

# When Reasoning Hurts Confidence: Demographic Calibration of Vision-Language Models in Medicine (Extended Abstract)

Sonnet Xu, Roxana Daneshjou

Stanford University  
sonnet@stanford.edu

## Abstract

Vision-language models (VLMs) are increasingly evaluated as diagnostic aids in medical imaging, yet clinical deployment requires confidence estimates that are reliable and equitable across patient demographics. We audit the demographic calibration of GPT-4o, R1-Onevision-7B, and Qwen2.5-VL-7B on dermatology and chest radiograph classification, stratified by skin tone and patient sex, across three confidence signals. In six of nine model-task combinations the best-calibrated signal is not the most equitable, with up to a  $2.6\times$  ECE ratio across skin tones, and temperature scaling reduces ECE by up to 96% without closing equity gaps. A same-backbone comparison shows that chain-of-thought training restructures every confidence signal, so no single method transfers reliably across architectures and tasks.

## Motivation

Deployed medical VLMs are ideally calibrated, meaning that they should be correct about 80% of the time they report 80% confidence, consistent across patient subgroups (Guo et al. 2017). Dermatology and radiology models show documented accuracy disparities across demographic groups (Daneshjou et al. 2022; Seyyed-Kalantari et al. 2021), yet calibration studies rarely stratify by demographics and fairness audits rarely examine uncertainty.

## Study Design

We evaluate GPT-4o (OpenAI 2024), the chain-of-thought reasoning model R1-Onevision-7B, and Qwen2.5-VL-7B (Bai et al. 2025). The open-source models we evaluate share one base architecture, isolating the effect of reasoning training. We measure benign/malignant classification on the 656-image DDI dataset (Daneshjou et al. 2022), stratified by Fitzpatrick skin type, and binary Pneumonia (3.4% prevalence) and Pleural Effusion (27.7%) detection on a 656-image CheXpert subset (Irvin et al. 2019), stratified by patient sex. From one constrained A/B prompt we extract token-level probability (TLP, a softmax over the two option logprobs), verbalized confidence (CE, a self-reported percentage; Xiong et al. 2024), and self-consistency (SC, agreement over 15 stochastic samples; Wang et al. 2023). We report per-group ECE (Naeini, Cooper, and Hauskrecht 2015)

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

| Model  | Dataset   | TLP         | CE          | SC <sub>0.5</sub> | SC <sub>1.0</sub> |
|--------|-----------|-------------|-------------|-------------------|-------------------|
| GPT-4o | DDI       | .035        | .070        | .051              | <b>.009</b>       |
|        | Pneumonia | .004        | .011        | <b>.000</b>       | .005              |
|        | Effusion  | <b>.022</b> | .056        | .035              | .041              |
| R1     | DDI       | .060        | .052        | <b>.038</b>       | .094              |
|        | Pneumonia | <b>.009</b> | .023        | .022              | .039              |
|        | Effusion  | <b>.004</b> | .029        | .021              | .012              |
| Qwen   | DDI       | <b>.022</b> | .109        | .117              | .117              |
|        | Pneumonia | .011        | <b>.004</b> | <b>.004</b>       | <b>.004</b>       |
|        | Effusion  | <b>.030</b> | .034        | .034              | .034              |

Table 1: Max ECE gap across demographic groups per (model, dataset, signal). Bold marks the most equitable signal per row, which coincides with the best-calibrated signal only for Qwen.

and AUROC with bootstrap intervals, summarize equity by the max ECE gap, and apply group-conditional temperature scaling.

## Findings

**Calibration and equity frequently conflict.** In six of nine model-task combinations, all involving GPT-4o and R1, the best-calibrated signal is not the most equitable (Table 1). On DDI, GPT-4o’s CE achieves its lowest mean ECE (0.073) but its largest skin-tone disparity (a  $2.6\times$  ratio), while SC at  $\tau=1.0$  nearly closes the gap at higher mean ECE. Group-conditional temperature scaling cuts aggregate ECE by up to 96% yet leaves the disparities essentially unchanged, and actively worsens anti-discriminative signals.

**Reasoning restructures every signal.** R1’s TLP achieves the strongest DDI discrimination (AUROC 0.690) with the worst calibration (ECE up to 0.456), then turns anti-discriminative on CheXpert (AUROC 0.341-0.436). Qwen shows the converse, with SC collapsing to a constant 1.0, CE to 95% on CheXpert samples, and has TLP as its only discriminative signal. R1’s most reliable signal at a high temperature is SC. Strategies did not transfer across architectures, highlighting that validation must be per signal, task, and group.

## Acknowledgments

Some of the computing for this project was performed on the Sherlock cluster. We thank Stanford University and the Stanford Research Computing Center for providing these resources.

## Ethics Statement

This study audits existing VLMs on two publicly released, de-identified datasets. IRB review was not required as we collected no new patient data.

## References

- Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*.
- Daneshjou, R.; Vodrahalli, K.; Novoa, R. A.; Jenkins, M.; Liang, W.; Rotemberg, V.; Ko, J.; Swetter, S. M.; Bailey, E. E.; Gevaert, O.; et al. 2022. Disparities in dermatology AI performance on a diverse, curated clinical image set. *Science Advances*, 8(31): eabq6147.
- Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On calibration of modern neural networks. In *International Conference on Machine Learning*, 1321–1330. PMLR.
- Irvin, J.; Rajpurkar, P.; Ko, M.; Yu, Y.; Ciurea-Illcus, S.; Chute, C.; Marklund, H.; Haghighi, B.; Ball, R.; Shpan-skaya, K.; et al. 2019. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01): 590–597.
- Naeini, M. P.; Cooper, G. F.; and Hauskrecht, M. 2015. Obtaining well calibrated probabilities using Bayesian binning into quantiles. In *AAAI Conference on Artificial Intelligence*, 2901–2907.
- OpenAI. 2024. GPT-4o system card. Technical report, OpenAI.
- Seyyed-Kalantari, L.; Zhang, H.; McDermott, M. B. A.; Chen, I. Y.; and Ghassemi, M. 2021. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine*, 27(12): 2176–2182.
- Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.; Chi, E.; Narang, S.; Chowdhery, A.; and Zhou, D. 2023. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.
- Xiong, M.; Hu, Z.; Lu, X.; Li, Y.; Fu, J.; He, J.; and Hooi, B. 2024. Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs. *arXiv preprint arXiv:2306.13063*.