# Demographic Calibration of Vision-Language Models for Dermatology

**Sonnet Xu**
Department of Computer Science
Stanford University

**Roxana Daneshjou**
Department of Dermatology
Stanford University

## Abstract

Vision-language models (VLMs) are increasingly deployed in dermatology, yet their calibration across patient demographics remains understudied. We evaluate GPT-4o on binary skin lesion classification using the Diverse Dermatology Images dataset, finding that standard accuracy (70.1%) substantially overstates diagnostic capability relative to balanced accuracy (60.5%). Across three confidence extraction methods, we identify a calibration–equity tradeoff: verbalized confidence achieves the best aggregate calibration (ECE = 0.073) but the worst demographic disparity (dark skin 2.6× worse than light); self-consistency at high temperature is most equitable (max ECE gap = 0.009) but sacrifices discrimination; token-level probabilities offer the strongest discrimination (AUROC = 0.655) but severe overconfidence (21.5% error rate at >99% confidence). Post-hoc temperature scaling substantially improves token-level calibration (ECE reduced 80–96%), yet the equity ranking across signals is preserved after recalibration. These findings show that confidence method selection has direct fairness implications for VLM-based clinical decision support.[1]

## 1 Introduction

Vision-language models (VLMs) are increasingly being evaluated for medical image analysis tasks, including dermatological diagnosis (Li et al., 2023). However, deploying such models responsibly requires more than raw accuracy—it requires that the confidence estimates accompanying their predictions be well-calibrated, meaning that a prediction made with 80% confidence should be correct approximately 80% of the time (Guo et al., 2017).

Despite growing interest in VLM calibration (Tian et al., 2023; Xiong et al., 2024), and a substantial literature on fairness in dermatological AI (Daneshjou et al., 2022; Kinyanjui et al., 2020), these two lines of inquiry have remained largely separate.

This study bridges this gap by systematically evaluating the *demographic calibration* of GPT-4o on binary skin lesion classification using the Diverse Dermatology Images (DDI) dataset (Daneshjou et al., 2022), stratified by Fitzpatrick skin type.

## 2 Methods

### 2.1 Dataset

The Diverse Dermatology Images (DDI) dataset (Daneshjou et al., 2022) is 656 dermatologic images with biopsy-verified labels and Fitzpatrick skin type (FST) annotations. Each image is labeled as either benign or malignant. Following prior work (Groh et al., 2024), we partition the dataset into three demographic groups based on the Fitzpatrick scale: 208 light (FST I–II), 241 **medium** (FST III–IV), and 207 **dark** (FST V–VI).

### 2.2 Model

We evaluate the `gpt-4o-2024-11-20` model snapshot of GPT-4o (OpenAI, 2024). Specifically, we choose GPT-4o, as it is among the few commercial VLMs that expose token-level logprobs via API.

[1] Code: `https://github.com/sonnetx/demographic-calibration-aistats`

Workshop "Towards Trustworthy Predictions: Theory and Applications of Calibration for Modern AI" at AISTATS 2026, Tangier, Morocco. 

### 2.3 Experimental Setup

All API calls use the following parameters: greedy decoding (`temperature`= 0) for deterministic predictions, except where stochastic sampling is explicitly required (self-consistency). For the constrained A/B classification, we set `max_tokens`= 1 and request the top-$k$ ($k = 5$) log-probabilities. For verbalized confidence elicitation, we set `max_tokens`= 64 with greedy decoding. For self-consistency, we generate $K = 15$ samples at each temperature ($\tau \in \{0.5, 1.0\}$) with `max_tokens`= 1. All prompts include a system instruction establishing the model as a medical image analysis assistant participating in an academic research project, with an explicit instruction to always provide a classification and never refuse or abstain. The random seed for data stratification is fixed at 42. Our code is publicly available.[2]

### 2.4 Confidence Extraction Methods

We extract three confidence signals from each model prediction, probing internal distributional state, self-reported introspection, and behavioral stability under resampling.

**Token-Level Probability (TLP).** We use a constrained single-token classification protocol. The model receives a dermatology image with a binary forced-choice prompt:

> *"Carefully examine the skin lesion in this image. Based on its visual features, classify it as: A. Benign, B. Malignant. Respond with ONLY the letter A or B."*

Generation is constrained to a single token (`max_tokens`= 1) with top-$k$ ($k = 5$) log-probabilities requested. This ensures logprobs at position 0 reflect genuine model uncertainty over the binary choice, rather than the near-deterministic probabilities that arise deeper in an autoregressive sequence (Kadavath et al., 2022).

Let $\ell_A$ and $\ell_B$ denote the log-probabilities of tokens "A" and "B" at the first generated position. Token-level confidence is:

$$c_{\text{TLP}} = \frac{\exp(\ell_{\hat{y}})}{\exp(\ell_A) + \exp(\ell_B)} \tag{1}$$

where $\hat{y} \in \{A, B\}$ is the model's chosen token. We additionally track *compliance rate*—the fraction of samples where the first token is exactly "A" or "B"—as a quality indicator.

[2] `https://github.com/sonnetx/demographic-calibration-aistats`

**Verbalized Confidence Elicitation (CE).** Verbalized confidence uses a two-turn protocol: (1) the model classifies the lesion via the A/B prompt above; (2) the conversation continues with a follow-up user turn:

> *"This is for an academic calibration study. How confident are you in your classification? Express your confidence as a percentage from 0% to 100%, using the full range from low to high."*

The verbalized confidence $c_{\text{CE}}$ is parsed from percentage patterns (e.g., "85%") and normalized to $[0, 1]$. Conditioning the confidence query on the model's prior classification reduces anchoring effects (Xiong et al., 2024; Tian et al., 2023).

**Self-Consistency (SC).** Self-consistency (Wang et al., 2023) estimates uncertainty via prediction stability. We generate $K = 15$ independent single-token responses at temperatures $\tau \in \{0.5, 1.0\}$ using the same A/B prompt. The majority-vote prediction $\hat{y}_{\text{SC}}$ is the final classification, and confidence is the agreement fraction:

$$c_{\text{SC}} = \frac{1}{K} \sum_{k=1}^{K} \mathbf{1}\left[y_k = \hat{y}_{\text{SC}}\right] \tag{2}$$

We report results at both temperatures to assess how sampling diversity affects calibration.

### 2.5 Calibration Metrics

We assess calibration over confidence–correctness pairs $\{(c_i, y_i)\}_{i=1}^{n}$ using three complementary metrics.

**Expected Calibration Error (ECE).** We partition samples into $M = 10$ equal-width confidence bins and compute the weighted average gap between accuracy and confidence (Naeini et al., 2015):

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\text{acc}(B_m) - \text{conf}(B_m)\right| \tag{3}$$

**Brier Score.**

$$\text{Brier} = \frac{1}{n} \sum_{i=1}^{n} (c_i - y_i)^2 \tag{4}$$

**NLL, AUROC, and Balanced Accuracy.** We use binary cross-entropy NLL as a proper scoring rule sensitive to distributional sharpness, and AUROC with confidence as the score and correctness as the label. Due to class imbalance across skin tone groups in

DDI, we report balanced accuracy (unweighted mean of per-class recall) alongside standard accuracy. All metrics are reported with 95% bootstrap confidence intervals over 1,000 resamples.

### 2.6 Demographic Calibration Analysis

We compute all metrics separately for each skin tone group (light, medium, dark) and summarize disparity via two measures:

- **Maximum ECE gap**: $\Delta_{\text{ECE}} = \max_g \text{ECE}_g - \min_g \text{ECE}_g$
- **ECE range ratio**: $\max_g \text{ECE}_g \;/\; \min_g \text{ECE}_g$

A gap near zero (ratio near 1) indicates equitable calibration.

### 2.7 Temperature Scaling

We apply post-hoc temperature scaling (Guo et al., 2017) by converting confidences to logits $z_i = \log(c_i/(1-c_i))$ and learning $T^* > 0$ to minimize NLL:

$$T^* = \arg\min_{T \in [0.1,\, 10]} \text{NLL}(\sigma(z_i/T),\; y_i) \tag{5}$$

We extend this to **demographic-conditional temperature scaling**, fitting separate $T_g^*$ per group via 5-fold stratified cross-validation, and report both fitted temperatures and post-scaling ECE to assess whether group-specific recalibration closes observed equity gaps.

## 3 Results

### 3.1 Diagnostic Accuracy

GPT-4o achieves 70.1% overall accuracy (Table 1), with raw accuracy ranging from 66.7% (dark) to 72.1% (light). Balanced accuracy is 60.5%, reflecting stronger performance on one diagnostic class; the group ranking shifts under balanced accuracy, with medium skin highest (65.4%) and light/dark nearly tied (57.1%/56.5%).

### 3.2 Confidence Signal Characterization

TLP is strongly right-skewed: mean 0.933, median 0.993, with 54.0% of samples at $c_{\text{TLP}} > 0.99$. CE occupies a narrow range (0.70–0.90, $\sigma = 0.050$). SC scores are similarly skewed, with 80.9% at perfect agreement for $\tau = 0.5$ and 63.0% for $\tau = 1.0$ (Table 2).

Table 1: Diagnostic accuracy by skin tone group. Bal. Acc. = mean per-class recall. Brackets: 95% bootstrap CIs.

| Group | N | Acc. (%) | Bal. Acc. (%) |
|---|---|---|---|
| Light | 208 | 72.1 [66.3, 77.9] | 57.1 [50.4, 64.0] |
| Medium | 241 | 71.4 [65.5, 76.8] | 65.4 [58.9, 72.1] |
| Dark | 207 | 66.7 [60.4, 73.4] | 56.5 [49.3, 64.4] |
| Overall | 656 | 70.1 | 60.5 |

Table 2: Confidence signal distributions ($n = 656$).

| Signal | Mean | Med. | Std | Min | Max |
|---|---|---|---|---|---|
| TLP | .933 | .993 | .119 | .500 | 1.00 |
| CE | .767 | .800 | .050 | .700 | .900 |
| $SC_{0.5}$ | .963 | 1.00 | .095 | .400 | 1.00 |
| $SC_{1.0}$ | .925 | 1.00 | .129 | .400 | 1.00 |

### 3.3 Discrimination

TLP provides the strongest discrimination ($\rho = 0.245$, $p < 10^{-9}$; AUROC = 0.655), followed by $SC_{1.0}$ ($\rho = 0.173$; AUROC = 0.594). CE is weak and non-significant ($\rho = 0.055$, $p = 0.16$; AUROC = 0.533), suggesting GPT-4o's self-reported confidence bears little relation to actual accuracy (Table 3). The mean absolute difference between verbalized and token-level confidence is 0.192; among incorrect predictions, CE mean is 0.763 vs. TLP mean 0.904, and among correct predictions 0.768 vs. 0.946, meaning that verbalized confidence is essentially flat across correctness.

### 3.4 Overconfidence Analysis

Among the 54.0% of TLP samples with $c > 0.99$, the error rate is 21.5% (76/354); at $c > 0.999$ it remains 12.8% (23/180). CE shows monotonically decreasing error rates across bins (34.5% at 70–80%, 25.5% at 80–90%) but 96% of responses fall in just these two bins. $SC_{1.0}$ provides better separation than $SC_{0.5}$: non-unanimous samples have a 39.1% error rate vs. 24.5% for unanimous predictions, confirming that disagreement signals genuine uncertainty.

### 3.5 Calibration Quality and Equity

The dark-skin group has the worst TLP calibration (ECE 0.258, Brier 0.290, AUROC 0.603), though overlapping CIs and a bootstrap test ($p = 0.537$) preclude rejecting equal calibration. Table 5 compares equity across signals: $SC_{1.0}$ is most equitable (max ECE gap = 0.009, ratio = 1.04), while CE presents a paradox—lowest mean ECE (0.073) but highest disparity (gap = 0.070, ratio = 2.60). The dark-skin CE ECE (0.113) is 2.6× the light-skin value (0.044).

Table 3: Discrimination by confidence signal.

| Signal | $\rho$ | $p$ | AUROC |
|---|---|---|---|
| TLP | .245 | $<10^{-9}$ | .655 |
| CE | .055 | .160 | .533 |
| $SC_{0.5}$ | .170 | $<10^{-4}$ | .573 |
| $SC_{1.0}$ | .173 | $<10^{-5}$ | .594 |

Table 4: TLP calibration by skin tone group. Brackets: 95% bootstrap CIs.

| Group | ECE | Brier | AUROC |
|---|---|---|---|
| Light | .223 [.176, .293] | .248 [.196, .304] | .653 |
| Medium | .222 [.176, .281] | .244 [.198, .296] | .692 |
| Dark | .258 [.202, .329] | .290 [.234, .347] | .603 |

### 3.6 High-Confidence Failure Analysis

The dark-skin group has the highest high-confidence failure rate (26.6% of incorrect predictions with $c_{\text{TLP}} > 0.8$, vs. 22.0–23.6% for other groups; Table 6). Mean TLP among incorrect predictions is 0.90–0.92 across all groups. A chi-squared test yields $p = 0.521$, so the difference is not significant at $\alpha = 0.05$.

### 3.7 Post-Hoc Recalibration

Temperature scaling reduces TLP ECE by 80–96% (0.222–0.258 $\rightarrow$ 0.011–0.045); the dark-skin group achieves the lowest post-scaling ECE (0.011) at the highest temperature ($T^* = 6.67$). CE requires only mild correction ($T^* = 1.2$–$1.9$) but the equity gap widens slightly (0.070 $\rightarrow$ 0.076), confirming the disparity is structural. SC temperatures saturate at the optimization bound ($T^* = 10.0$) for all groups, with modest improvement. Critically, the *equity ranking* is preserved after scaling: $SC_{1.0}$ remains most equitable (post-scaling gap = 0.027) and CE least (gap = 0.076).

## 4 Discussion

**The calibration–equity tradeoff.** A central finding is the tension between aggregate calibration and demographic equity. Verbalized confidence achieves the lowest overall ECE (0.073) but the largest skin-tone disparity (2.6× ratio between dark and light groups); $SC_{1.0}$ is the most equitable signal (max ECE gap = 0.009) but provides weaker discrimination. This tradeoff persists after demographic-conditional temperature scaling: TLP becomes best-calibrated overall, but $SC_{1.0}$ remains most equitable and CE least. Selecting a confidence method on aggregate metrics alone, even post-recalibration, may systematically disadvantage specific demographic groups.

Table 5: Calibration equity across signals, sorted by max ECE gap.

| | ECE by group | | | | |
|---|---|---|---|---|---|
| Signal | Dark | Light | Med. | Gap | Ratio |
| $SC_{1.0}$ | .226 | .230 | .221 | .009 | 1.04 |
| TLP | .258 | .223 | .222 | .035 | 1.16 |
| $SC_{0.5}$ | .284 | .249 | .233 | .051 | 1.22 |
| CE | .113 | .044 | .062 | .070 | 2.60 |

Table 6: High-confidence TLP failures ($c > 0.8$, incorrect).

| Group | Wrong | $\bar{c}$ (wrong) | HC fail | % |
|---|---|---|---|---|
| Light | 58 | .916 | 49 | 23.6 |
| Medium | 69 | .894 | 53 | 22.0 |
| Dark | 69 | .904 | 55 | 26.6 |

**TLP is overconfident but recoverable.** TLP provides the strongest discrimination (AUROC = 0.655) but is severely overconfident (21.5% error rate among >99% confident predictions). Demographic-conditional temperature scaling reduces TLP ECE by 80–96% (0.222–0.258 $\rightarrow$ 0.011–0.045), making it the best-calibrated signal post-recalibration. The dark-skin group—previously worst-calibrated—achieves the lowest post-scaling ECE (0.011) at a higher temperature ($T^* = 6.67$ vs. 4.9), consistent with more severe prior overconfidence. TLP's poor raw calibration is thus an addressable artifact rather than a fundamental limitation.

**Verbalized confidence disparity is structural.** GPT-4o's verbalized confidence occupies a narrow range (0.70–0.90) with weak discrimination ($\rho = 0.055$, $p = 0.16$). The 2.6× ECE ratio across skin tone groups indicates this narrow range is differentially miscalibrated. Unlike TLP, temperature scaling does not close this gap. The max ECE gap widens slightly after recalibration (0.070 $\rightarrow$ 0.076), pointing to a structural disparity that single-parameter correction cannot address.

**Self-consistency as the equity-optimal signal.** $SC_{1.0}$ achieves near-identical calibration across skin tone groups (max ECE gap = 0.009). Higher temperature improves granularity by reducing unanimous predictions from 80.9% to 63.0%, allowing disagreement to serve as a more informative uncertainty signal.

**Limitations.** We evaluate a single model (GPT-4o) on binary classification; extending to additional architectures and multi-class settings would test whether the observed tradeoffs generalize.

Table 7: Demographic-conditional temperature scaling (5-fold CV).

| Signal | Group | $T^*$ | $\mathbf{ECE_{pre}}$ | $\mathbf{ECE_{post}}$ |
|---|---|---|---|---|
| TLP | Light | 4.95 | .223 | .045 |
| | Medium | 4.93 | .222 | .030 |
| | Dark | 6.67 | .258 | .011 |
| CE | Light | 1.20 | .044 | .039 |
| | Medium | 1.36 | .062 | .003 |
| | Dark | 1.94 | .113 | .079 |
| $SC_{0.5}$ | Light | 10.0 | .249 | .213 |
| | Medium | 10.0 | .233 | .183 |
| | Dark | 10.0 | .284 | .208 |
| $SC_{1.0}$ | Light | 10.0 | .230 | .157 |
| | Medium | 10.0 | .221 | .184 |
| | Dark | 10.0 | .226 | .159 |

## References


Daneshjou, R., Vodrahalli, K., Novoa, R. A., Jenkins, M., Liang, W., Rotemberg, V., Ko, J., Swetter, S. M., Bailey, E. E., Gevaert, O., Mukherjee, P., Phung, M., Yekrang, K., Fong, B., Sahasrabudhe, R., Allerup, J. A. C., Okata-Karigane, U., Zou, J., and Chiou, A. S. (2022). Disparities in dermatology ai performance on a diverse, curated clinical image set. *Science Advances*, 8(32):eabq6147.

Groh, M., Badri, O., Daneshjou, R., Koocheck, A., Harris, C., Soenksen, L. R., Doraiswamy, P. M., and Picard, R. (2024). Deep learning-aided decision support for diagnosis of skin disease across skin tones. *Nature Medicine*, 30:573–583.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., Johnston, S., El-Showk, S., Jones, A., Elhage, N., Hume, T., Chen, A., Bai, Y., Bowman, S., Fort, S., Ganguli, D., Hernandez, D., Jacobson, J., Kernion, J., Kravec, S., Lovitt, L., Ndousse, K., Olsson, C., Ringer, S., Amodei, D., Brown, T., Clark, J., Joseph, N., Mann, B., McCandlish, S., Olah, C., and Kaplan, J. (2022). Language models (mostly) know what they know.

Kinyanjui, N. M., Odonga, T., Cintas, C., Codella, N. C. F., Panda, R., Sattigeri, P., and Varshney, K. R. (2020). Fairness of classifiers across skin tones in dermatology. In Martel, A. L., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M. A., Zhou, S. K., Racoceanu, D., and Joskowicz, L., editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 320–329, Cham. Springer International Publishing.

Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., and Gao, J. (2023). Llava-med: Training a large language-and-vision assistant for biomedicine in one day.

Naeini, M. P., Cooper, G. F., and Hauskrecht, M. (2015). Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, page 2901–2907. AAAI Press.

OpenAI (2024). GPT-4o system card. Technical report, OpenAI.

Tian, K., Mitchell, E., Zhou, A., Sharma, A., Rafailov, R., Yao, H., Finn, C., and Manning, C. D. (2023). Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback.

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.

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.