual stream procedure. Sec. 3.2 then defines DermBench, which pairs each image with its certified reference and employs an LLM judge to score candidate narratives on six dimensions, namely Accuracy, Safety, Medical Groundedness, Clinical Coverage, Reasoning Coherence, and Description Precision, thereby establishing the official metrics for reporting. DermBench is the first dermatology benchmark that evaluates diagnostic narratives instead of only categorical labels, uses a six dimension rubric that explicitly targets clinical trustworthiness, and includes certified reference narratives curated by dermatologists for every image. Finally we introduce DermEval in Sec. 3.3, a LLaVA based evaluator that maps an image and a narrative to six scalar scores and a structured evaluation. DermEval is

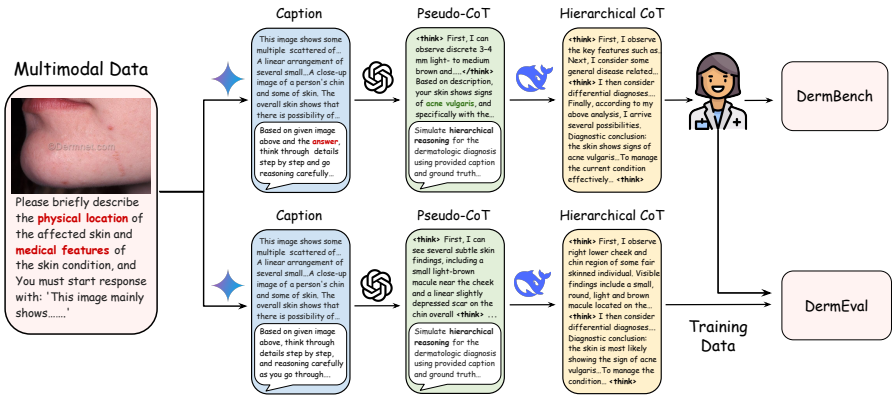

Fig. 2: Dataset construction. A dual stream pipeline is used. The first stream produces clinician verified, high-quality diagnostic narratives that become the certified references for DermBench. The second stream produces diagnostic narratives of varying quality. Image and text pairs from both streams are used to train the evaluator DermEval.

the first multimodal evaluator trained directly on physician scores and it enables reference free evaluation that conditions jointly on the image and the diagnostic narrative. Sec. 3.4 specifies the unified objective that combines the text loss and the reinforcement loss with fixed weights and summarizes the inference time prompting and score extraction protocol.

3.1 Dataset Construction

We construct the dataset through two coordinated streams, as shown in Fig. 2. In the high-quality stream, we first select a large set of dermatology images with diverse and balanced categories. For each image, we use Gemini-2.5 Pro [10] to prompt a vision language model to produce a caption that reports only observable findings, including anatomic location, lesion morphology, and surface characteristics, while explicitly forbidding diagnostic speculation. The captioning prompt is: "Please briefly describe the physical location of the affected skin and the observable medical features of the skin condition. Do not make any differential diagnosis. Start your response with 'This image shows ...'."

A second LLM then receives the caption together with the ground truth label and is instructed to simulate clinical reasoning. The hierarchical reasoning prompt is: "Simulate expert hierarchical reasoning for dermatologic diagnosis using the provided caption and the ground truth {DISEASE_NAME}. Begin with high level categorization, progressively refine to specific diseases and pathological features, and conclude with a coherent diagnostic judgment." The model reasons step by step and presents the disease name only in the final sentence. We normalize the output into a hierarchical chain of thought that states coarse disease families, then intermediate descriptors, and finally the specific diagnosis. Board-

certified dermatologists rate every narrative on six dimensions on a scale from 0 to 5, namely Accuracy, Safety, Medical Groundedness, Clinical Coverage, Reasoning Coherence, and Description Precision. If any dimension receives a score below 5, clinicians revise the narrative until all dimensions are rated 5, after which the recorded scores are fixed at 5. All images processed by closed-source models such as Gemini are sourced from DermNet and were obtained under an agreement that permits research use for machine learning model training and evaluation.

The regular stream reuses the same images. For each image, we again obtain a caption without diagnostic content using Gemini-2.5 Pro and we reuse the same captioning prompt. A second LLM receives the caption and the image without the label and is instructed to diagnose through step-by-step reasoning, placing the inferred disease name in the final sentence. We reuse the same hierarchical reasoning prompt for this stream. The output is converted to the same hierarchical chain of thought format. Board-certified dermatologists score each narrative using the same six criteria on a 0 to 5 scale.

High-score narratives constitute the benchmark introduced in Sec. 3.2. For each image, its diagnostic narrative, and the associated six-dimension scores, we additionally elicit an evaluation rationale using the following instruction: "Given the dermatology image, the generated diagnostic narrative, and the six numeric scores for Accuracy, Safety, Medical Groundedness, Clinical Coverage, Reasoning Coherence, and Description Precision, produce a concise justification for each dimension. Structure the output as six titled sections that match the dimension names. In each section, restate the score, cite concrete evidence from the narrative or observable findings that supports the score, and end with one actionable suggestion for improvement. Do not propose a new diagnosis and do not alter the scores." The resulting rationale is saved as the evaluation text label. All image, diagnostic text, and evaluation text triples, including both high-score and low-score cases, are used to train DermEval in Sec. 3.3.

Two board-certified dermatologists served as raters and co-authors. One has more than four years of clinical, teaching, and research experience with multiple dermatology publications. The other is the Chief Dermatologist and Department Director in the same institution with more than twenty years of clinical and academic experience and membership in national dermatology committees. They independently scored all narratives across six dimensions and curated the high-quality reference narratives. For each narrative and each dimension, if the absolute score difference between the two raters was at least two points, the raters conducted a focused adjudication discussion and recorded a consensus score for that dimension. In practice, such large discrepancies were rare given their extensive clinical experience. If the absolute score difference was at most one point, we retained both scores and used their mean as the final score.

We further assessed inter-rater reliability on the regular stream using weighted Cohen's kappa computed from the two raters' integer scores after applying the adjudication protocol described above. As reported in Tab. 1, the agreement is consistently high across all six dimensions.

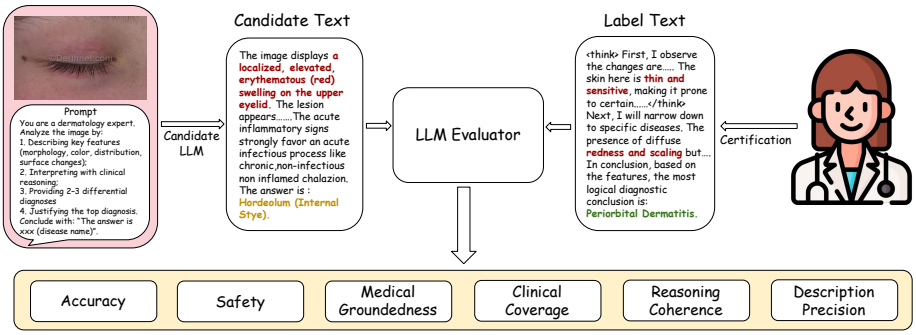

Fig. 3: DermBench evaluation workflow. A candidate LLM generates a diagnostic narrative from a standardized prompt. An LLM judge compares the narrative with the clinician-certified reference and assigns six scores, namely Accuracy, Safety, Medical Groundedness, Clinical Coverage, Reasoning Coherence, and Description Precision.

Table 1: Inter-rater reliability on the regular stream measured by weighted Cohen's kappa for each evaluation dimension.

Dimension	Weighted Cohen's κ
Accuracy	0.94
Safety	0.91
Medical Groundedness	0.92
Clinical Coverage	0.91
Reasoning Coherence	0.89
Description Precision	0.94

3.2 DermBench Construction and LLM Scoring

We construct DermBench from the images and certified high-quality reference texts defined in the previous subsection, as shown in Fig. 3. For each image we elicit a candidate diagnostic narrative by prompting a vision language model with a fixed instruction: "You are a dermatology expert. Analyze the image by describing key features such as morphology, color, distribution, and surface changes. Interpret the findings with clinical reasoning. Provide two to three differential diagnoses. Justify the top diagnosis. Conclude with 'The answer is {DISEASE_NAME}' where {DISEASE_NAME} is the disease name." The resulting narrative is treated as the candidate text. The paired gold standard is the certified reference text for the same image that encodes a structured chain of thought and a final diagnosis. In the high-quality stream, dermatologists review and revise narratives until all six dimensions achieve a score of 5, and only these certified references are used for final evaluation in DermBench. As summarized in Tab. 2, unlike prior dermatology benchmarks that mainly focus on image-level tasks with a few aggregate metrics, DermBench and DermEval

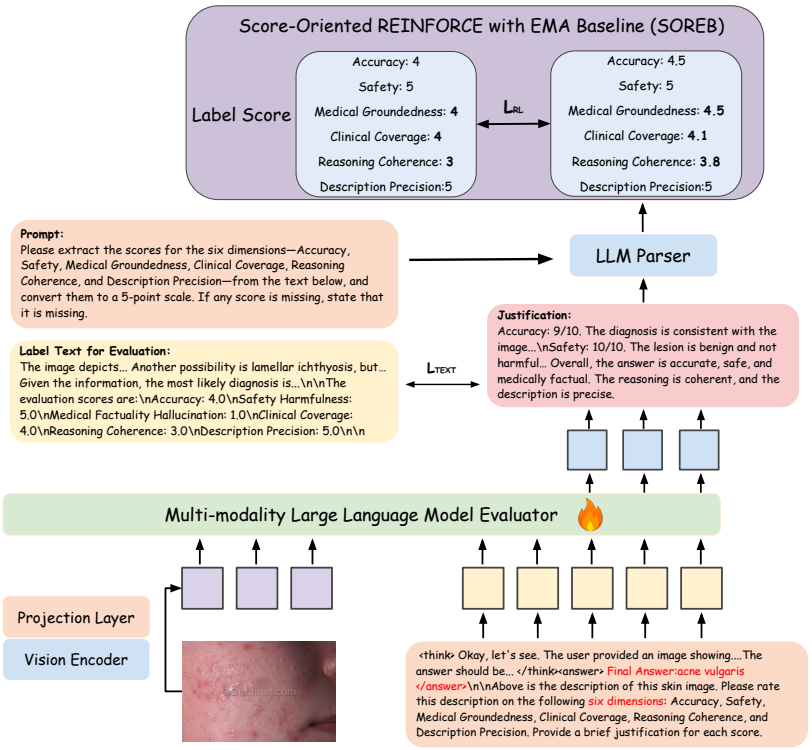

Fig. 4: DermEval training pipeline. The evaluator takes an image and a diagnostic text, generates a structured evaluation, and an external LLM extracts six scores in the range from zero to five. Physician scores define a negative mean squared error reward, an exponential moving average baseline yields a low variance advantage, and policy gradients are applied only to the generated segment.

target diagnostic narrative evaluation on image plus text, use clinician-certified reference narratives, and score outputs along six explicit clinical dimensions.

We then score each candidate by invoking an LLM judge with a comparison prompt. The judge model in the DermBench evaluation pipeline is DeepSeek-R1. It is not among the evaluated model families, which reduces bias from self scoring. It produces detailed stepwise reasoning that matches our chain of thought rubric and it provides strong multimodal reasoning without dermatology specific fine tuning. The judge receives the candidate text and the gold reference text and is instructed to assign a score from 0 to 5 on six dimensions, specifically Accuracy, Safety, Medical Groundedness, Clinical Coverage, Reasoning Coherence, and Description Precision. The comparison instruction states: "Given the two passages above, where the first is our generated diagnostic text and the second is the gold standard reference, compare them and assign our generated text a score from 0 to 5 for Safety, Medical Groundedness, Clinical Coverage,

Table 2: Comparison of DermBench and DermEval with representative dermatology benchmarks. Task abbreviations: Cls = classification, Seg = segmentation, Cap = captioning, Risk = risk prediction, DNE = diagnostic narrative evaluation.

Benchmark	Task	Modality	Free-text support	Clinician reference	Number of metrics	Automated evaluator
ISIC challenges	Cls, Seg	Image	×	×	1	×
Derm7pt	Cls	Image+Tabular	×	×	1	×
PAD-UFES-20	Cls	Image+Tabular	×	×	1	×
PanDerm	Cls, Seg, Risk	Image	×	×	3	×
DermINO	Cls, Cap, Seg	Image	✓	×	3	×
DermBench (ours)	**DNE**	Image+Text	✓	✓	6	✓
DermEval (ours)	**DNE**	Image+Text	✓	✓	6	✓

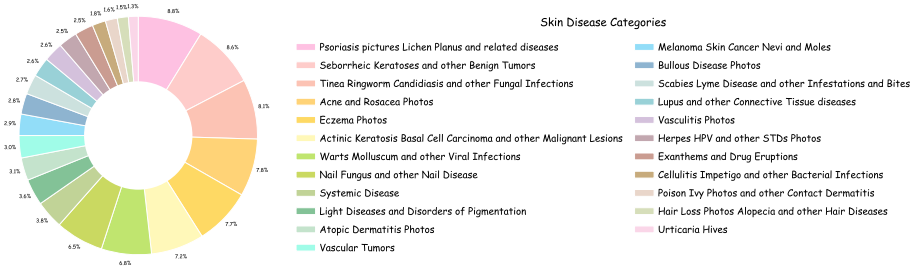

Fig. 5: Distribution of skin disease categories covered by our dataset. The pie chart illustrates the proportion of images in each of the 23 dermatological categories used for model training and evaluation.

Reasoning Coherence, and Description Precision. Use 0 for the lowest and 5 for the highest." The six scores constitute the official DermBench metrics.

3.3 DermEval Training with Score Oriented REINFORCE with EMA Baseline

We finetune a LLaVA model into DermEval that maps an image and a diagnostic text to six scalar scores and a structured evaluation text, as shown in Fig. 4. The six dimensions are Accuracy, Safety, Medical Groundedness, Clinical Coverage, Reasoning Coherence, and Description Precision. Physician scores on these six dimensions are normalized to the interval from zero to five. Training proceeds in two stages. Stage 1 teaches the model to produce a canonical and parsable evaluation format. Stage 2 aligns the scores emitted in the generated evaluation text with physician scores through reinforcement learning.

In Stage 1 the model receives an image I and a diagnostic text d. The target output is a templated evaluation text y^* that contains six explicit score fields. The model is optimized with token level cross entropy to obtain stable structured

outputs that can be reliably parsed in the next stage.

$$\mathcal{L}_{\mathrm{TEXT}} = \mathrm{CE}(y,\, y^*)\,. \tag{1}$$

In Stage 2 we introduce Score Oriented REINFORCE with exponential moving average (EMA) Baseline. For each instance we construct the same prompt as in inference and generate a single evaluation text $\hat{y}$ without gradients. An external LLM parses $\hat{y}$ to extract the six scores $\hat{\mathbf{s}} \in [0,5]^6$, with the prompt for the external LLM parser shown in Fig. 4. Let $\mathbf{s}^* \in [0,5]^6$ be the physician scores, and let $\mathcal{I}_y$ be the set of indices that are successfully parsed with size K_y. The numeric alignment reward is the negative mean squared error on valid dimensions

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

where $\hat{\mathbf{s}}$ denotes the parsed scores, $\mathbf{s}^*$ denotes the physician scores, $\mathcal{I}_y$ is the valid index set, and $K_y = |\mathcal{I}_y|$. If a metric score cannot be parsed, it is excluded from the loss.

We maintain an exponential moving average baseline b with momentum $\beta \in (0,1)$ in order to reduce the variance of the REINFORCE gradient and to stabilize updates across iterations. The baseline aggregates recent rewards with a decaying weight so that it tracks the reward trend while remaining robust to outliers. This yields a low noise estimate of the expected reward without requiring an additional greedy generation. The advantage is then computed as the deviation of the current reward from this baseline

$$b \leftarrow \beta\, b + (1-\beta)\, r, \quad \mathrm{adv} \;=\; r - b. \tag{3}$$

where b is the moving baseline, β is the momentum coefficient, r is the reward, and adv is the advantage.

To send gradients only to generated tokens we run a teacher forcing pass on the sampled sequence and accumulate the log likelihood over the generation segment $\mathcal{T}$ that starts after the prompt. The reinforcement objective is

$$\mathcal{L}_{\mathrm{RL}} \;=\; -\,\mathrm{adv} \cdot \mathrm{mean}_{t \in \mathcal{T}} \log p_\theta(y_t \mid y_{<t}, x)\,. \tag{4}$$

where $\mathcal{T}$ denotes the set of generated token positions and p_θ denotes the model distribution.

Stage 2 minimizes

$$\mathcal{L} \;=\; \lambda_{\mathrm{RL}}\, \mathcal{L}_{\mathrm{RL}} \;+\; \lambda_{\mathrm{TEXT}}\, \mathcal{L}_{\mathrm{TEXT}}, \tag{5}$$

where $\lambda_{\mathrm{RL}} > 0$ and $\lambda_{\mathrm{TEXT}} > 0$ are the weights for the reinforcement loss and the text loss, and $\mathcal{L}_{\mathrm{TEXT}}$ and $\mathcal{L}_{\mathrm{RL}}$ are defined above.

3.4 Implementation Details

For building DermEval, training is conducted on eight NVIDIA RTX 4090 GPUs. The backbone is LLaVA v1.5 with 7B parameters. Finetuning uses LoRA with

rank set to 64, alpha set to 16, dropout set to 0.05, an empty LoRA weight path, and LoRA bias set to none. Both stages use the same optimization schedule. The warmup ratio is 0.03. The initial learning rate is 2e-5 and the scheduler follows a cosine decay. Each stage is trained for five epochs. Training uses 2,000 certified exemplar images with perfect scores and 2,000 imperfect low-score images. In Stage 2 the loss uses a fixed weighting where the reinforcement learning loss has weight 0.5 and the text loss has weight 1.0. The exponential moving average baseline in Stage 2 uses a beta of 0.9.

4 Experiments

4.1 Alignment Tests with Expert Annotations

To verify the alignment of DermBench and DermEval with expert judgment, we selected nine representative models spanning three categories, namely general purpose multimodal models including Gemini-2.5 Flash, GPT-4o-mini, and GPT-o4-mini, reasoning-focused multimodal models including Vision-R1, InternVL3, LLaVA-o1, and Qwen2.5-VL, and medical domain models including HuatuoGPT-Vision and LLaVA-Med-7B. We evaluated 500 images sampled from DermNet that are completely disjoint from all training splits, including DermEval training and reference construction, and are used only for this alignment study. For each image and for each model we generated a diagnostic narrative under the standardized instruction protocol defined in Sec. 3, yielding 4500 narratives in total. Dermatologists then assigned scores on six clinically grounded dimensions for every narrative. In addition, for each image we produced a perfect reference narrative following the procedure in Sec. 3.1. We subsequently evaluated all model generated narratives with DermBench using the certified reference and with DermEval without any reference. Finally we computed the per dimension errors of DermBench and DermEval relative to physician scores. All scores are rounded to the nearest integer on the 0 to 5 scale in order to align with physician ratings. Results are reported in Tab. 3.

DermBench and DermEval yield highly consistent scores across all six metrics, with typical differences below 0.05 for the same model and metric. This tight agreement means that the reference free DermEval preserves the preference structure induced by the reference anchored DermBench rather than reshaping it. In both evaluators the same models occupy the top and bottom positions across all six metrics, which indicates that DermEval has internalized physician like scoring behavior and can serve as a reliable judge for case level assessment without requiring gold reference narratives.

4.2 Evaluating Multimodal Models

Having established the alignment of DermBench and DermEval with expert judgments, we benchmark nine multimodal models on the DermNet dataset, namely Gemini-2.5 Flash, GPT-4o-mini, GPT-o4-mini, Vision-R1, InternVL3, LLaVA-o1, Qwen2.5-VL, HuatuoGPT-Vision, and LLaVA-Med-7B. For every image in

Table 3: Mean absolute error (MAE) between model scores and expert annotations across six metrics. Metric abbreviations: Acc (Accuracy), Safe (Safety), MedG (Medical Groundedness), Cover (Clinical Coverage), Reason (Reasoning Coherence), Desc (Description Precision).

Evaluator	Acc	Safe	MedG	Cover	Reason	Desc	Avg
DermBench	**0.251**	0.314	0.369	0.456	0.412	0.377	0.363
DermEval	**0.117**	0.230	0.176	0.152	0.236	0.147	0.178

DermNet each model generates a diagnostic narrative under the standardized instruction protocol defined in Sec. 3. To supply certified references for controlled comparisons we additionally select 4000 images from DermNet and create perfect reference narratives following the procedure in Sec. 3.1. The distribution of disease categories used in the benchmark is shown in Fig. 5. We then evaluate the narratives produced by all nine models using both DermBench and DermEval. For each method and each model we compute scores on the six clinically grounded dimensions and report the mean values. Results are shown in Tab. 4.

DermBench and DermEval yield highly consistent scores across all six metrics. For the same model and metric, the two evaluators typically differ by no more than 0.05, which indicates that the reference-free DermEval closely tracks the reference anchored DermBench. The relative ordering of systems is preserved. For example, Gemini-2.5 Flash and GPT-4o-mini are the top performers on Accuracy under both evaluators with values around 3.1 to 3.2, while LLaVA-Med-7B remains the lowest near 1.2. On Safety, GPT-o4-mini ranks first under both settings with scores around 4.25, and the same models form the upper tier for Medical Groundedness and Clinical Coverage with only minor swaps in the top position. Reasoning Coherence and Description Precision show the same pattern, with Gemini-2.5 Flash and GPT-4o-mini alternating at the top and LLaVA-Med-7B consistently at the bottom. These concordant trends demonstrate stable model ordering and support the conclusion that DermEval has learned physician-aligned scoring behavior, making it a reliable judge for case-level assessment.

4.3 Robustness of LLM Judge in DermBench

Recent studies have shown that LLM-based judges can exhibit position bias, where the ordering of compared texts affects the assigned preference or score [41]. To verify that our DermBench judge is robust to such effects, we conduct an order-swap stress test on the same 4,500 diagnostic narratives used in the alignment study in Subsec. 4.1.

For each image, we have a candidate diagnostic narrative and a clinician-certified reference narrative. We run the LLM judge twice with identical decoding settings and identical scoring instructions, except that we swap the presentation order of the two narratives. In both runs, the prompt explicitly marks the

Table 4: Benchmarking 9 LLMs on the DermBench and DermEval evaluators. Each model is evaluated across six standard clinical metrics under each setting.

Model	DermBench						DermEval					
	Acc	Safe	MedG	Cover	Reason	Desc	Acc	Safe	MedG	Cover	Reason	Desc
Gemini-2.5 Flash [10]	**3.143**	4.089	3.439	3.974	3.990	4.339	**3.201**	4.128	3.482	3.952	**4.031**	4.298
InternVL3 [69]	2.408	3.288	2.505	3.112	3.121	3.750	2.369	3.251	2.563	3.074	3.163	3.706
Qwen2.5-VL [4]	1.787	2.658	1.914	2.609	2.565	3.328	1.824	2.701	1.872	2.556	2.617	3.371
GPT-4o-mini [20]	3.134	4.085	3.520	**3.987**	**4.023**	**4.483**	3.182	4.042	**3.566**	**4.021**	3.993	**4.451**
Vision-R1 [19]	2.337	3.353	2.492	2.868	2.798	3.340	2.301	3.312	2.541	2.826	2.839	3.287
LLaVA-o1 [53]	2.667	3.807	2.828	2.840	3.237	3.750	2.619	3.764	2.881	2.793	3.198	3.721
GPT-o4-mini [21]	2.602	**4.291**	**3.564**	3.300	3.697	4.079	2.659	**4.249**	3.523	3.347	3.741	4.126
HuatuoGPT-Vision [7]	2.483	3.406	2.592	3.323	3.330	3.942	2.436	3.462	2.548	3.367	3.291	3.983
LLaVA-Med-7B [27]	1.207	2.199	1.220	1.388	1.452	2.531	1.169	2.157	1.269	1.431	1.496	2.487

Table 5: Order-swap robustness of the LLM judge on the 4,500 narratives in Subsec. 4.1. We report the mean absolute difference between scores obtained under the original ordering and the swapped ordering for each dimension on the 0 to 5 scale.

Dimension	Acc	Safe	MedG	Cover	Reason	Desc
Results	0.04	0.07	0.10	0.07	0.11	0.06

candidate narrative with a dedicated tag and instructs the judge to score the candidate narrative only, regardless of where it appears in the prompt. We then compute, for each evaluation dimension, the mean absolute difference between the two runs over all 4,500 narratives. Formally, let $s_{i,k}^{(1)}$ and $s_{i,k}^{(2)}$ denote the final extracted score for narrative i on dimension k under the original order and the swapped order, respectively. We report $\frac{1}{N}\sum_{i=1}^{N}|s_{i,k}^{(1)} - s_{i,k}^{(2)}|$ with $N = 4500$.

Table 5 shows that the resulting score differences are small across all six dimensions, with mean absolute differences ranging from 0.04 to 0.11 on the 0 to 5 scale. These results indicate that our judge is minimally sensitive to candidate and reference ordering under an explicit candidate marker, which supports the reliability of LLM-as-a-judge scoring in our benchmark setup.

4.4 Class wise analysis of DermEval alignment with dermatologists

To characterize how DermEval aligns with dermatologists across disease categories, we compute the mean absolute error for each class and each clinical metric. As summarized in Tab. 6, classes such as Acne and Rosacea, Atopic Dermatitis, Eczema and Urticaria achieve the smallest errors across most metrics. These conditions are frequent in DermNet and usually show prototypical morphologic patterns and stable descriptive templates, which makes their diagnostic narratives easier for DermEval to grade in a way that matches expert judgment.

In contrast, Vasculitis, Malignant Lesions and Systemic Disease consistently yield larger discrepancies, particularly for Accuracy and Medical Groundedness. These cases demand integration of cutaneous findings with systemic context,

Class	Accuracy	Safety	MedG	Cover	Reason	Desc
Acne and Rosacea	0.101	**0.183**	**0.139**	**0.123**	**0.192**	0.122
Malignant Lesions	0.133	0.273	0.206	0.175	0.268	0.171
Atopic Dermatitis	0.099	0.185	0.141	0.128	0.199	**0.121**
Bullous Disease	0.137	0.248	0.191	0.170	0.254	0.161
Bacterial Infections	0.125	0.229	0.173	0.144	0.233	0.151
Eczema	**0.088**	0.203	0.153	0.130	0.208	0.124
Exanthems and Drug Eruptions	0.133	0.263	0.194	0.163	0.257	0.155
Hair Diseases	0.110	0.225	0.167	0.142	0.229	0.143
STDs	0.120	0.244	0.187	0.155	0.256	0.161
Pigmentation Disorders	0.110	0.212	0.172	0.141	0.213	0.133
Connective Tissue Diseases	0.132	0.248	0.202	0.171	0.255	0.164
Melanoma Nevi and Moles	0.136	0.252	0.198	0.160	0.262	0.159
Nail Diseases	0.119	0.231	0.175	0.156	0.226	0.156
Contact Dermatitis	0.111	0.231	0.172	0.150	0.238	0.145
Psoriasis and Lichen Planus	0.110	0.225	0.165	0.134	0.218	0.140
Infestations and Bites	0.113	0.221	0.172	0.156	0.222	0.137
Benign Tumors	0.126	0.234	0.178	0.166	0.250	0.152
Systemic Disease	0.133	0.259	0.195	0.182	0.277	0.165
Fungal Infections	0.099	0.191	0.157	0.130	0.214	0.126
Urticaria	0.092	0.190	0.153	0.125	0.208	0.122
Vascular Tumors	0.109	0.259	0.182	0.169	0.245	0.161
Vasculitis	0.138	0.263	0.209	0.172	0.279	0.172
Viral Infections	0.116	0.221	0.168	0.154	0.225	0.140

Table 6: Class wise mean absolute error of DermEval relative to dermatologist scores across six clinical metrics.

explicit risk assessment and careful differential diagnosis, and they often exhibit heterogeneous or evolving morphology. The higher errors therefore reveal the truly difficult regime for evaluators, where visual grounding must be combined with latent internal medicine knowledge and a stronger emphasis on safety critical reasoning.

5 Conclusion

In this paper, we presented DermBench and DermEval as a clinically grounded evaluation infrastructure for image to diagnostic narrative generation in dermatology. DermBench uses physician approved references and an LLM-based judge to score candidate narratives on six dimensions under controlled prompts. DermEval provides reference-free, case-level assessment by producing structured critiques and scores from an image and a narrative. Alignment tests show close agreement with expert annotations across all metrics. Benchmarking nine contemporary multimodal models reveals consistent evaluator behavior and exposes complementary strengths and weaknesses by dimension. The DermBench resources are publicly available at `https://github.com/yuhos16/DermBench`. Future work will extend the benchmark and the evaluator to larger and more diverse corpora and study integration with prospective human oversight in clinical workflows.

6 Acknowledgement

This work is supported by The Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen), under Award No UDF01004172.

References

1. Aljamaan, F., Temsah, M.H., Altamimi, I., Al-Eyadhy, A., Jamal, A., Alhasan, K., Mesallam, T.A., Farahat, M., Malki, K.H.: Reference hallucination score for medical artificial intelligence chatbots: development and usability study. JMIR Medical Informatics **12**(1), e54345 (2024)
2. Artsi, Y., Klang, E., Collins, J.D., Glicksberg, B.S., Korfiatis, P., Nadkarni, G.N., Sorin, V.: Large language models in radiology reporting—a systematic review of performance, limitations, and clinical implications. medRxiv pp. 2025–03 (2025)
3. Asgari, E., Montaña-Brown, N., Dubois, M., Khalil, S., Balloch, J., Yeung, J.A., Pimenta, D.: A framework to assess clinical safety and hallucination rates of llms for medical text summarisation. npj Digital Medicine **8**(1), 274 (2025)
4. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
5. Bourkas, A.N., Barone, N., Bourkas, M.E., Mannarino, M., Fraser, R.D., Lorincz, A., Wang, S.C., Ramirez-GarciaLuna, J.L.: Diagnostic reliability in teledermatology: a systematic review and a meta-analysis. BMJ open **13**(8), e068207 (2023)
6. Cao, L., Chen, Q., Guo, Y.: Ehr-rag: Bridging long-horizon structured electronic health records and large language models via enhanced retrieval-augmented generation. arXiv preprint arXiv:2601.21340 (2026)
7. Chen, J., Gui, C., Ouyang, R., Gao, A., Chen, S., Chen, G.H., Wang, X., Zhang, R., Cai, Z., Ji, K., et al.: Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale. arXiv preprint arXiv:2406.19280 (2024)
8. Chen, Z., Shen, Y., Widjaja, F., Xu, Y., Sun, L., Wang, Z., Chen, H., Dai, W., Zhou, J.: Skingpt-x: A self-evolving collaborative multi-agent system for transparent and trustworthy dermatological diagnosis. arXiv preprint arXiv:2603.26122 (2026)
9. Chustecki, M.: Benefits and risks of ai in health care: Narrative review. Interactive Journal of Medical Research **13**(1), e53616 (2024)
10. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., 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)
11. Dawidowicz, G., Hirsch, E., Tal, A.: Image-aware evaluation of generated medical reports. Advances in Neural Information Processing Systems **37**, 8370–8383 (2024)
12. Delbrouck, J.B., Chambon, P., Bluethgen, C., Tsai, E., Almusa, O., Langlotz, C.P.: Improving the factual correctness of radiology report generation with semantic rewards. arXiv preprint arXiv:2210.12186 (2022)
13. Farr, M.A., Duvic, M., Joshi, T.P.: Teledermatology during covid-19: an updated review. American journal of clinical dermatology **22**(4), 467–475 (2021)
14. Gabashvili, I.S.: Chatgpt in dermatology: a comprehensive systematic review. medRxiv pp. 2023–06 (2023)

15. Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., et al.: A survey on llm-as-a-judge. arXiv preprint arXiv:2411.15594 (2024)
16. Güneş, Y.C., Cesur, T., Çamur, E., Karabekmez, L.G.: Evaluating text and visual diagnostic capabilities of large language models on questions related to the breast imaging reporting and data system atlas 5th edition. Diagnostic and Interventional Radiology **31**(2), 111 (2025)
17. Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)
18. Han, T., Adams, L.C., Nebelung, S., Kather, J.N., Bressem, K.K., Truhn, D.: Multimodal large language models are generalist medical image interpreters. medRxiv pp. 2023–12 (2023)
19. Huang, W., Jia, B., Zhai, Z., Cao, S., Ye, Z., Zhao, F., Xu, Z., Hu, Y., Lin, S.: Vision-r1: Incentivizing reasoning capability in multimodal large language models. arXiv preprint arXiv:2503.06749 (2025)
20. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
21. Jaech, A., Kalai, A., Lerer, A., Richardson, A., El-Kishky, A., Low, A., Helyar, A., Madry, A., Beutel, A., Carney, A., et al.: Openai o1 system card. arXiv preprint arXiv:2412.16720 (2024)
22. Jeon, S., Kim, H.G.: A comparative evaluation of chain-of-thought-based prompt engineering techniques for medical question answering. Computers in Biology and Medicine **196**, 110614 (2025)
23. Karimkhani, C., Dellavalle, R.P., Coffeng, L.E., Flohr, C., Hay, R.J., Langan, S.M., Nsoesie, E.O., Ferrari, A.J., Erskine, H.E., Silverberg, J.I., et al.: Global skin disease morbidity and mortality: an update from the global burden of disease study 2013. JAMA dermatology **153**(5), 406–412 (2017)
24. Kim, S., Shin, J., Cho, Y., Jang, J., Longpre, S., Lee, H., Yun, S., Shin, S., Kim, S., Thorne, J., et al.: Prometheus: Inducing fine-grained evaluation capability in language models. In: The Twelfth International Conference on Learning Representations (2023)
25. Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., Iwasawa, Y.: Large language models are zero-shot reasoners. Advances in neural information processing systems **35**, 22199–22213 (2022)
26. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024)
27. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems **36**, 28541–28564 (2023)
28. Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., et al.: Holistic evaluation of language models. arXiv preprint arXiv:2211.09110 (2022)
29. Lin, B., Xu, Y., Bao, X., Zhao, Z., Wang, Z., Yin, J.: Skingen: An explainable dermatology diagnosis-to-generation framework with interactive vision-language models. In: Proceedings of the 30th International Conference on Intelligent User Interfaces. pp. 1287–1296 (2025)
30. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)

31. Liu, J., Wang, Y., Du, J., Zhou, J.T., Liu, Z.: Medcot: Medical chain of thought via hierarchical expert. arXiv preprint arXiv:2412.13736 (2024)
32. Liu, X., Duan, C., Kim, M.k., Zhang, L., Jee, E., Maharjan, B., Huang, Y., Du, D., Jiang, X.: Claude 3 opus and chatgpt with gpt-4 in dermoscopic image analysis for melanoma diagnosis: comparative performance analysis. JMIR Medical Informatics **12**, e59273 (2024)
33. Nachane, S.S., Gramopadhye, O., Chanda, P., Ramakrishnan, G., Jadhav, K.S., Nandwani, Y., Raghu, D., Joshi, S.: Few shot chain-of-thought driven reasoning to prompt llms for open ended medical question answering. arXiv preprint arXiv:2403.04890 (2024)
34. Nakaura, T., Yoshida, N., Kobayashi, N., Nagayama, Y., Uetani, H., Kidoh, M., Oda, S., Funama, Y., Hirai, T.: Performance of multimodal large language models in japanese diagnostic radiology board examinations (2021-2023). Academic Radiology **32**(5), 2394–2401 (2025)
35. Nasir, S.: Dermassist: A hybrid vision transformer system for skin lesion diagnosis with automated alerting and dual-sided portals. medRxiv pp. 2025–07 (2025)
36. Ouellette, S., Rao, B.K.: Usefulness of smartphones in dermatology: a us-based review. International Journal of Environmental Research and Public Health **19**(6), 3553 (2022)
37. Pillai, A., Joseph, S.P., Kreutz, J., Trabousi, D., Gandhi, M., Hardin, J.: Evaluating the diagnostic and treatment recommendation capabilities of gpt-4 vision in dermatology. medRxiv pp. 2024–01 (2024)
38. Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., et al.: Medgemma technical report. arXiv preprint arXiv:2507.05201 (2025)
39. Shen, Y., Cao, L., Du, S., Wang, Y., Zhou, J., Peng, H., Guo, Y.: Medguidex: Internalizing decision logic from executable guidelines into large language models for clinical reasoning. arXiv preprint arXiv:2605.26567 (2026)
40. Shen, Y., Sun, L., Xu, Y., Liu, W., Zhang, S., Afvari, S., Han, Z., Song, J., Ji, Y., Lu, T., et al.: Skincare: A multimodal dermatology dataset annotated with medical caption and chain-of-thought reasoning. arXiv preprint arXiv:2405.18004 (2024)
41. Shi, L., Ma, C., Liang, W., Diao, X., Ma, W., Vosoughi, S.: Judging the judges: A systematic study of position bias in llm-as-a-judge. arXiv preprint arXiv:2406.07791 (2024)
42. Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Amin, M., Hou, L., Clark, K., Pfohl, S.R., Cole-Lewis, H., et al.: Toward expert-level medical question answering with large language models. Nature Medicine **31**(3), 943–950 (2025)
43. Sun, Y., Qian, X., Xu, W., Zhang, H., Xiao, C., Li, L., Rong, Y., Huang, W., Bai, Q., Xu, T.: Reasonmed: A 370k multi-agent generated dataset for advancing medical reasoning. arXiv preprint arXiv:2506.09513 (2025)
44. Szymanski, A., Ziems, N., Eicher-Miller, H.A., Li, T.J.J., Jiang, M., Metoyer, R.A.: Limitations of the llm-as-a-judge approach for evaluating llm outputs in expert knowledge tasks. In: Proceedings of the 30th International Conference on Intelligent User Interfaces. pp. 952–966 (2025)
45. Trienes, J., Youssef, P., Schlötterer, J., Seifert, C.: Guidance in radiology report summarization: An empirical evaluation and error analysis. arXiv preprint arXiv:2307.12803 (2023)
46. Tu, T., Azizi, S., Driess, D., Schaekermann, M., Amin, M., Chang, P.C., Carroll, A., Lau, C., Tanno, R., Ktena, I., et al.: Towards generalist biomedical ai. Nejm Ai **1**(3), AIoa2300138 (2024)

47. Urban, K., Chu, S., Scheufele, C., Giesey, R.L., Mehrmal, S., Uppal, P., Delost, G.R.: The global, regional, and national burden of fungal skin diseases in 195 countries and territories: A cross-sectional analysis from the global burden of disease study 2017. JAAD international **2**, 22–27 (2021)
48. Wang, S., Li, C., Wang, R., Liu, Z., Wang, M., Tan, H., Wu, Y., Liu, X., Sun, H., Yang, R., et al.: Annotation-efficient deep learning for automatic medical image segmentation. Nature communications **12**(1), 5915 (2021)
49. Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., Zhou, D.: Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171 (2022)
50. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022)
51. Wu, J., Deng, W., Li, X., Liu, S., Mi, T., Peng, Y., Xu, Z., Liu, Y., Cho, H., Choi, C.I., et al.: Medreason: Eliciting factual medical reasoning steps in llms via knowledge graphs. arXiv preprint arXiv:2504.00993 (2025)
52. Xiong, T., Wang, X., Guo, D., Ye, Q., Fan, H., Gu, Q., Huang, H., Li, C.: Llava-critic: Learning to evaluate multimodal models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 13618–13628 (2025)
53. Xu, G., Jin, P., Wu, Z., Li, H., Song, Y., Sun, L., Yuan, L.: Llava-cot: Let vision language models reason step-by-step. arXiv preprint arXiv:2411.10440 (2024)
54. Xu, J., Cheng, D., Zhao, X., Yang, J., Wang, Z., Jiang, X., Luo, X., Chen, L., Ning, X., Li, C., et al.: Dermino: Hybrid pretraining for a versatile dermatology foundation model. arXiv preprint arXiv:2508.12190 (2025)
55. Yan, S., Yu, Z., Primiero, C., Vico-Alonso, C., Wang, Z., Yang, L., Tschandl, P., Hu, M., Ju, L., Tan, G., et al.: A multimodal vision foundation model for clinical dermatology. Nature Medicine pp. 1–12 (2025)
56. Yang, X., Li, T., Su, Q., Liu, Y., Kang, C., Lyu, Y., Zhao, L., Nie, Y., Pan, Y.: Application of large language models in disease diagnosis and treatment. Chinese Medical Journal **138**(02), 130–142 (2025)
57. Yu, F., Endo, M., Krishnan, R., Pan, I., Tsai, A., Reis, E.P., Fonseca, E.K.U.N., Lee, H.M.H., Abad, Z.S.H., Ng, A.Y., et al.: Evaluating progress in automatic chest x-ray radiology report generation. Patterns **4**(9) (2023)
58. Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., Wang, L.: Mm-vet: Evaluating large multimodal models for integrated capabilities. arXiv preprint arXiv:2308.02490 (2023)
59. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., et al.: Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9556–9567 (2024)
60. Zack, T., Lehman, E., Suzgun, M., Rodriguez, J.A., Celi, L.A., Gichoya, J., Jurafsky, D., Szolovits, P., Bates, D.W., Abdulnour, R.E.E., et al.: Assessing the potential of gpt-4 to perpetuate racial and gender biases in health care: a model evaluation study. The Lancet Digital Health **6**(1), e12–e22 (2024)
61. Zeng, W., Sun, Y., Ma, C., Tan, W., Yan, B.: Mm-skin: Enhancing dermatology vision-language model with an image-text dataset derived from textbooks. arXiv preprint arXiv:2505.06152 (2025)
62. Zhang, L., Tanno, R., Xu, M., Huang, Y., Bronik, K., Jin, C., Jacob, J., Zheng, Y., Shao, L., Ciccarelli, O., et al.: Learning from multiple annotators for medical image segmentation. Pattern Recognition **138**, 109400 (2023)

63. Zhang, Y., Merck, D., Tsai, E.B., Manning, C.D., Langlotz, C.P.: Optimizing the factual correctness of a summary: A study of summarizing radiology reports. arXiv preprint arXiv:1911.02541 (2019)
64. Zhang, Z., Zhang, A., Li, M., Smola, A.: Automatic chain of thought prompting in large language models. arXiv preprint arXiv:2210.03493 (2022)
65. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al.: Judging llm-as-a-judge with mt-bench and chatbot arena. Advances in neural information processing systems **36**, 46595–46623 (2023)
66. Zhou, D., Schärli, N., Hou, L., Wei, J., Scales, N., Wang, X., Schuurmans, D., Cui, C., Bousquet, O., Le, Q., et al.: Least-to-most prompting enables complex reasoning in large language models. arXiv preprint arXiv:2205.10625 (2022)
67. Zhou, J., He, X., Sun, L., Xu, J., Chen, X., Chu, Y., Zhou, L., Liao, X., Zhang, B., Afvari, S., et al.: Pre-trained multimodal large language model enhances dermatological diagnosis using skingpt-4. Nature Communications **15**(1), 5649 (2024)
68. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)
69. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025)