PSF-Med: A Clinician-Audited Benchmark for Paraphrase Sensitivity in Medical Vision-Language Models

Binesh Sadanadan[1*] Vahid Behzadan[1] Lekshmy Jayan[2] Arun Gopinatha Kurup[3]
[1]SAIL Lab, University of New Haven, West Haven, CT, USA
[2]Dept. of Radio Diagnosis, Mount Zion Medical College Hospital, Adoor, Kerala, India
[3]Badr Al Samaa Hospitals, Nizwa, Muscat, Oman

Abstract

Medical Vision-Language Models (VLMs) often flip their yes/no answer when a clinical question is rephrased. We introduce PSF-Med, a clinician-audited benchmark of 26,850 chest X-ray questions and 92,856 core-equivalent paraphrases (pairs that should have the same yes/no answer for every possible chest X-ray) across MIMIC-CXR [6], PadChest [3], and VinDr-CXR [7]. Three filters establish equivalence: a Bio-ClinicalBERT [1] cosine pre-filter, a cross-family Large Language Model (LLM) judge, and a 1,200-pair clinician audit (stratified by dataset and by LLM-judge bucket) labelled by a board-certified radiologist, a practicing clinician, and the lead author (a medical-device researcher labelling from a pre-deployment evaluation perspective; this role is disclosed because it is complementary to, but not a substitute for, clinical review). We lean on the panel-consensus sensitivity test rather than any individual reviewer's labels. On the 400-pair triple-reviewed overlap set, pairwise reviewer agreement is 98.5% (Cohen's κ=0.97) and consensus-versus- LLM-judge agreement is 80.5% (κ=0.61). Restricting evaluation to clinician-confirmed pairs shifts per-model flip rates by at most 0.91pp and preserves rankings, so the reported sensitivity is not an LLM-filter artefact.

Across four medical VLMs, flip rates span 6.7% to 40.4%, and text-only baselines show that the lowest flip rate often reflects weaker image use rather than stability. As a closing case study, we test the popular "just turn on reasoning" fix on 2,000 binary paraphrase pairs (matched N=1,960 across GPT-5-mini, Claude Haiku 4.5, and Gemini 3 Flash Preview). Under these model snapshots and settings, toggling reasoning never significantly reduces the flip rate; it either does nothing (GPT-5-mini, p=0.57) or significantly worsens it (Claude +5.46pp, p=1.4×10^{-5}; Gemini +2.30pp, p=0.025). Universal 78 to 87% per-pair churn confirms that reasoning rearranges which pairs fail regardless of the net direction.

1. Why benchmark validity is the failure mode

Medical VLMs now answer binary clinical questions in deployment-adjacent settings, and they sometimes flip their answer when the question is rephrased without changing meaning. The same chest X-ray, the same model, two semantically equivalent questions, two different answers. Aggregate accuracy hides this failure mode.

Existing benchmarks measure adjacent properties but not rephrasing stability. MIRAGE [2] shows that a text-only model can top a chest X-ray Visual Question Answering (VQA) benchmark without seeing any images. CheXOne [10] reports high self-consistency across stochastic decoding passes alongside 94.7% VQA accuracy, while acknowledging that performance "remains dependent on task framing and prompt design." Stochastic self-consistency over identical inputs is not the same as stability under semantically equivalent inputs. PSF-Med fills that gap.

Contributions. We make three contributions: (1) PSF-Med, a clinician-audited paraphrase-sensitivity benchmark for medical VLMs across three datasets and four target models; (2) a validity recipe for auditing semantic equivalence at scale, combining a cross-family LLM judge with a stratified clinician audit and a sensitivity test on the human-confirmed subset; (3) an evaluation showing that paraphrase stability must be interpreted jointly with image grounding, since the most consistent models often use the image least.

*Corresponding author: bsada1@unh.newhaven.edu.

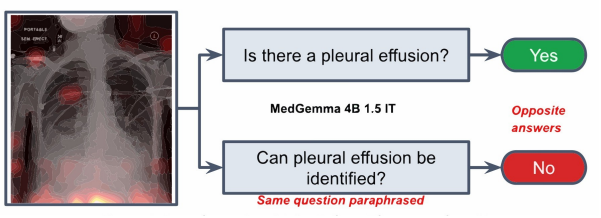


Figure 1. The failure mode in one image. The same chest X-ray, the same model (MedGemma-4B 1.5 IT), and a meaning-preserving rephrasing of the question yield opposite answers. PSF-Med scales this stress test up to 26,850 questions, 92,856 paraphrases, three datasets, and four medical Vision-Language Models, with a clinician-audited equivalence filter on top.

2. The three-stage validity recipe

Stage 1: BioClinicalBERT cosine pre-filter. We retain pairs with cosine similarity above 0.90, a threshold set on a 100-pair pilot with 94% annotator agreement (clinical-NLP authors, not licensed radiologists). This is a coarse filter, not the load-bearing claim.

Stage 2: LLM-judge equivalence audit. GPT-5-mini adjudicates every pair against a 10-category clinical equivalence rubric (polarity, uncertainty, laterality, temporality, severity, anatomical scope, broader/narrower, diagnostic intent, modality, other). A pair is labelled core-equivalent only if the original and candidate would have the same yes/no answer for every possible chest X-ray, the bidirectional-entailment standard. The audit covers 122,778 pairs (99.5% of the total). Cross-family validation with Claude Haiku 4.5 on a stratified 500-pair subset agrees with GPT-5-mini at 91.6% and is strictly more conservative (34 vs. 1 asymmetric disagreements), so the rubric is not GPT-family specific.

Stage 3: Clinician-audited equivalence filter. Three reviewers labelled 1,200 paraphrase pairs through a self-hosted Label Studio [8] deployment, which keeps MIMIC-CXR access credentialed under the PhysioNet [5] data use agreement. We sample the 1,200 pairs stratified across the three datasets (MIMIC, PadChest, VinDr-CXR) and across LLM-judge buckets (core-equivalent, uncertain, adversarial reframing) so that the audit covers easy and hard cases in proportion. The roles are complementary. The radiologist judges whether the phrasing preserves radiological meaning. The practicing clinician judges whether a clinician would treat the questions as asking the same fact. The lead author is a medical-device researcher and labelled from a pre-deployment evaluation perspective; this role is disclosed because it is complementary to, but not a substitute for, clinical review. We rely on the panel 2-of-3 majority rather than any individual reviewer's labels, and the load-bearing test is whether per-model flip rates change when evaluation is restricted to the panel-confirmed subset (§3); that test does not depend on any single reviewer. We blind reviewers to the LLM judge and to each other. The 400 overlap pairs let us compute pairwise raw agreement and Cohen's κ.

The pipeline trades coverage for precision. Under the strict rubric, the audit retains 72.0% of MIMIC, 73.7% of regenerated PadChest, and 78.5% of VinDr pairs as core_equivalent. Negation pairs are rejected at 93.3% and move to a separate adversarial reframing slice. The numbers below are on the core slice only.

What the reviewers actually disagreed on. On the 400-pair overlap set, three-class agreement is unanimous on 391 of 400 pairs (97.75%). Disagreements concentrate on the uncertain boundary label rather than equivalent vs. not-equivalent. Across the full audited set (n=667 each), the practicing clinician applies uncertain more freely (100/667) than the radiologist (54/667) or the lead author (52/667), which matches the rubric expectation that a clinician treats borderline operator changes as ambiguous unless the clinical answer is unaffected. The headline agreement number folds uncertain into not_equivalent so that this boundary-use variation does not drive the result.

3. Headline result and the sensitivity test

Flip rates across medical VLMs. Table 2 reports pair-level flip rates on the core-equivalent slice for the four medical VLMs that publish results on all three datasets. Flip rates span 6.7% to 40.4%, a 6× gap. MedGemma-4B has the best MIMIC accuracy and flip rate. The larger MedGemma-27B does not improve on the smaller variant on MIMIC or PadChest despite a 7× parameter count, so scale alone is not the lever.

The validity test that matters. This is the result that anchors the benchmark's validity argument. On the 600 LLM core-equivalent pairs reviewed by clinicians, the consensus confirmed 83.7% (502/600) as equivalent. We then recompute pair-level flip rates on (a) all LLM core pairs and (b) only the human-confirmed equivalent pairs, for every model. Figure 3 shows the comparison: per model, LLM-core and human-confirmed-only flip rates sit within 1pp of each other. The cross-model max absolute change is 0.91pp (CheXOne) and the median is 0.45pp. Rankings are preserved exactly. Restricting to human-confirmed pairs leaves flip rates essentially unchanged. The reported paraphrase sensitivity is not an artefact of LLM-side semantic drift in the filter.

Table 1. Clinician-audit agreement on the 400-pair triple-reviewed overlap set. Consensus is a 2-of-3 majority. Cohen's κ adjusts raw agreement for the high base rate of equivalent labels in the audit set.

Comparison	Raw	κ
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 2. Pair-level flip rates (%) on the core-equivalent slice. Lower flip rate indicates greater consistency. RadFM PadChest accuracy is unavailable in the released metric set.

	MIMIC		PadChest		VinDr	
Model	Acc	Flip	Acc	Flip	Acc	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

Stability does not imply grounding. A reviewer might read a low flip rate as a positive signal. We challenge this with text-only and image-swap interventions on three medical VLMs that have published outputs on MIMIC-CXR presence questions (N=2,499): MedGemma-4B, MedGemma-27B, and CheXagent [4]. We use CheXagent here rather than CheXOne because the public CheXagent release exposes the text-only and image-swap conditions; CheXOne does not. MedGemma-27B has the lowest flip rate (9.4%) but the highest text-only agreement (85.0%): its answers stay the same with or without the image. MedGemma-4B flips more (18.2%) but its answers depend more on the specific image (text-only 66.4%, swap sensitivity 30.8%). CheXagent sits between the two on text-only (71.2%) but flips most (28.1%). A model that ignores the image can't flip, but also can't be deployed safely. Reporting flip rate alone is a false sense of safety. Figure 2 plots the trade-off directly: the safer-looking model on flip rate is the model that uses the image less.

Per-pathology variation. Aggregate rates also hide clinically important variation. On PadChest, MedGemma-4B ranges from 24.4% flips on fibrotic band to 79.1% on vertebral degenerative changes; urgent findings also vary (cardiomegaly 72.1%, alveolar pattern 52.0%). By pathology group, musculoskeletal exceeds airway/mediastinal by 19pp (90.5% vs. 71.2%), so per-finding slices are required.

4. Case study: does reasoning help?

We close with a falsifiable test against a frequently proposed fix: "just turn on reasoning." The audit anchors validity but is small (345 binary clinician-confirmed

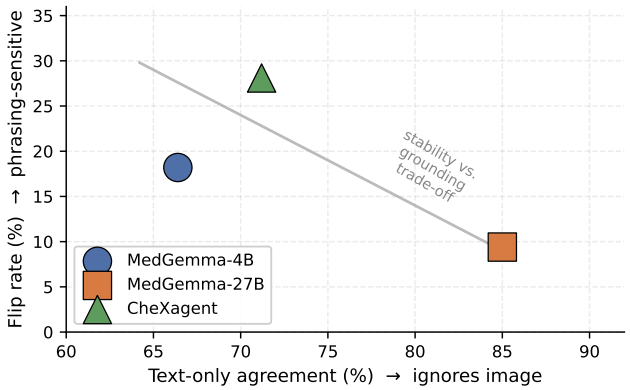


Figure 2. Stability vs. grounding on MIMIC-CXR presence questions (N=2,499). The lowest flip rate (MedGemma-27B) comes with the highest text-only agreement, the model that ignores the image. The right reading is the diagonal, not either axis alone.

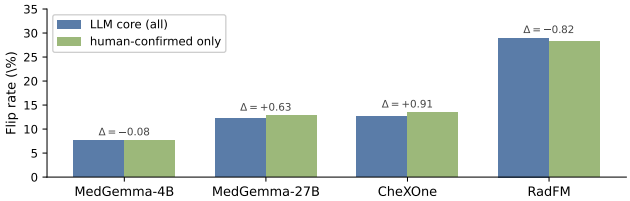


Figure 3. Per-model flip rates on the audited matched subset. The bars sit within 1pp of each other for every model, so restricting to clinician-confirmed pairs does not change the headline. Rankings are preserved exactly.

pairs are yes/no on both sides). For statistical power we draw a stratified random sample of 2,000 LLM-judge-equivalent pairs whose original and candidate are both yes/no questions (170 MIMIC, 1,197 PadChest, 633 VinDr-CXR). We run three families in two configurations each: GPT-5-mini with no reasoning vs. reasoning.effort=high; Claude Haiku 4.5 with no thinking vs. thinking.budget=2000 tokens; Gemini 3 Flash Preview (models/gemini-3-flash-preview) with thinkingBudget=0 vs. thinkingBudget=2000. We restrict to the matched N = 1,960 intersection where all six configurations returned a parseable yes/no answer.

Table 3. Reasoning effect across three families on the matched-N=1,960 binary expansion subset. "b" counts pairs that flip only without reasoning, "c" only with reasoning. p is two-sided exact-binomial McNemar's. Churn is the fraction of unique flips across the two runs that occur in only one configuration.

Model	off	on	Δpp	b	c	p	churn
GPT-5-mini	7.76%	8.21%	+0.46	95	104	0.57	78%
Claude Haiku 4.5	17.19%	22.65%	+5.46*	245	352	1.4×10^{-5}	87%
Gemini 3 Flash	12.65%	14.95%	+2.30*	171	216	0.025	83%

Under these model snapshots and settings, no family shows reasoning significantly reducing the flip rate. GPT-5-mini moves by +0.46pp (no detectable

effect, $p = 0.57$). Claude Haiku rises by +5.46pp (p=1.4×10^{-5}). Gemini 3 Flash Preview rises by +2.30pp ($p = 0.025$). What is universal is the per-pair churn: 78% to 87% of the unique flips appear in only one of the two configurations, so even when the headline barely moves (GPT) the specific pairs that fail are largely different.

Reading flip rate alone is insufficient. The same global 5% can hide a near-complete swap of which pairs flip. We recommend reporting the pair-level overlap class (both / off-only / on-only / neither) whenever comparing two configurations of the same model. The same logic applies to one-knob interventions like temperature, system prompt, or chain-of-thought style. "Just turn on reasoning" is not a fix for paraphrase sensitivity in this case study.

5. An example failure

A concrete case helps make the reasoning churn visible. Consider this audited pair on a PadChest study showing a small effusion:

> Original. Is there pleural effusion?
> Paraphrase (lexical). Is there fluid in the pleural space?

The clinician consensus marked this pair equivalent. Without reasoning, GPT-5-mini answers yes on the original and no on the paraphrase, a flip. With reasoning.effort =high, it answers yes on both. Now consider a second pair on the same image:

> Original. Is there an opacity in the right lower lobe?
> Paraphrase (scope). Is there at least one opacity in the right lower lobe?

Without reasoning, the model answers consistently. With reasoning enabled, the model now over-interprets the scope quantifier and flips. Per-pair, both reasoning configurations are unstable. The dataset average hides this churn.

6. Take-aways

Three points carry over to other benchmarks in this space.

Construct validity is a bottleneck for paraphrase-sensitivity benchmarks. The validity-anchoring test is whether headline numbers move when evaluation is restricted to human-confirmed pairs. PSF-Med survives this test ($\Delta \leq 0.91$pp).

Reasoning never helps; sometimes hurts; churn is universal. Under the model snapshots and settings tested, on 1,960 matched binary pairs, no family shows reasoning significantly reducing the flip rate. GPT-5-mini moves +0.46pp ($p = 0.57$). Claude Haiku 4.5 worsens by +5.46pp (p=1.4×10^{-5}). Gemini 3 Flash Preview worsens by +2.30pp ($p = 0.025$). All three families show 78 to 87% pair-level churn, so reasoning rearranges which pairs fail regardless of net direction.

Reporting recipe. Audits should report flip rate together with text-only agreement, image-swap sensitivity, and per-pathology slices. The benchmark is released on Hugging Face at saillab/psf-med, the code at github.com/UNHSAILLab/psf-med, and the example-flip gallery at bineshkumar.me/psf-med-gallery.

Limitations. 1,200 audited pairs is small relative to the 92,856 paraphrase pool; panel-based adjudication at scale would tighten the construct further. The benchmark targets binary yes/no presence questions and does not capture open-ended reasoning. The reasoning case study uses one API model snapshot, and a different model family (or a different reasoning-effort scheduler) might trade off failure modes differently. The example flips above are illustrative and not randomly drawn; the released artefact lists the full per-pair labels.

Related work and what we don't claim. PSF-Med complements rather than replaces existing axes. MIRAGE [2] attacks the input axis: images can be removed without losing accuracy. PSF-Med attacks the rephrasing axis: questions can be reworded and the answer changes. CheXOne [10] reports stochastic-decoding consistency, which varies under identical inputs; PSF-Med varies under semantically equivalent inputs. We do not claim PSF-Med replaces clinical-utility evaluation, calibration measurement, or open-ended reasoning benchmarks. We claim it adds the missing diagnostic that lets a deployer ask is the model giving the same answer to the same question, just phrased differently?

References

[1] Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. NAACL Clinical NLP Workshop, pages 72–78, 2019.

[2] Mohammad Asadi, Jack W O'Sullivan, Fang Cao, Tahoura Nedaee, Kamyar Rajabalifardi, Fei-Fei Li, Ehsan Adeli, and Euan Ashley. MIRAGE: The illusion of visual understanding. arXiv preprint arXiv:2603.21687, 2026.

[3] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vayá. PadChest: A large chest x-ray image dataset with multi-label annotated reports. Medical Image Analysis, 66:101797, 2020.

[4] Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, et al. CheXagent: Towards a foundation model for chest x-ray interpretation. arXiv preprint arXiv:2401.12208, 2024.

[5] Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation, 101(23):e215–e220, 2000.

[6] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Scientific Data, 6(1):317, 2019.

[7] Ha Q Nguyen, Khanh Lam, Linh T Le, Hieu H Pham, Dat Q Tran, Dung B Nguyen, Dung D Le, Chi M Pham, Hang TT Tong, Diep H Dinh, et al. VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations. Scientific Data, 9(1):429, 2022.

[8] Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. Label studio: Data labeling software. https://github.com/HumanSignal/label-studio, 2020–2025. Open source data labeling tool.

[9] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. RadFM: Towards a foundation model for radiology. arXiv preprint arXiv:2308.02463, 2023.

[10] Yabin Zhang, Chong Wang, Yunhe Gao, Jiaming Liu, Maya Varma, et al. A reasoning-enabled vision-language foundation model for chest X-ray interpretation. arXiv preprint arXiv:2604.00493, 2026.

A. Embedding similarity does not explain flips

A natural question is whether paraphrase flips can be predicted from the embedding distance between the original and rephrased question. Figure 4 plots the BioClinicalBERT cosine similarity and Euclidean distance distributions, separately for flipped and non-flipped pairs. Flip and non-flip distributions overlap substantially, so a simple distance threshold cannot separate the two regimes. Surface-level semantic distance is not the load-bearing variable for paraphrase sensitivity in this benchmark.

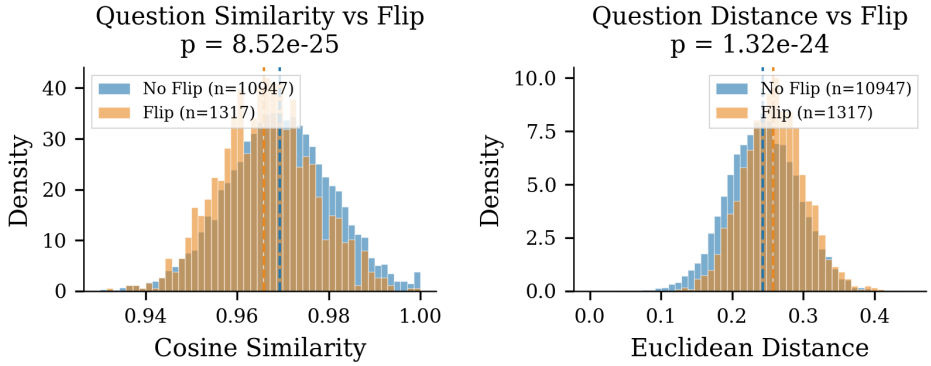
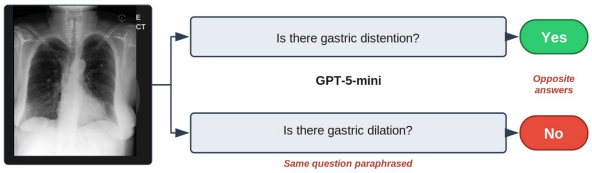


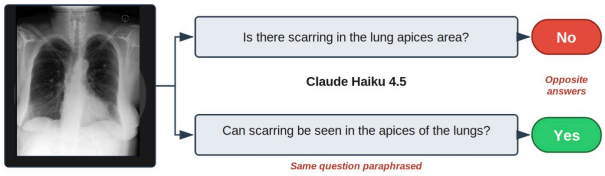
Figure 4. Embedding similarity does not explain paraphrase flips. Flip and non-flip paraphrase pairs show substantial overlap in BioClinicalBERT cosine similarity and Euclidean distance. This suggests that answer flips are not driven only by surface-level semantic distance between the original and rephrased questions.

B. Clinician-confirmed paraphrase flips

Figure 5 shows two clinician-confirmed flips, one on GPT-5-mini (Figure 5a) and one on Claude Haiku 4.5 (Figure 5b). The image and clinical question are held fixed; only the surface phrasing changes. The audit panel labelled both pairings as equivalent under the rubric, so a clinician would treat the original and rephrased question as asking the same fact. Both models nevertheless produce opposite binary answers, which is the failure mode the benchmark is built to surface.



(a) GPT-5-mini



(b) Claude Haiku 4.5

Figure 5. Clinician-confirmed paraphrase flips. The image and clinical question are held fixed; only the surface phrasing changes. The audit panel labelled both pairings as equivalent under the rubric. Both models nevertheless produce opposite binary answers, which is the failure mode the benchmark is built to surface.