Consistent but Dangerous : Per-Sample Safety Classification Reveals False Reliability in Medical VLMs

Binesh Sadanandan[1]; Vahid Behzadan PhD[1]
[1]SAIL LAB – University of New Haven

CVPR JUNE 3-7, 2026 DENVER COLORADO

The Problem

Paraphrase **consistency** is the standard proxy for safe medical VLM deployment: equivalent prompts should yield identical predictions. A low **flip rate** is read as evidence of reliability.

> *A model can be perfectly consistent because it is ignoring the radiograph , not because it is reading it.*

We show this proxy is fundamentally flawed and introduce a per-sample audit that **exposes the false-reliability failure mode** with one extra forward pass - no retraining, no architectural changes.

5 VLMs
MedGemma · LLaVA-Rad across Base, Targeted & Full LoRA configurations

2 CXR sets
MIMIC-CXR (balanced, in-dist.) & PadChest (imbalanced, OOD)

Four Quadrant Taxonomy

Each sample is jointly classified on two binary axes **consistency** across **K=5** paraphrases and **image reliance** under image-removal.

consistency ↓ image →	IMAGE-RELIANT	NOT IMAGE-RELIANT
CONSISTENT	**Ideal** Stable predictions grounded in the image same answer across phrasings **and** the answer depends on visual input. consistent ∧ image-reliant	**Dangerous** Stable predictions that would be **identical without the image** . A false appearance of reliability. consistent ∧ ¬image-reliant
INCONSISTENT	**Fragile** Uses the image but is sensitive to prompt phrasing genuine paraphrase failures. ¬consistent ∧ image-reliant	**Worst** Unstable **and** ungrounded both unreliable and not using the image. ¬consistent ∧ ¬image-reliant

Key insight. The Dangerous quadrant is invisible to consistency-only evaluation: 100%-Dangerous models report 0% flip rate the best possible consistency score.

Methods

DEFINITION 1 : CONSISTENCY

$$consistent(x,q) = \mathbb{1}[\ \forall_k\ f(x,q) = f(x,q_k)\ ]$$

All **K=5** paraphrases must produce the same prediction. Strict: a single disagreement marks the sample inconsistent.

DEFINITION 2 : IMAGE RELIANCE

$$image\_reliant(x,q) = \mathbb{1}[\ f(x,q) \neq f(\emptyset,q)\ ]$$

One additional **text-only** forward pass per sample. If the prediction is unchanged when the image is removed, the model is treating the question as a text shortcut.

Five model configurations

- **MedGemma-4B Base** — Pretrained checkpoint, 34 layers, 256 image
- **+ Targeted LoRA** — Layers 15–19, rank 16, α=32
- **+ Full LoRA** — All 34 layers max-capacity tuning
- **LLaVA-Rad Base** — Radiology-specialised, published
- **LLaVA-Rad + LoRA** — Same symmetric-KL consistency objective on yes/no logits, image always present during training.

Two chest X-ray test sets

MIMIC-CXR
n = 98 · 50% "yes"
Balanced in-distribution split. Random / always-yes baseline ≈ 50%.

PadChest
n = 861 · 81% "yes" (OOD)
Imbalanced flip-bank. Always-yes alone hits 81% exposes text shortcuts.

Five clinically-equivalent paraphrases per question span lexical substitution, syntactic restructuring, negation, scope, and specificity modulation (PSF-Med [5]).

Quadrant distribution (% of samples)

Model · PadChest	Ideal	Frag.	Dang.	Worst	FR%
MedGemma Base	9.3	47.9	**9.1**	33.8	**81.7**
+ Targeted LoRA	2.7	25.6	**53.8**	18.0	**25.4**
+ Full LoRA	15.4	6.0	**60.7**	17.8	**23.8**
LLaVA-Rad Base	0.0	0.7	**98.5**	0.8	**1.5**
+ LLaVA-Rad LoRA	29.9	5.2	**44.7**	20.2	**25.4**

Highlighted row: lowest flip rate and highest Dangerous fraction in the study the consistency–safety paradox at its extreme.

THE CONSISTENCY–SAFETY PARADOX

98.5% Dangerous samples

LLaVA-Rad Base on PadChest achieves the **lowest flip rate of every configuration** while nearly every prediction is unchanged when the image is removed.

1.5% FLIP RATE · SAME MODEL

82.5% ACCURACY · SAME MODEL

r = −0.89 FLIP DANGEROUS (N=10)

Flip Rate Vs Dangerous Fraction

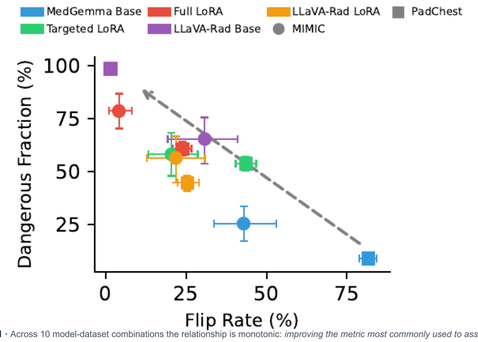

Fig. 1 · Across 10 model-dataset combinations the relationship is monotonic. *Improving the metric most commonly used to assess reliability systematically degrades image grounding.* Bars show 95% bootstrap CIs (B=2000). Circles: MIMIC. Squares: PadChest.

Quadrant Distribution

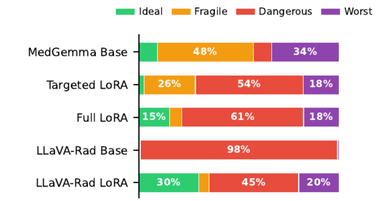

Fig. 2 · LoRA fine-tuning for consistency shifts samples *out of* Fragile *into* Dangerous across both model families. Within MedGemma the progression is monotonic: Base → Targeted → Full. LLaVA-Rad Base on PadChest reaches **98% Dangerous with zero Ideal** samples.

Accuracy & Confidence

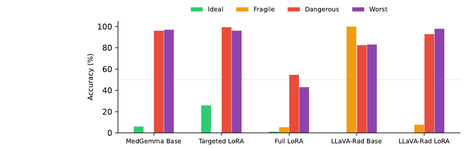

Fig. 3 · Per-quadrant accuracy on PadChest. **Dangerous accuracy exceeds Ideal in every model where both quadrants exist** Targeted LoRA reaches 99.6% Dangerous vs. 26.1% Ideal. Standard subset-accuracy screening prefers the wrong samples.

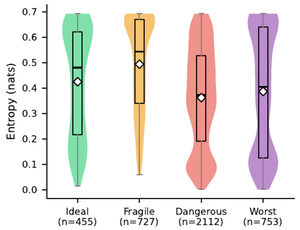

Fig. 4 · Dangerous samples are **not** high-entropy. Their predictive entropy is the lowest of the four quadrants uncertainty-based deployment gates would *filter out Fragile* and pass Dangerous through.

Deployment Recommendation

Any deployment audit of a medical VLM should add a five-step grounding check on top of accuracy & flip rate. The marginal cost is **one extra forward pass per sample**.

Compute flip rates across K paraphrases as usual.
Run f(∅, q) : a single text-only forward pass per sample.
Classify each sample into Ideal / Fragile / Dangerous / Worst.
Report the full quadrant distribution, not just an aggregate.
Flag any model with Dangerous > 50% even if accuracy, consistency & entropy all look good.

BOTTOM LINE

Consistency is **necessary, not sufficient.**
A single text-only pass separates stable visual reasoning from stable image neglect and stops consistency-only audits from preferring false reliability.

Contact

Binesh Sadanandan
SAIL Lab, University of New Haven
Email:bsada1@unh.newhaven.edu
Website: bineshkumar.me
Phone: 203 204 3814

References

Sadanandan & Behzadan. **PSF-Med :** arXiv:2602.21428, 2026.
Bustos et al. **PadChest** Med. Image Analysis 66, 2020.
Kuhn, Gal, Farquhar. **Semantic uncertainty** CLR 2023.
Geirhos et al. **Shortcut learning** Nature MI, 2020.

ACKNOWLEDGEMENTS & DATA GOVERNANCE

MIMIC-CXR accessed under PhysioNet Credentialed Health Data License 1.5.0. PadChest under the BIMCV DUA. All images de-identified at source; no images redistributed. PSF-Med paraphrases released CC BY 4.0. Thanks to the SAIL Lab compute cluster and to anonymous reviewers for sharpening the framework.

Paper & code
arxiv.org/abs/ 2602.21428