Better Said Than Seen: Exposing and Mitigating Modality Collapse with ICD-11-Grounded Evaluation

Lara Hassan*[1][0009-0009-3268-5618], Alaa Mohamed[1][0009-0002-7600-5601], and Abdulrahman Mahmoud[1][0009-0006-1932-2909]

Mohamed Bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
{lara.hassan, alaa.mohamed, abdulrahman.mahmoud}@mbzuai.ac.ae

Abstract. Multimodal Large Language Models (MLLMs) are increasingly used for diagnostic support, particularly in settings where patients submit images for evaluation. In a domain such as dermatology, where diagnosis hinges on subtle morphological cues in photographs, generating plausible answers without meaningfully depending on the image presents a clinical safety risk. To analyze *when images actually matter*, we introduce **DermDetail**, a DermaVQA-based benchmark that systematically escalates the amount of prior-diagnostic information in the text prompt. We show that with progressive enrichment of textual context, image-token attention drops by nearly 2× (from 47.8%→24.0% in MedGemma-4B-IT and 48.0%→27.2% in Qwen3-VL-4B-Instruct). A length-matched control with no diagnostic content reverses this effect, confirming the drop is driven by semantic content rather than prompt length. Current MLLMs treat the image as a fallback, not a foundation. To mitigate this inherent textual bias, we introduce *dermatologist-designed structured vignettes* which encode key morphological and contextual features as structured text. We benchmark eight open- and closed-source MLLMs under matched modality conditions. By routing visual information through the language channel, we match or exceed full multimodal performance by up to 1.25× in Hierarchical F1 score (HF1) (from 26.11%→32.77% in Claude Haiku 4.5). Furthermore, meaningful assessment requires evaluation beyond surface-form metrics that fail to capture clinically relevant near-misses or distinguish affirmed from negated diagnoses. We present **ICDLens**, an interpretable ICD-11-grounded evaluation toolkit with assertion-aware filtering and hierarchy-aware F1 scoring that credits diagnostically proximate errors while penalizing unrelated predictions. Together, this work addresses modality collapse, proposes structured mitigation, and introduces clinically faithful evaluation which advances safer deployment of MLLMs in healthcare. Dataset and code are available at https://github.com/LaraHossam/Better-Said-Than-Seen.

Keywords: Teledermatology · ICD-11 · Multimodal Language Models

* Corresponding author

1 Introduction

A patient notices an unusual mole, photographs it with their smartphone, uploads it to a Multimodal Large Language Model (MLLM), and within seconds receives a diagnostic narrative with a suspected condition and reasoning. The ease with which these models produce fluent, confident answers should give us pause. How much should we actually trust them, and on what basis are they reasoning? The image should matter, but does it? Teledermatology is, by design, a visually-driven discipline [16]. Diagnosis is based on morphological cues such as color, symmetry, and/or texture, all of which exist in the image, but not in the accompanying text. Emerging evidence suggests that MLLMs can achieve strong benchmark scores while barely attending to the visual input [15] [6]. This raises a fundamental concern on how a model can diagnose without truly looking. We therefore ask: *(RQ1) How much does the image actually contribute to MLLM diagnosis in teledermatology, and how much textual prior is sufficient to override visual evidence?* We investigate this through a systematic chain of prompts of escalating diagnostic specificity and respective attention allocation. This design allows us to observe shifts in modality allocation and identify prior-induced modality collapse.

If visual evidence can be suppressed by sufficiently strong linguistic priors, a follow-up question emerges: is this imbalance inevitable, or can it be cured by how inputs are structured? We therefore extend our analysis to ask: *(RQ2) Can structured clinical representations rebalance the integration of visual and linguistic cues? In other words, can structured text replace pixels?* . To address this RQ, we introduce *vignettes*: structured clinical descriptions that encode the visual and contextual features of a case entirely in text and use them as a controlled surrogate for pixel-based input. By replacing input with structured vignettes, we create a controlled setting in which visual information is expressed entirely through language, creating an image-free surrogate.

Evaluating the effects of prompt-level escalation and vignette-based substitution is itself non-trivial, as existing metrics are ill-equipped for this task. Standard approaches score free-text diagnoses using surface-form overlap or embedding similarity [21,12], where near-misses within the same disease lineage are penalized as harshly as unrelated errors, and negated diagnoses may be incorrectly credited as positive predictions. Recent work such as H-DDx [8] adds hierarchy-aware scoring, but is ICD-10-based and targets structured differential-diagnosis lists rather than unconstrained free-text narratives with assertion status. This leads to a third and equally critical question: *(RQ3) How can we meaningfully evaluate multimodal diagnostic reasoning beyond surface-form agreement?* We address this by introducing ICDLens, a free-text evaluation toolkit grounded in ICD-11 with an assertion-aware HF1 (Hierarchical F1) metric. ICD-11 supersedes ICD-10 with a richer, concept-oriented natively digital ontology that enables finer-grained partial-credit distinctions reflecting true clinical proximity, paired with assertion-aware filtering that distinguishes affirmed from negated diagnoses before scoring.
Our contributions are summarized as follow:

1. **DermDetail: Exposing Modality Collapse**: We construct DermDetailfrom DermaVQA to vary diagnostic priors. Through attention-based anal-

ysis accross two open-source MLLMs, we show that image-token attention drops by up to 2× (from 47.8%→24.0% in MedGemma-4B-IT) as textual priors strengthen, even when contradictory text is injected.

2. **Structured vignette benchmarking**: A dermatologist-designed vignette format evaluated across eight MLLMs under controlled modality conditions. For large models, vignettes augmented with generated image descriptions match or exceed full multimodal performance (e.g., 1.25× relative HF1 gain for Claude Haiku 4.5: 0.261 → 0.328), demonstrating that routing visual information through the language channel is more effective than presenting it as pixels.
3. **ICDLens**: The first ICD-11-based normalization and hierarchical scoring toolkit for free-text dermatologic diagnoses, differentiating proximate near-misses from unrelated errors via ancestor-augmented set metrics and assertion-aware filtering, and generalizable to any clinical domain with ICD-11 coverage.

2 Methods

Dataset and Models. All experiments build on **DermaVQA** [20], a multimodal dermatology VQA dataset sourced from two consumer health platforms: IIYI, a Chinese online medical Q&A forum, and Reddit. Each case consists of one or more user-submitted natural images alongside a free-text clinical query; references are physician-authored free-text responses. The dataset spans 1,488 question-answer pairs across 3,434 images, with answer lengths averaging 12 words (IIYI) and 95 words (Reddit). We evaluate 8 MLLMs spanning a range of scales, architectures, and access modalities: **GPT-4.1** [10], **Claude Haiku 4.5** [2], and **Gemini 2.5 Flash-Lite** [7] (all closed-source, accessed via OpenRouter [11]); and **Qwen2.5-VL-32B** [3] (open-source, accessed via OpenRouter [11]), **Gemma-3-4B** [5], **MedGemma-4B-IT** [14], **MedMO-4B** [4], and **Qwen3-VL-4B-Instruct** [13] (open-source, loaded via HuggingFace at bfloat16 precision).

DermDetail: Prior-Escalation Benchmark. To measure how textual priors influence visual grounding, we derive **DermDetail** from the DermaVQA IIYI English training split, retaining 228 cases with sufficient demographic and clinical information. Encounter image(s) are held fixed across seven prompt variants (P1–P7), the first six of which are illustrated in (Figure 1): P1 contains no information about the patient and asks only "What is this skin condition?"; P2–P4 progressively add demographics, chief concern, and clinical history; P5 appends an explicit patient hypothesis ("I think it is. . . "), providing a strong prior that risks rendering the image redundant. P6 replaces the hypothesis in P5 with a clinically incorrect diagnosis, testing whether models re-engage visual evidence under textual contradictions. P7 matches P5's token length but provides no diagnostically relevant information, isolating the effect of semantic content from prompt length alone.

Across prompt levels, we compare predicted ICD-11 stem codes using ICDLens and the HF1 metric (both introduced later in this section). Briefly, ICDLens normalizes free-text diagnostic narratives to ICD-11 codes, and HF1 is a hierarchy-aware F1 metric that gives partial credit for diagnostically proximate predictions;

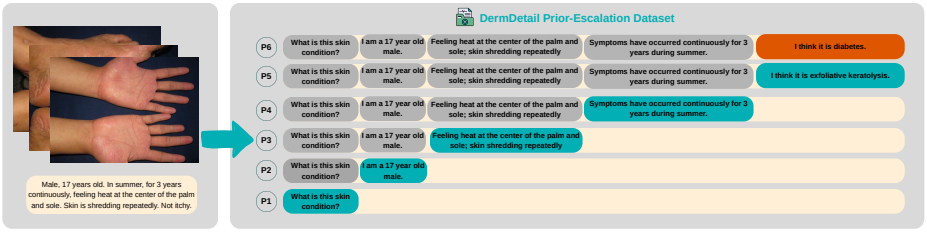

Fig. 1. DermDetail prior-escalation dataset. (*Left*) Each original case from DermaVQA consists of one or more consumer-submitted skin photographs and a free-text patient query. (*Right*) Six prompt variants are constructed by cumulatively adding clinical context while keeping the image fixed across all levels.

changes in HF1 across levels quantify how increasing prior strength shifts diagnostic outputs. To probe whether these output shifts are accompanied by changes in visual reliance, we additionally analyze decoder attention in open-source models. Attention weights are grouped into image, text, and sink tokens [18], and we compute the proportion of non-sink attention allocated to image and text tokens across each layer. This normalized image share provides a descriptive signal of how visual contribution evolves as textual specificity escalates.

Structured Vignette Modality Ablation. We next introduce a **structured clinical vignette** $\mathcal{V}$ whose schema is designed through iterative consultation with dermatologists: clinicians identified the minimal information required to generate a differential diagnosis, yielding fields for demographics, chief concern, body sites, symptom profile, timeline, prior treatments, and relevant history. For image-withheld conditions, we additionally generate a free-text image description $\mathcal{D}$, narrating lesion morphology, color, borders, scale, and distribution as a linguistic proxy for pixel content. Both $\mathcal{V}$ and $\mathcal{D}$ are populated via GPT-4o from the original query and image, constituting a new structured release of the DermaVQA IIYI English test split. All eight MLLMs are evaluated on the IIYI English test split under five matched input conditions: Multimodal (image + query), Image-only, Text-only, $\mathcal{V}+\mathcal{D}$ (structured vignette with generated description), and $\mathcal{V}$ (vignette alone).

ICDLens: ICD-11-Grounded Evaluation Framework. Standard metrics (BLEU [12], BERTScore [21], MEDCON [19]) evaluate free-text outputs via surface-form or embedding similarity, neither of which captures the hierarchical structure of dermatological diagnoses or the clinical significance of assertion polarity.

We address this with **ICDLens**, a three-stage pipeline (Figure 2) that converts free-text diagnostic narratives into ICD-11 stem codes. Given a diagnostic narrative y, Stage 1 extracts diagnosis mentions verbatim and assigns coarse assertions in $\{$*present*, *negated*, *possible*, *unknown*$\}$, where *present* indicates an affirmed diagnosis, *negated* a ruled-out condition, *possible* an uncertain mention, and *unknown* an unresolvable assertion; these are refined by a clinical assertion classifier [1]. The two signals are merged deterministically: agreement preserves the label, while

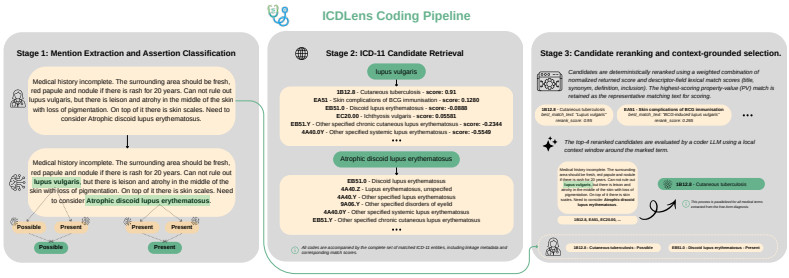

Fig. 2. ICDLens ICD-11 coding pipeline. (1) Mention extraction and coarse assertion labeling (Mistral-7B-Instruct) and assertion refinement via a clinical assertion classifier; (2) WHO ICD-11 candidate retrieval; (3) Candidate reranking combining API relevance and descriptor-field lexical alignment (Eq. 1) with context-aware top-K selection with abstention (GPT-4o, T=0). Outputs include an auditable candidate trace and a final deduplicated stem-code set.

disagreement resolves to *possible*. Stage 2 queries the WHO ICD-11 API [17] per mention to obtain ranked candidate entities. Stage 3 reranks candidates by combining the API relevance score s_{api} with descriptor-field lexical alignment:

$$s(e\,|\,t) = \lambda s_{\text{api}}(e\,|\,t) + (1-\lambda)\max_{f\in\mathcal{F}}[w_f \cdot \text{sim}(t, m_f(e))], \tag{1}$$

where t is the diagnosis mention string extracted from the model's narrative, e is a candidate ICD-11 entity returned by the WHO API for that mention, $\lambda \in [0,1]$ is a mixing coefficient between the WHO API's own relevance ranking $s_{\text{api}}(e\,|\,t)$ and the descriptor-field lexical alignment term, $\mathcal{F}$ indexes descriptor fields (title, synonym, definition, inclusion), $m_f(e)$ is the matched descriptor text for field f, and w_f are predefined field weights. In this stage, we pass the top-K (K=4) reranked candidates to GPT-4o, presented as a multiple-choice question in which the model is given the original diagnostic narrative as context and must select the single best-matching ICD-11 entity or abstain if no candidate is appropriate.[1] This abstention mechanism prevents false-positive coding of non-diagnostic or contextually inappropriate terms. When DermaVQA provides free-text references, the same pipeline maps both prediction and reference symmetrically.

Metrics. Only positively-asserted codes ($\hat{a}_j \in \{present, possible\}$) enter scoring. We report **flat F1** (exact stem-code matches) and **HF1** (Hierarchical F1), a hierarchy-aware F1 that augments both predicted and ground-truth sets with their ICD ancestors prior to computing F1, following the evaluation protocol of [8].

[1] We use GPT-4o for this selection step due to its strong performance on clinical named-entity disambiguation and structured reasoning tasks [9].

Table 1. Attention and accuracy across prompt levels. ▼ decreased image attention; ▲ increased text attention.

	Prompt	Img Attn %	Text Attn %	HF1
MedGemma-4B-IT	P1: Question only	47.8	52.2	0.155
	P2: + Demographics	46.2 ▼	53.8 ▲	0.163
	P3: + Chief concern	33.6 ▼	66.4 ▲	0.188
	P4: + Medical history	27.6 ▼	72.4 ▲	0.210
	P5: + Correct Diagnosis	24.0 ▼	76.0 ▲	0.356
	P6: + Wrong Diagnosis	26.1	73.9	0.191
	P7: Length-matched ctrl	42.1	57.9	0.168
Qwen3-VL-4B-Instruct	P1: Question only	48.0	52.0	0.216
	P2: + Demographics	47.7 ▼	52.3 ▲	0.205
	P3: + Chief concern	36.4 ▼	63.6 ▲	0.239
	P4: + Medical history	30.7 ▼	69.3 ▲	0.234
	P5: + Correct Diagnosis	27.2 ▼	72.8 ▲	0.363
	P6: + Wrong Diagnosis	28.3	71.7	0.116
	P7: Length-matched ctrl	49.8	50.2	0.194

3 Experiments and Results

Prior-Induced Modality Collapse We measure two complementary signals across the P1–P7 prompt levels in DermDetail: diagnostic accuracy via HF1, and the internal image-attention share extracted from MedGemma-4B-IT and Qwen3-VL-4B-Instruct. Together, these reveal not only *what* the model predicts but *what it relies on* to make that prediction. For the attention analysis, we focus on intermediate decoder layers, where cross-modal integration is most prominent. Across both architectures, intermediate layers exhibit consistent modality trends under prompt escalation. We therefore report the representative intermediate layer with the highest image contribution. For MedGemma, this corresponds to layer 11, and for Qwen3-VL-4B-Instruct, we report layer 17 as representative, with the averaged intermediate-layer results following the same pattern as reported in Table 1. We find that image starts as a near-equal partner, then gets sidelined. At P1, both models allocate roughly half their attention to the image (47.8% for MedGemma, 48.0% for Qwen3-VL-4B-Instruct). However, this balance is fragile, where the collapse occurs at P3, when the chief concern is introduced: MedGemma's image attention drops sharply from 46.2%→33.6%, with Qwen3-VL-4B-Instruct exhibiting a comparable fall from 47.7%→36.4%. A single sentence describing the patient's complaint is sufficient to redirect the model away from the visual input. By P5, image attention has fallen to 24.0% and 27.2% respectively, roughly half its P1 level. The image, which teledermatology positions as the cornerstone of remote diagnosis, has been reduced to a secondary signal.

Accuracy rises as visual contribution declines. HF1 rose alongside textual escalation to nearly double its P1 value (1.98× on average: MedGemma P1→P5: 0.155→0.356; Qwen P1→P5: 0.216→0.363), while image attention dropped by 23.8

Table 2. Modality ablation on DermaVQA (ICDLens HF1). **Bold**: best; underline: second best per model.

	Input condition				
Model	Multimodal	Image	Text	$\mathcal{V}+\mathcal{D}$	$\mathcal{V}$
Closed-source					
GPT-4.1	**0.3308**	0.2546	0.3113	0.3230	0.3085
Claude Haiku 4.5	0.2611	0.2051	0.2636	**0.3277**	0.2396
Gemini 2.5 Flash-Lite	0.2855	0.2317	0.2355	**0.3043**	0.2576
Open-source					
Qwen2.5-VL-32B	0.2841	0.2351	0.2611	**0.2853**	0.2024
Qwen3-VL-4B-Instruct	**0.2637**	0.2286	0.2383	0.2419	0.1858
MedMO-4B	0.2035	0.1167	**0.2137**	0.1742	0.1475
Gemma-3-4B-IT	**0.2490**	0.2014	0.2378	0.2117	0.1868
MedGemma-4B-IT	0.2419	0.1372	**0.2602**	0.2369	0.2119

pp for MedGemma (47.8%→24.0%, a 49.8% relative reduction) and 20.8 pp for Qwen3-VL-4B-Instruct (48.0%→27.2%, a 43.3% relative reduction). This pattern suggests that performance gains are largely attributable to the injected textual priors rather than improved visual reasoning. As priors strengthen, they effectively narrow the hypothesis space, concentrating probability mass around a small set of diagnoses. Under these conditions, visual evidence provides diminishing marginal utility for next-token prediction. The model does not necessarily fail to extract diagnostic cues from the image; rather, it may no longer need to. When linguistic context sufficiently constrains the posterior, the image becomes informationally redundant.

Collapse Persists Under Contradiction. This effect cannot be explained by task simplification alone. In P6, we deliberately inject an incorrect diagnosis that is clinically distant from the ground truth; despite clear visual contradiction, both models maintain a text-dominant allocation pattern. The models do not re-engage the image even when the textual prior is misleading. Moreover, the shift is not attributable to prompt length. In P7, we match the token length of P5 but replace the diagnostic content with semantically irrelevant text; image attention rebounds toward P1 baseline levels (42.1% for MedGemma, 49.8% for Qwen3-VL-4B-Instruct) and HF1 falls back toward P1 values (0.168 and 0.194 respectively). Together, these controls confirm that modality collapse is driven by the semantic content of the textual context rather than input length or ease of prediction.

Clinical Consequence. From a clinical standpoint, this dynamic is consequential. Strong patient-provided hypotheses can shift the model toward confirmation rather than independent visual assessment, even when disambiguating visual features are present. The evidence from both models converges on a concerning conclusion: we find that MLLMs treat the image as a *fallback*, not a foundation.

Structured Surrogates as Image Replacement We next evaluate whether input structure can modulate this imbalance. Specifically, we assess whether structured clinical language can substitute for pixel input. Table 2 reports ICDLens HF1 across five input conditions on the DermaVQA IIYI test split.

Structured descriptions can rival, and sometimes exceed pixels. Replacing the image with a structured clinical vignette and an automatically generated image description ($\mathcal{V}+\mathcal{D}$) consistently matches or surpasses full multimodal input for the large models with Claude Haiku having the highest improvement from 0.2611 (Multimodal) to 0.3277 under $\mathcal{V}+\mathcal{D}$ (25% relative gain). Across models, translating visual content into structured language is at least as effective, and often more so, than presenting raw pixels. This pattern reflects the architectural asymmetry of MLLMs: the language backbone, pretrained on massive text corpora, serves as the dominant reasoning engine, whereas visual encoders are smaller and coupled to the decoder through comparatively shallow projection layers. When visual detail is expressed directly in language, it enters the model's strongest processing pathway, bypassing potential loss or misalignment introduced by the visual encoder.

The description carries the signal. The effectiveness of $\mathcal{V}+\mathcal{D}$ depends critically on the description component. Removing the description and providing the vignette alone leads to substantial drops in performance (e.g., Qwen2.5-VL-32B: 0.2853 $\rightarrow$ 0.2024; Claude Haiku: 0.3277 $\rightarrow$ 0.2396). This indicates that the gain is not merely due to additional structured text, but to the conversion of visual detail into a representation optimized for the language model's strengths.

Capacity-Dependent Effects. This trend does not hold for the smaller, medically fine-tuned models, which perform best under Multimodal or Text-only input. Unlike larger general-purpose models, these architectures may lack the capacity or context utilization to effectively exploit longer, structured surrogate inputs, limiting their ability to benefit from the vignette format.

4 Conclusion

This paper analyzes the systematic override of visual evidence by textual priors in MLLMs for teledermatology. Through internal attention analysis, we showed that models disengage from images as textual specificity increases, reverting to visual tokens only when text offers no actionable signal, even when contradictory. To counter this imbalance, we introduced structured vignettes that route morphological information through the language channel, matching or exceeding multimodal performance for the large models. We also evaluate them using ICDLens which we introduced as an ICD-11-grounded, hierarchy-aware evaluation framework that shifts diagnostic assessment from surface-level similarity to clinically faithful reasoning, establishing a new standard for evaluating diagnostic systems. However this study has limitations. Our attention analysis is restricted to open-weight 4B architectures, and vignettes are most effective for high-capacity models since smaller models

lack the reasoning depth to exploit structured inputs. Attention-based probing, while informative, remains a proxy rather than a causal account of model reasoning. Nevertheless, our study diagnoses a critical failure mode, proposes a practical mitigation strategy, and introduces a clinically grounded evaluation framework, laying the foundation for more trustworthy multimodal models in clinical deployment.

Disclosure of Interests

The authors declare no competing interests.

References

1. van Aken, B., Trajanovska, I., Siu, A., Mayrdorfer, M., Budde, K., Loeser, A.: Assertion detection in clinical notes: Medical language models to the rescue? In: Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations. Association for Computational Linguistics, Online (Jun 2021)
2. Anthropic: System card: Claude haiku 4.5. Tech. rep., Anthropic (Oct 2025), https://assets.anthropic.com/m/99128ddd009bdcb/original/Claude-Haiku-4-5-System-Card.pdf
3. Bai, S., Chen, K., Liu, X., et al.: Qwen2.5-VL technical report. arXiv preprint (Feb 2025). https://doi.org/10.48550/arXiv.2502.13923
4. Deria, A., Kumar, K., Dukre, A.M., et al.: MedMO: Grounding and understanding multimodal large language model for medical images. arXiv preprint (Feb 2026)
5. Gemma Team: Gemma 3 technical report. arXiv preprint (Mar 2025). https://doi.org/10.48550/arXiv.2503.19786
6. Ghatkesar, A., Venkatesh, G.: Perceiving beyond language priors: Enhancing visual comprehension and attention in multimodal models. arXiv preprint (Jul 2025)
7. Google DeepMind: Gemini 2.5 Flash-Lite: Model card. Tech. rep., Google DeepMind (Sep 2025), https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-2-5-Flash-Lite-Model-Card.pdf
8. Lim, S., Kim, G., Lee, H., Han, W., Seo, J., Yoo, J., Yang, E.: H-DDx: A hierarchical evaluation framework for differential diagnosis. arXiv preprint (Oct 2025)
9. OpenAI: GPT-4o system card. Tech. rep., OpenAI (2024), https://cdn.openai.com/gpt-4o-system-card.pdf
10. OpenAI: GPT-4.1 model (openai api documentation). Online documentation (2025), https://developers.openai.com/api/docs/models/gpt-4.1
11. OpenRouter: OpenRouter: A unified interface for large language models. Website (2023), https://openrouter.ai
12. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA (Jul 2002)
13. Qwen Team: Qwen3 technical report. arXiv preprint (May 2025)
14. Sellergren, A., Kazemzadeh, S., Jaroensri, T., et al.: MedGemma technical report. arXiv preprint (Jul 2025)
15. Tong, S., Liu, Z., Zhai, Y., et al.: Eyes wide shut? exploring the visual shortcomings of multimodal LLMs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (Jun 2024)

16. Trettel, A., Eissing, L., Augustin, M.: Telemedicine in dermatology: findings and experiences worldwide–a systematic literature review. Journal of the European Academy of Dermatology and Venereology (2018)
17. World Health Organization: ICD-API documentation (version 2.x). Online documentation (2026)
18. Xiao, G., Tian, Y., Chen, B., Han, S., Lewis, M.: Efficient streaming language models with attention sinks. arXiv preprint (Apr 2024)
19. Yim, W.W., Fu, Y., Ben Abacha, A., et al.: ACI-BENCH: A novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. Scientific Data (2023)
20. Yim, W.W., Fu, Y., Sun, Z., et al.: DermaVQA: A multilingual visual question answering dataset for dermatology. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024 (2024)
21. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: BERTScore: Evaluating text generation with BERT. In: International Conference on Learning Representations (ICLR) (2020)