PSF-Med: A Clinician-Audited Benchmark for *Paraphrase Sensitivity* in Medical VLMs

Binesh Sadanandan[1], Vahid Behzadan[1] · Lekshmy Jayan[2] · Arun Gopinatha Kurup[3]
[1]SAIL Lab, Univ. of New Haven, CT · [2]Mount Zion Medical College Hospital, Kerala · [3]Badr Al Samaa Hospitals, Muscat

CVPR JUNE 3-7, 2026 DENVER COLORADO

1. The Problem

Medical Vision-Language Models (VLMs) can reverse a diagnostic answer when a clinician rephrases the same question. A model might answer "yes" to "Is there evidence of pleural effusion?" but "no" to "Can pleural effusion be identified?" on the same chest X-ray (Figure 1). We call this Paraphrase Sensitivity Failure (PSF). Standard benchmarks test fixed prompts and cannot detect it. Worse, optimizing for consistency can reduce visual grounding rather than improve image understanding.

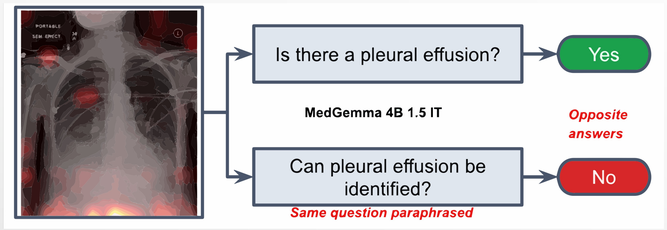

Figure 1: Paraphrase Sensitivity Failure. The same chest X-ray produces opposite answers under rephrasing

A natural hypothesis is that flips occur when paraphrases are semantically distant from the original. Cosine similarity and Euclidean distance distributions for flip and non-flip pairs overlap substantially. The failure must arise from how models process linguistic structure, not surface-level dissimilarity.

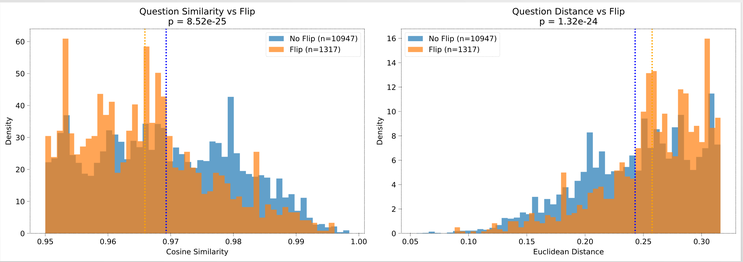

Figure 2: Embedding distance distributions for flip vs. non-flip pairs

2. Methods and Materials

STAGE 1
BioClinicalBERT cosine pre-filter
Retain pairs with cosine > 0.90, threshold set on a 100-pair pilot (94% pilot agreement).

coarse filter

STAGE 2
Cross-family LLM judge
GPT-5-mini adjudicates every pair against a 10-category clinical equivalence rubric Bidirectional-entailment standard.

122,778 pairs (99.5%)

STAGE 3
Clinician-audited filter
1,200 pairs reviewed via Label Studio by a **board-certified radiologist** , a **practicing clinician**, and the **lead author** (medical-device researcher). Blinded; 2-of-3 majority.

load-bearing test

RETENTION UNDER RUBRIC · CORE-EQUIVALENT

72.0% MIMIC	73.7% PADCHEST	78.5% VINDR-CXR

Cross-family validation with Claude Haiku 4.5 agrees with GPT-5-mini at **91.6%** and is strictly more conservative , the rubric is not GPT-family specific.

Clinician-audit agreement

COMPARISON	RAW AGREEMENT	COHEN'S K
Radiologist vs. clinician	98.5%	0.9693
Radiologist vs. author	98.5%	0.9693
Clinician vs. author	98.5%	0.9693
Consensus vs. LLM judge	**80.5%**	**0.6085**

TABLE 1 *κ adjusts raw agreement for the high base rate of equivalent labels. Three-class agreement is unanimous on **391 / 400** pairs (97.75%). Disagreements concentrate at the uncertain boundary, not equivalent vs. non-equivalent.*

3. Results

Headline: flip rates across four medical VLMs

MODEL	MIMIC ACC	MIMIC FLIP	PADCHEST ACC	PADCHEST FLIP	VINDR ACC	VINDR FLIP
MedGemma-4B	**88.3**	**6.7**	49.4	13.1	57.2	12.7
MedGemma-27B	76.7	10.6	45.6	17.8	57.6	12.4
CheXOne [10]	65.4	13.3	51.8	**10.3**	47.3	14.7
RadFM [9]	33.8	17.2	n/a	27.0	48.2	**40.4**

*Flip rates span **6.7%** → **40.4%** a 6× gap. The 7×-larger MedGemma-27B does not improve over the 4B variant; scale alone is not the lever.*

4. Discussion

The validity-anchoring test

If the LLM judge is over-aggressive, headline flip rates would shift when we restrict to **clinician-confirmed** pairs. They don't.

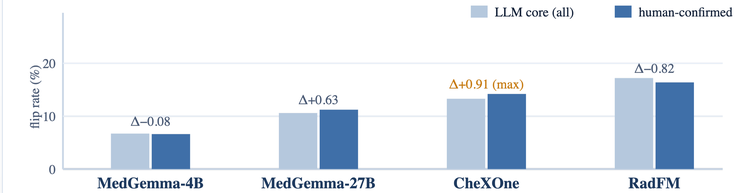

FIG 3 *Bars sit within **1 pp** for every model; max shift **0.91 pp (CheXOne)**, median **0.45 pp**. **Rankings preserved exactly.***

Stability ≠ grounding

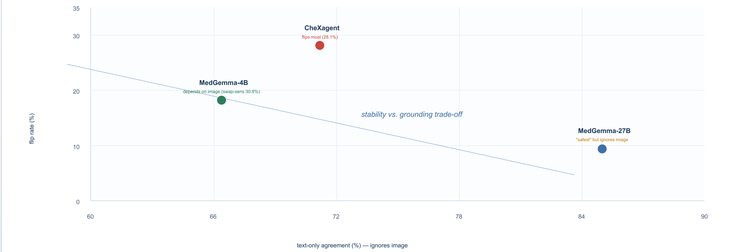

FIG 2 *N = 2,499 MIMIC-CXR presence questions. MedGemma-27B has the lowest flip rate (9.4%) and the highest text-only agreement (85.0%) — its answers stay the same with or without the image. **Flip rate alone is a false sense of***

Key Takeaways:

✓ **Consistency is a false safety signal.** A stable model often gets there by ignoring the image.

✓ **One text-only forward pass catches it.** No architecture changes, no new image runs.

✓ **A tiny LoRA fixes most of it.** Tuning 0.1% of parameters gives the best consistency-grounding balance.

Contact
Binesh Sadanandan
SAIL Lab – University of New Haven
bsada1@unh.Newhaven.edu
Webbineshkumar.me/psf-med-gallery

References

1. Alsentzer *et al.* Publicly available clinical BERT embeddings. *NAACL ClinNLP*, 2019.
2. Asadi *et al.* MIRAGE: The illusion of visual understanding. *arXiv:2603.21687*, 2026.
3. Bustos *et al.* PadChest: A large chest x-ray image dataset. *MedIA*, 2020.
4. Chen *et al.* CheXagent: Towards a foundation model for chest x-ray interpretation. *arXiv:2401.12208*, 2024.
5. Goldberger *et al.* PhysioNet. *Circulation*, 2000.
6. Johnson *et al.* MIMIC-CXR. *Scientific Data*, 2019.
7. Nguyen *et al.* VinDr-CXR. *Scientific Data*, 2022.
8. Tkachenko *et al.* Label Studio. 2020–2025.
9. Wu *et al.* RadFM. *arXiv:2308.02463*, 2023.

BENCHMARK
huggingface.co/ saillab/psf-med

CODE
github.com/UNHSAILLab/ psf-med

FLIP GALLERY
bineshkumar.me/ psf-med-gallery