

Auditing and Mitigating Paraphrase Fragility in Medical Vision-Language Models

Syed Faizan¹ Anuradha Chandramohan² Parameswari Krishnamurthy^{1*}

¹International Institute of Information Technology (IIIT) Hyderabad, India

²Christian Medical College (CMC), Vellore, India

syed.faizan@ihub-data.iiit.ac.in

anuradha.chandramohan@cmcvellore.ac.in

param.krishna@iiit.ac.in

Abstract

Paraphrases express the same intent in different words, so a robust vision-language model (VLM) should preserve its clinical conclusion when a question is rephrased. Medical VLMs often violate this expectation. We audit this failure on VQA-RAD using nine clinically grounded linguistic operations and 9,723 answer-preserving paraphrases. Across eight checkpoints from five architecture stacks (2.2B–27.4B), Negation Reframing is consistently the most fragile category, with clinical flip rates from 30.7% to 95.3%. A locked, category-conditioned LoRA reduces this failure by 59.0–75.8% across all eight checkpoints; on MedGemma-27B, paraphrase accuracy rises from 68.4% to 75.5%, although base accuracy falls from 79.6% to 73.9%. A pooled adapter performs nearly as well, so we use category-conditioned adapters primarily as diagnostic instruments. Two clinicians agree on 97.8% of a 500-pair calibration sample ($\kappa=0.901$), whereas agreement with an LLM equivalence judge is only $\kappa\approx 0.14$. Sparse-autoencoder clamping yields only a 5.8% reduction, and activation patching provides suggestive rather than conclusive localization evidence. Together, these results provide a reproducible robustness audit for radiology VLMs and make its dataset, generator, evaluation, and deployment limits explicit.

1 Introduction

Paraphrases are alternative ways of expressing the same underlying meaning (Bhagat and Hovy, 2013). In visual question answering, “*Is there pneumothorax?*” and “*Does the image show pneumothorax?*” should lead to the same clinical conclusion. This expectation is especially important for medical VLMs because clinicians can describe the same finding with different vocabulary, syntax, levels of certainty, and use of negation. A model

that responds to these surface choices rather than to the image and clinical meaning is difficult to trust, reproduce, or safely integrate into decision support.

Negation makes the problem particularly clear. Suppose a radiologist asks, “*Is there pneumothorax?*”, and the model answers “*Yes.*” If the question is rephrased as “*Has pneumothorax been ruled out?*”, the clinically corresponding answer is now “*No.*” A model that again answers “*Yes.*” has preserved the literal token while reversing the clinical meaning. This example also shows why paraphrase robustness is not a single, uniform capability. On MEDGEMMA-4B, clinical flip rates range from 8.09% for Hedging to 45.65% for Negation Reframing, a $5.6\times$ spread (Table 11). We therefore measure robustness by linguistic operation rather than only through one aggregate score.

Prior work has established paraphrase sensitivity in language models (Ribeiro et al., 2020; Elazar et al., 2021; Wahle et al., 2023) and begun to quantify it in medical VLMs (Sadanandan and Behzadan, 2026b). Two questions remain central: which linguistic operations fail consistently across model families, and can the dominant failure be reduced without merely collapsing the model to one answer? We address these questions with a clinically grounded, operation-level audit across multiple VLM families and capacities, followed by controlled LoRA and training-free interventions. Our core contribution combines a taxonomy, controlled evaluation, cross-model evidence, and an analysis protocol.

Contributions.

1. A nine-category taxonomy and 9,723 generated paraphrases for 230 VQA-RAD (Lau et al., 2018) clinical-finding yes/no questions, with explicit provenance and answer-preservation criteria (Section 3).
2. A frozen nine-category audit across eight checkpoints from five architecture stacks (2.2B–

*Corresponding author.

27.4B), including six models added after review; Negation Reframing is the highest-flip category for every checkpoint (Section 3).

3. A locked category-conditioned LoRA evaluation that reduces Cat 7 flips by 59.0–75.8% across all eight checkpoints, plus a pooled single-adaptor alternative and a training-free SAE baseline (Section 5).
4. Human validation showing 97.8% clinician agreement ($\kappa=0.901$) but only $\kappa\approx 0.14$ clinician–judge agreement, together with polarity-consistent pre/post accuracy and calibration analysis (Sections 7–8).
5. Exploratory layer ablation, activation patching, and SAE attribution, with causal and transfer limitations stated explicitly (Section 6).

2 Related Work

Paraphrase sensitivity in (medical) VLMs. General paraphrase research decomposes meaning-preserving changes into lexical, syntactic, and polarity operations (Bhagat and Hovy, 2013; Kovatchev et al., 2018; Wahle et al., 2023). We adapt these distinctions to clinical VQA and test whether the resulting operations expose different failure profiles. The closest medical VLM study, PSF-Med (Sadanandan and Behzadan, 2026b), reported a 31% flip-rate reduction through single-layer SAE clamping on MEDGEMMA-4B. A subsequent study (Sadanandan and Behzadan, 2026a) reported a 69.6% relative reduction using LoRA on a different dataset and severity regime. Because those conditions are not directly comparable, we use our within-setup reimplementations (layers 15–19: 55.3% reduction) as the primary comparator.

Calibration and uncertainty in medical AI. Guo et al. (2017) showed that modern networks are often overconfident; Naeini et al. (2015) introduced expected calibration error (ECE), and Ovidia et al. (2019) documented degradation under shift. In medical AI, miscalibration directly threatens deployment (Kompa et al., 2021). We therefore report accuracy, ECE, and polarity-specific error rates alongside flip rate.

Mechanistic interpretability for VLMs. Sparse autoencoders (Bricken et al., 2023; Templeton et al., 2024; McDougall et al., 2025) decompose residual streams into features; activation patching (Meng et al., 2022) tests whether replacing

an activation changes a prediction. Our probes use Gemma Scope 2 SAEs trained on Gemma 3, not MedGemma, so we treat them as exploratory and report the failed clamping intervention prominently.

Parameter-efficient adaptation. LoRA (Hu et al., 2022) injects low-rank updates into attention and MLP projections. We apply the same locked recipe separately by category for diagnosis and jointly across categories for a deployable single-adaptor baseline.

3 Phase 1: Quantifying Paraphrase Sensitivity

3.1 Taxonomy

We refine established operation-level paraphrase typologies (Bhagat and Hovy, 2013; Kovatchev et al., 2018; Wahle et al., 2023) for clinical VQA. The nine categories are: (1) *Clinical Synonyms*, anchored to same-concept terminology in UMLS (Bodenreider, 2004); (2) *Latinate vs. Vernacular*, reflecting the professional–consumer vocabulary gap (Zeng and Tse, 2006); (3) *Syntactic Restructuring*, an established structural paraphrase operation (Bhagat and Hovy, 2013; Kovatchev et al., 2018); (4) *Hedging*, grounded in biomedical uncertainty cues and scope (Vincze et al., 2008); (5) *Register Shift*, restricted to clinical shorthand rather than lay terminology (Sadanandan and Behzadan, 2026b); (6) *Quantifier Insertion*, separating scope/quantity changes from hedging (Kovatchev et al., 2018; Sadanandan and Behzadan, 2026b); (7) *Negation Reframing*, grounded in clinical negation triggers and scope (Chapman et al., 2001); (8) *Abbreviation Expansion*, restricted to standardized, unambiguous forms (Moon et al., 2014); and (9) *Certainty Modulation*, grounded in radiology certainty lexicons and their documented variability (Panicek and Hricak, 2016; Shinagare et al., 2019). Full definitions and exemplars appear in Appendix A.

Three criteria governed inclusion. *Clinical realism* means the expression is attested or accepted by the radiologist reviewer. *Mechanistic distinctness* means one linguistic operation is varied while the clinical concept is held fixed; overlaps were audited and relabelled. *Answer preservation* means the image-level clinical state is unchanged; Cat 7 uniquely requires inversion of the displayed answer token. The first two authors designed and revised the taxonomy, with the radiologist substantively re-

vising six categories and proposing Certainty Modulation.

3.2 Data and protocol

VQA-RAD’s test split contains 251 literal yes/no questions. We deterministically excluded 21 orientation, imaging-plane, modality-identification, modality-property, and image-evaluability questions using the published regular-expression list, leaving 230 clinical-finding questions. GPT-5.5 generated five candidates per admitted question and category using its default sampling configuration. The released file has 9,843 rows: 9,723 non-empty paraphrases and 120 explicit `empty_by_design` records; Abbreviation Expansion contributes 533 paraphrases because fewer questions admit a standardized expansion.

Claude Sonnet 4 labelled each pair using the question: “*would a radiologist give the same yes/no clinical answer to both questions when viewing the same medical image?*” We do not treat this judgement as ground truth. Two clinicians independently relabelled a stratified 500-pair sample: they agreed on 97.8% ($\kappa=0.901$), whereas each agreed with the judge at only $\kappa\approx0.14$ (Section 7). Cat 7 mitigation remains large under clinician-gold and no-filter evaluation.

We define two metrics. The *literal flip rate* is $\Pr[\hat{y}_{\text{para}} \neq \hat{y}_{\text{base}}]$ on the model’s argmax over $\{\text{yes}, \text{no}\}$. The *clinical flip rate* accounts for the *polarity convention* of Cat 7: a base “*Is there X?*” with answer *yes* maps to a paraphrase “*Has X been ruled out?*” whose clinically correct answer is *no*; a literal-flip is the desired behaviour, and a literal-non-flip is the clinical failure. For Cats 1–6, 8–9 the two metrics coincide.

3.3 Results

Table 11 (Appendix D) reports per-category flip rates with bootstrap 95% CIs clustered by base question. Figure 1 visualises the result.

Negation Reframing dominates. On MEDGEMMA-4B, Cat 7 has a clinical flip rate of 45.65% (95% CI [41.0, 50.4]); on QWEN2.5-VL-3B, 44.2% [40.9, 47.5]. The same frozen nine-category queue, prompt, parser, retained-response rule, and question-cluster bootstrap were used for SmolVLM2-2.2B, Granite-Vision-3.3-2B, LLaVA-OneVision-7B, Qwen2.5-VL-7B, the medically adapted Lingshu-7B (Xu et al., 2025), and MedGemma-27B. Cat 7 is

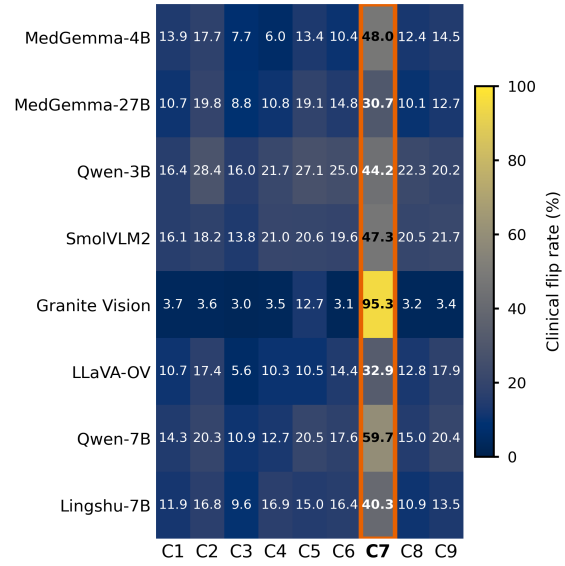


Figure 1: Per-category clinical flip rate across all eight checkpoints. Columns C1–C9 follow the taxonomy in Section 3; annotated cells are percentages. CAT 7 (Negation Reframing) is the maximum for every model.

the highest-flip category for all eight checkpoints, spanning 30.7% to 95.3%; MedGemma-27B is the low endpoint (30.7%, 95% CI [27.0, 34.6]) but remains substantially sensitive (Table 1; complete category grid in Appendix Table 8). Within the MedGemma family, the 27B checkpoint is 14.7 points lower than 4B on the full Cat 7 audit. This replication supports a cross-family, 2.2B–27.4B claim. However, differences between checkpoints cannot be attributed solely to medical pretraining because architecture, scale, and instruction tuning are confounded.

Literal vs. clinical metric and judge stratification. Cat 7 changes which literal token is clinically correct. For example, if “*Is there X?*” is answered *yes*, then the equivalent question “*Has X been ruled out?*” should be answered *no*. The token changes, but the clinical conclusion does not. Consequently, a literal flip can be correct rather than an error: on MEDGEMMA-4B, the literal flip rate is 54.3%, whereas the clinical failure rate is 45.65%, a difference of about nine percentage points. For Cats 1–6 and 8–9, the two metrics coincide (Figure 8, App. C). The Claude Sonnet 4 judge’s verdict also separates flip-prone from stable paraphrases at the population level (Figure 9, App. C), although Section 7 shows that its per-pair reliability is far below clinician reliability.

Table 1: Frozen nine-category baseline and locked Cat 7 LoRA evaluation. Flip rates are clinical-polarity percentages; both pre- and post-LoRA 95% CIs use question-cluster bootstrap. Accuracy is reported on the paired yes/no slice used for the pre/post comparison.

Model	Params	Base acc.	Overall flip	Cat 7 pre (95% CI)	Cat 7 post (95% CI)	Rel. reduction
MedGemma-4B	4.0B	79.6	17.8	48.0 [43.1, 52.9]	12.5 [9.4, 16.1]	73.9
MedGemma-27B	27.4B	79.6	15.7	30.7 [26.9, 34.6]	10.5 [7.7, 13.7]	65.7
Qwen2.5-VL-3B-Instruct	3.0B	71.7	24.7	44.2 [40.9, 47.5]	12.9 [9.4, 16.3]	70.9
SmolVLM2-2.2B-Instruct	2.2B	67.8	22.2	47.3 [43.5, 51.0]	19.4 [15.7, 23.4]	59.0
Granite-Vision-3.3-2B	3.0B	56.5	15.3	95.3 [92.3, 97.8]	23.4 [18.9, 28.1]	75.5
LLaVA-OneVision-7B	7.0B	83.9	14.8	32.9 [28.3, 38.0]	11.8 [8.6, 15.4]	64.0
Qwen2.5-VL-7B-Instruct	7.0B	71.3	21.7	59.7 [54.6, 64.8]	14.4 [10.8, 18.5]	75.8
Lingshu-7B	7.0B	76.5	17.5	40.3 [36.2, 44.2]	10.4 [7.3, 13.7]	74.1

4 Per-Category LoRA Methodology

Phase 1 identifies Negation Reframing as the dominant failure across architectures. We next ask whether low-rank adaptation can reduce this failure while preserving performance on the original questions. To keep recipe development separate from evaluation, we first summarize the exploratory sequence and then specify the locked configuration used for the main results.

4.1 Recipe development (v1–v6)

Here, v1–v6 denote successive LoRA training recipes, each specifying a layer window, target modules, training duration, and consistency-loss weight. v1 uses layers 15–19 with attention-only adaptation for three epochs; v2 adds six-epoch KL consistency; v3 broadens the layer window to 12–21; v4 adds MLP targets to the matched three-epoch setup; v5 tests heterogeneous rank and self-curriculum; and v6 combines the broader window, attention and MLP targets, six epochs, and $\lambda_{\text{KL}}=1.0$ for the main evaluation. These development-set-informed ablations are exploratory; their complete configurations and results appear in Appendix Table 15. We locked v6 after completing this sequence.

4.2 Locked recipe (artifact label v6)

We train one LoRA adapter *per category*, on the VQA-RAD *train*-split paraphrases generated by the same pipeline (Section 3). Each adapter targets:

- **Layers** 12–21 (MEDGEMMA-4B) or 13–22 (QWEN2.5-VL-3B). Broadening the MedGemma window from layers 15–19 to 12–21 produced the largest gain during recipe development; the Qwen window preserves a comparable relative depth in its 36-layer stack.
- **Modules** q , k , v , o (attention) and gate, up, down

(MLP) – 70 adapted modules per adapter.

- **Rank** $r=16$, $\alpha=32$, dropout 0.05.
- **Optimisation** AdamW, $\text{lr}=2\times 10^{-4}$, six epochs, effective batch size 8 via grad accumulation, bf16.
- **Consistency loss** $\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda_{\text{KL}} \cdot D_{\text{KL}}(p_{\text{base}}||p_{\text{para}})$ with $\lambda_{\text{KL}}=1.0$, computed over $\{yes, no\}$ at the first generated token. For Cat 7, p_{para} is on the polarity-flipped target.

A **v6lite** variant (three epochs, $\lambda_{\text{KL}}=0.3$, attention-only) is reported for canonical Cats 3 and 4. This assignment was made after inspecting test results and is therefore a post-hoc diagnostic, not a locked confirmatory result.

4.3 Cat 7 polarity convention

For Cat 7, the paraphrase-side target token is *flipped* relative to the base-side: if the base answer is *yes*, the paraphrase target is *no*, because the negation-reframing operation should preserve the *clinical* answer, not the literal token. All other categories keep the same target on both sides. This is the single most consequential design choice of the recipe.

5 Phase 2 Results

5.1 MedGemma per-category results

Figure 2 and Table 12 show the per-category reductions on MEDGEMMA-4B. On the strictly paired yes/no slice, locked v6 reduces Cat 7 from 47.99% to 12.55%, or **73.85%** relative. The 74.99% layers-10–14 result uses the ablation-slice denominator and is explicitly test-selected (Section 5.4). Canonical Cats 6 and 9 also improve significantly after Benjamini–Hochberg false discovery rate (BH-FDR) correction; Cat 4 regresses descriptively under post-hoc v6lite.

At $\alpha=0.05$, this correction retains canonical Cats 6, 7, and 9; Cat 7 Cohen’s $d = -0.619$, and a 5,000-shuffle paired permutation test gives $p < 0.001$.

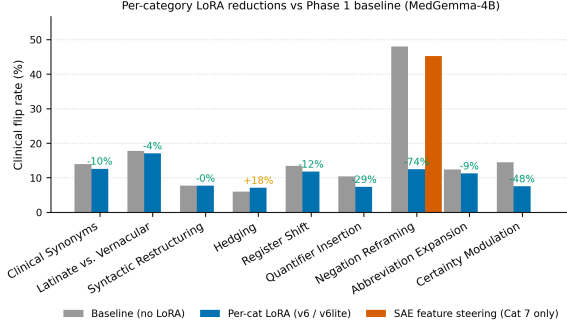


Figure 2: Canonical per-category clinical flip rates before/after LoRA (MEDGEMMA-4B). On the strictly paired analysis Cat 7 falls 73.85% relative. Cats 3–4 use post-hoc v6lite; Cats 6, 7, and 9 survive BH-FDR.

5.2 Cross-family Cat 7 transfer and accuracy

We transferred the locked six-epoch recipe, scaling the layer window proportionally to depth, to all eight checkpoints in Table 1. Cat 7 reductions range from 59.0% (SmolVLM2) to 75.8% (Qwen-7B), so the effect is not isolated to MedGemma-4B/Qwen-3B. On the common paired slice, base/paraphrase accuracy changes are: MEDGEMMA-4B 79.6/55.1 $\rightarrow$ 77.8/75.9, MedGemma-27B 79.6/68.4 $\rightarrow$ 73.9/75.5, QWEN2.5-VL-3B 71.7/57.7 $\rightarrow$ 75.2/71.7, SmolVLM2 67.8/53.4 $\rightarrow$ 73.9/68.8, Granite 56.5/48.2 $\rightarrow$ 70.9/68.2, LLaVA-OneVision 83.9/63.2 $\rightarrow$ 84.3/76.0, Qwen-7B 71.3/53.0 $\rightarrow$ 70.9/71.2, and Lingshu 76.5/65.0 $\rightarrow$ 76.1/75.7. Thus every checkpoint gains paraphrase accuracy, although MedGemma-27B also loses 5.7 points of base accuracy and its ECE changes from 0.221 to 0.240; flip reduction is not a sufficient safety certificate. Appendix Table 10 reports ECE, Brier, FPR, and FNR for every checkpoint. For Lingshu, three independent seeds give post-flip $11.06 \pm 1.81\%$, relative reduction $72.55 \pm 4.45\%$, and post-paraphrase accuracy $74.90 \pm 1.85\%$ (Appendix Table 9). Appendix Table 14 gives category-wise base and paraphrase accuracy for MedGemma-4B and Qwen2.5-VL-3B.

The meaningful prior-work comparison is within setup: our layers-15–19 reimplementing reduces Cat 7 flips by 55.3%, versus 73.9% for locked v6 on the strict paired slice. The published 69.6%

figure used a different dataset, baseline severity, and question count, so it is external context rather than a competitive threshold.

5.3 Recipe ablation

Table 15 in App. F traces the recipe’s evolution from a PSF-Med-style baseline (v1: layers 15–19, attn-only, 3 epochs, no KL) to v6 (Figure 10 in App. F visualises the progression). The largest single-knob gain is broadening the layer window from 15–19 to 12–21 (+10 pp); the second largest is adding MLP to attention-only (+14 pp); heterogeneous rank and self-curriculum (v5) hurt by -4.8 pp; raising λ_{KL} from 0.5 to 1.0 helps Cat 7 specifically because the polarity convention forces the paraphrase-side argmax to mirror a polarity-flipped base.

5.4 Layer ablation

Table 2 and Figure 3 present a five-window, single-run Cat 7 layer ablation. The test-selected window 10–14 reaches a 74.99% reduction, 17 points above 15–19 in our implementation; layers 25–29 yield 21.9%, while 0–9 already yields 65.21%. Prior work also found that early layers could outperform 15–19 on its margin metric, so 15–19 should not be read as a published optimum. Because windows were selected on this evaluation set and the layer-window ablation was not repeated across seeds, we present the geometry as exploratory rather than a general layer-optimality claim.

Table 2: Single-run layer-window ablation on MedGemma-4B Cat 7 ($n=1,150$). Each window was trained from scratch. The 10–14 window was selected on this evaluation set, while locked v6 uses layers 12–21. The table reports the ablation-slice denominator; on the strict paired evaluation, locked v6 achieves a 73.85% relative reduction.

Layer range	Post-LoRA flip %	Rel. reduction %	Notes
0–9	16.70	65.21	
10–14	12.00	74.99	best
15–19	20.00	58.32	Mech-Guided LoRA recipe
20–24	32.26	32.77	
25–29	37.48	21.90	
v6 default (12–21)	12.43	74.09	reference

5.5 Transfer matrix and monolithic baseline

A 9×9 transfer matrix (each adapter evaluated on all categories at $n=50/\text{cat}$; Figure 11) supports a narrow specialization result for Cat 7: its own-category change is -43.8 points, whereas off-diagonal changes range from -6.2 to $+8.0$ points. It does not show that every adapter is

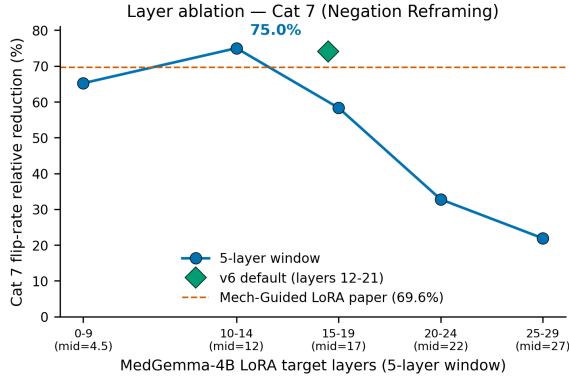


Figure 3: Single-run Cat 7 layer-window ablation (MEDGEMMA-4B). The test-selected 10–14 window reaches 74.99% reduction; the locked result uses layers 12–21.

best on its own category. A *pooled* LoRA trained once over all nine categories (9,540 pairs) achieves 71.08% relative Cat 7 reduction, only 3.0 points below locked category-conditioned v6. Without a validated router, the pooled adapter is the practical deployment baseline and the nine category-conditioned adapters are diagnostic instruments.

6 Mechanistic Analysis

The behavioral experiments establish that adaptation changes the model’s answers, but not where the paraphrase-induced difference is represented or how the adapter acts on it. We therefore use activation patching and sparse-autoencoder attribution as complementary probes. Because neither probe directly validates the adapted model’s internal mechanism, we treat both analyses as exploratory.

6.1 Activation patching

For each *divergent* (base, paraphrase) pair on MEDGEMMA-4B we cache the base-side residual at layer L , replace the paraphrase-side residual at the same layer and token, and decode. The per-layer *reversion rate* (fraction of patches that restore the base-side answer; $N=300$ stratified, 216 pairs admitted) measures how causally responsible each layer is for the paraphrase-induced disagreement. Per-layer numbers are tabulated in Appendix L (Table 16).

The informative part of the reversion curve is its mid-stack transition: 0% for layers 0–14, 12.5% at layer 15, 25% at layer 17, 62.5% at layers 18–19, and approximately 87% by layers 24–30. Near-total reversion at layers 31–33 is expected when patching close to the answer readout and is not it-

self a localization result. Because the probe uses the base model rather than the adapted model, the transition is only consistent with upstream adaptation whose effects are read out later; it does not establish that LoRA creates invariant representations.

6.2 Sparse-autoencoder attribution

We use a Gemma Scope 2 SAE (McDougall et al., 2025) at layer 17 (residual post; 16,384 features, JumpReLU). These SAEs were trained on Gemma 3, not on MedGemma’s medically adapted multimodal distribution; we performed no reconstruction-loss or distribution-transfer validation. For each category, we score features by mean activation difference between flip and non-flip pairs ($N=100/\text{category}$), so the evidence is attributional.

Feature **14026** is active in all nine categories. Its sign *bisects the taxonomy*: positive Δ on lexical-substitution categories (Cats 1, 2, 6, 8) and negative Δ on structural/pragmatic categories (Cats 3, 4, 5, 7, 9). Features 12744 (active in 7/9 categories) and 14479 (active in Cat 7 plus several others) complete the multi-feature steering set. Figure 4 shows the per-category attribution heatmap; the directional bisection of feature 14026 and a partial replication of PSF-Med’s feature 3818 appear in App. H (Figures 12, 13).

6.3 SAE causal steering: an informative negative

Clamping the multi-feature set $\{14026, 12744, 14479\} \rightarrow 0$ at layer 17 over Cat 7 ($n=1150$) reduces the flip rate from 47.99% to 45.21%, only 5.8% relative versus 73.9% for locked LoRA on the strict paired slice. This is our training-free baseline and a negative result: the failed intervention does not causally validate the feature interpretation, although it shows that this simple single-layer steering method is insufficient in our setup.

7 Inter-Annotator Agreement and Filter Robustness

The audit depends on paraphrases preserving the intended clinical meaning. An automatic equivalence judge is convenient at scale, but an unreliable judge can bias both evaluation and filtering. We therefore compare the LLM judge with a dual-clinician calibration sample and then test whether the headline LoRA result survives different filtering regimes.

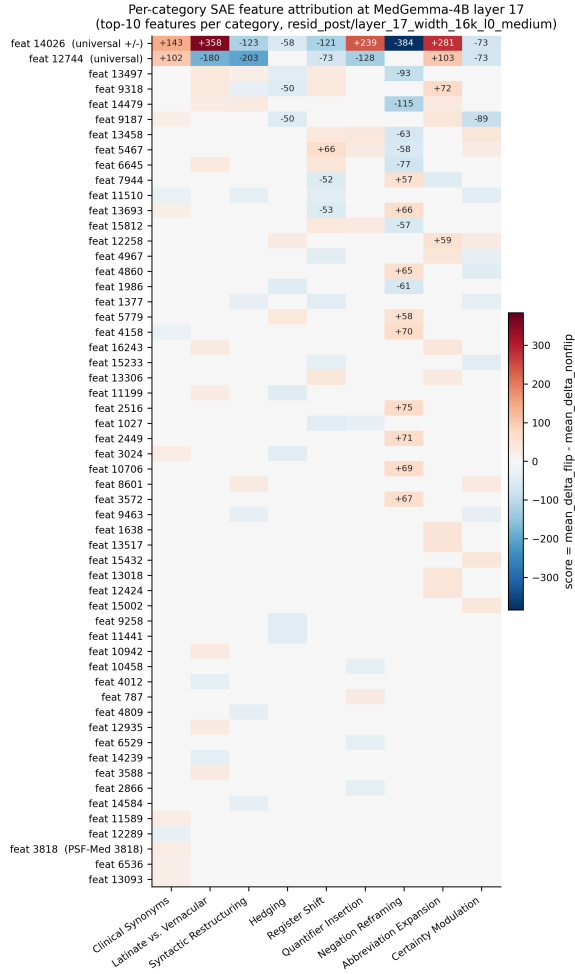


Figure 4: Top-K SAE features (Gemma Scope 2, layer 17) by per-category attribution score. Feature 14026 is active in all 9 categories; the sign of Δ activation bisects the taxonomy into lexical (+) and structural (−) families.

Table 3: Inter-annotator agreement on the 500-pair calibration set. A1 = Radiologist, A2 = M.B.B.S.-qualified physician, Judge = Claude Sonnet 4 LLM judge. 95% CIs from 2,000 bootstrap replicates over pair indices. κ interpretation per Landis and Koch (1977).

Pair	Agreement %	Cohen’s κ	95% CI	Strength
A1 vs A2	97.8%	0.901	[0.841, 0.955]	almost perfect
A1 vs Judge	63.0%	0.135	[0.063, 0.209]	slight
A2 vs Judge	64.0%	0.149	[0.082, 0.223]	slight
3-way Fleiss’ κ	—	0.265	—	fair
3-way Krippendorff’s α	—	0.510	—	—

7.1 Two-clinician gold standard

Two clinician co-authors—a board-certified radiologist and an M.B.B.S. physician—independently labelled a 500-pair sample stratified across the nine categories for clinical equivalence. The same pairs were labelled by the Claude Sonnet 4 judge during quality control. Because both clinicians are co-authors and one helped design the taxonomy, this is expert adjudication rather than independent external validation.

Headline finding. A1 and A2 agree on **97.8%** of pairs with Cohen’s $\kappa = 0.901$ (95% CI [0.841, 0.955]). The prevalence-adjusted bias-adjusted kappa (PABAK) is 0.956, showing that the high equivalence prevalence does not conceal low raw agreement. Each clinician’s agreement with the LLM judge is $\kappa \approx 0.14$, with 95% CIs [0.063, 0.209] and [0.082, 0.223]. Figure 5 visualises the pairwise matrix; per-category results appear in Appendix I.

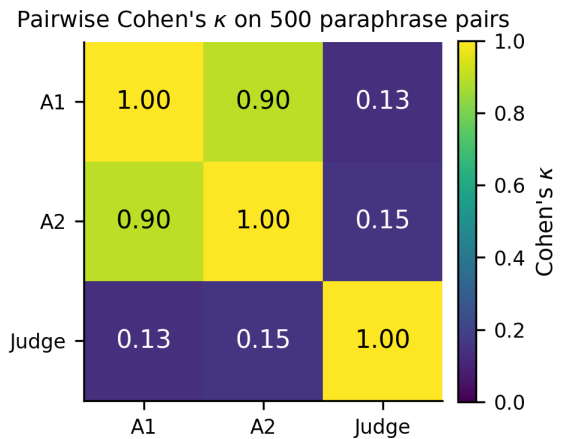


Figure 5: Pairwise Cohen’s κ matrix among A1 (radiologist), A2 (M.B.B.S.-qualified physician), and the Claude Sonnet 4 LLM judge. Clinician–clinician $\kappa = 0.901$; clinician–judge $\kappa \approx 0.14$.

Judge-clinician disagreement is asymmetric.

Of the 200 pairs the judge labelled NOT-EQUIVALENT, 158/200 (A1) and 161/200 (A2) were in fact clinically equivalent. The judge systematically *over-flags* clinically equivalent paraphrases as non-equivalent. This has a direct downstream consequence for any pipeline (including [Sadanandan and Behzadan, 2026b](#)) that filters paraphrases on judge verdicts before computing flip rates: the kept set is biased toward judge-easy pairs, inflating measured robustness.

7.2 Filter robustness

Is our -74% headline an artefact of judge filtering? Table 4 and Figure 6 apply three equivalence-filter regimes to the Cat 7 portion of the calibration set ($n = 50$): the LLM judge, the two-clinician gold (intersection of A1=Y and A2=Y), and no filter.

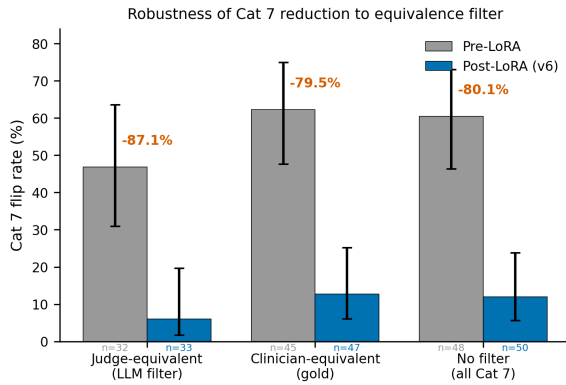


Figure 6: Cat 7 v6 LoRA reduction under three equivalence-filter regimes. The judge inflates to -87% by removing harder pairs; the honest clinician-gold subset gives -79.48% ; no filter gives -80.14% .

The LoRA’s effect is *robust to filter choice*. The judge inflates to -87.07% by removing pairs it disagrees on (systematically the harder ones); the clinician-gold filter gives -79.48% ; no filter gives -80.14% . None of the 95% CIs overlap with zero; the LoRA effect is not a judge artefact.

8 Accuracy and Calibration

Flip rate alone cannot distinguish robustness from answer collapse. We therefore compute literal-token base and paraphrase accuracy, and map Cat 7 predictions, probabilities, and labels back to the same underlying clinical polarity before computing ECE (10 equal-width bins), MCE, Brier score, FPR, and FNR. This polarity mapping is essential: the displayed Cat 7 token is the inverse of the clinical state.

On MEDGEMMA-4B Cat 7, paraphrase accuracy rises from 55.1% to 75.9% on the paired yes/no slice while base accuracy changes from 79.6% to 77.8%. In clinical polarity, ECE improves from 0.403 to 0.208, Brier from 0.411 to 0.221, and FNR from 0.73 to 0.21; FPR increases from 0.19 to 0.27. Thus the headline flip reduction is not answer collapse, although the FPR trade-off remains clinically relevant. The earlier draft’s opposite conclusion resulted from comparing raw $P(\text{displayed yes})$ with the uninverted clinical label; the released analysis script now enforces one polarity convention.

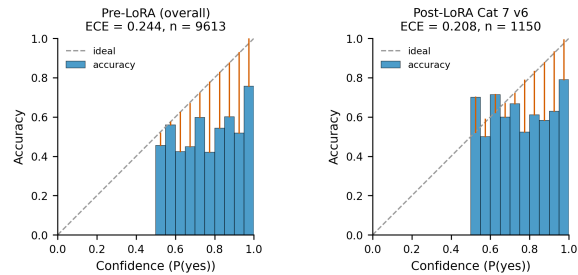


Figure 7: CAT 7 reliability in underlying clinical polarity. Post-LoRA calibration improves; per-category ECE appears in Figure 16.

Across the other categories, ΔECE ranges from -0.049 to $+0.014$ (Table 5). These are evaluation results, not evidence of deployment safety: VQA-RAD is small and not a prospective clinical cohort, and no adapter should be used without site-specific validation and clinician oversight.

9 Discussion

Three findings organize the evidence. First, paraphrase fragility is structured by linguistic operation and model family rather than captured by one aggregate rate. Second, LoRA substantially reduces the dominant Negation Reframing failure, but accuracy and calibration trade-offs prevent us from equating fewer flips with safety. Third, the evaluation protocol matters: clinician adjudication exposes a systematic bias that an LLM-only equivalence filter would conceal.

Category-level effects are model-dependent.

On MEDGEMMA-4B, only canonical Categories 6, 7, and 9 survive BH-FDR and Category 4 regresses descriptively. On QWEN2.5-VL-3B, all nine categories improve descriptively after the corrected name-based mapping, but we did not repeat the confirmatory significance analysis. We therefore

Table 4: Robustness of the Cat 7 v6 LoRA reduction to the choice of paraphrase-equivalence filter. Computed on the 500-pair calibration subset (Cat 7 portion). Wilson 95% CIs in brackets.

Equivalence filter	N	Pre-LoRA flip % [95% CI]	Post-LoRA flip % [95% CI]	Δ rel %	95% CI overlap
Judge (Claude Sonnet 4)	32	46.88 [30.9, 63.6]	6.06 [1.7, 19.6]	+87.07	no
Both clinicians (gold)	45	62.22 [47.6, 74.9]	12.77 [6.0, 25.2]	+79.48	no
No filter	48	60.42 [46.3, 73.0]	12.00 [5.6, 23.8]	+80.14	no

Table 5: Calibration before and after per-category LoRA. ECE = Expected Calibration Error (10 equal-width bins on max-class confidence, Guo et al. 2017). MCE = Maximum Calibration Error. Brier = mean squared error on $P(\text{yes})$. For Cat 7, probabilities, labels, and FPR/FNR are mapped to the underlying clinical state before comparison. FPR / FNR use a 0.5 threshold on paraphrase questions. 95% CIs from 500 bootstrap replicates. Statistically significant ECE increases (post-LoRA worse than pre-LoRA) are highlighted with $\uparrow$.

Cat	Name	ECE pre	ECE post (95% CI)	MCE pre $\rightarrow$ post	Brier pre $\rightarrow$ post	FPR $\rightarrow$	FNR $\rightarrow$	Δ ECE
1	Clinical Synonyms	0.219	0.202 [0.182, 0.227]	0.47 $\rightarrow$ 0.43	0.222 $\rightarrow$ 0.212	0.25 $\rightarrow$ 0.22	0.20 $\rightarrow$ 0.26	-0.016
2	Latinate vs. Vernacular	0.238	0.189 [0.170, 0.215]	0.50 $\rightarrow$ 0.30	0.245 $\rightarrow$ 0.206	0.31 $\rightarrow$ 0.25	0.21 $\rightarrow$ 0.20	-0.049
3	Syntactic Restructuring	0.219	0.195 [0.176, 0.221]	0.47 $\rightarrow$ 0.62	0.223 $\rightarrow$ 0.208	0.29 $\rightarrow$ 0.24	0.19 $\rightarrow$ 0.21	-0.024
4	Hedging	0.207	0.207 [0.184, 0.235]	0.42 $\rightarrow$ 0.61	0.210 $\rightarrow$ 0.222	0.22 $\rightarrow$ 0.30	0.22 $\rightarrow$ 0.19	-0.000
5	Register Shift	0.244	0.220 [0.195, 0.246]	0.49 $\rightarrow$ 0.43	0.249 $\rightarrow$ 0.232	0.38 $\rightarrow$ 0.35	0.14 $\rightarrow$ 0.15	-0.024
6	Quantifier Insertion	0.228	0.242 [0.218, 0.269]	0.37 $\rightarrow$ 0.38	0.233 $\rightarrow$ 0.253	0.31 $\rightarrow$ 0.29	0.19 $\rightarrow$ 0.28	+0.014
7	Negation Reframing	0.403	0.208 [0.187, 0.236]	0.52 $\rightarrow$ 0.30	0.411 $\rightarrow$ 0.221	0.19 $\rightarrow$ 0.27	0.73 $\rightarrow$ 0.21	-0.195
8	Abbreviation Expansion	0.180	0.137 [0.111, 0.173]	0.56 $\rightarrow$ 0.38	0.183 $\rightarrow$ 0.151	0.25 $\rightarrow$ 0.18	0.15 $\rightarrow$ 0.16	-0.043
9	Certainty Modulation	0.245	0.214 [0.194, 0.240]	0.62 $\rightarrow$ 0.43	0.248 $\rightarrow$ 0.220	0.35 $\rightarrow$ 0.25	0.17 $\rightarrow$ 0.21	-0.030

do not treat weak MedGemma effects in six categories as universal or claim a confirmed category-by-model interaction.

What the layer geometry tells us. The single-run ablation favors an earlier intervention than our within-setup layers-15–19 comparator, while patching changes most sharply through the middle-to-late stack. This is consistent with, but does not prove, an upstream-adaptation/late-readout account. Testing the adapted model directly and repeating the layer-window experiment across seeds are required before making a stronger causal claim.

When per-category beats monolithic. The pooled adapter reaches 71.08% Cat 7 reduction, close to the locked category-conditioned adapter’s 73.85% on the strict paired slice. Since the category of an incoming question is unknown and we do not validate a router, pooled LoRA is the practical single-adapter baseline. Category-conditioned adapters remain useful for controlled diagnosis and transfer analysis, not as a complete deployment proposal.

The clinician-gold filter as the right evaluation. Our IAA result implies that any evaluation that uses an LLM judge as the equivalence ground truth is systematically biased. Numbers inflate when the judge removes harder pairs and deflate when the judge keeps easier ones. The fix is to build a clinician-labelled gold for at least a calibration sam-

ple, report *both* judge-filtered and clinician-filtered numbers, and use the gap as a sanity check. Our Cat 7 LoRA’s -87% (judge) vs. -79% (clinician) gap is small enough to retain the headline finding, but other interventions may show much larger gaps.

10 Conclusion

Paraphrase robustness asks a basic but consequential question: does a model preserve its clinical conclusion when the wording changes but the meaning does not? Across eight checkpoints from five architecture stacks, our operation-level audit shows that the answer depends strongly on how the question is rephrased, with Negation Reframing consistently the most fragile operation. A locked LoRA reduces this failure by 59.0–75.8% and improves paraphrase accuracy on every checkpoint, but the MedGemma-27B base-accuracy loss confirms that flip reduction alone is not a safety certificate. A pooled adapter provides a practical alternative to category-conditioned adapters, which are most useful as diagnostic tools. Clinician adjudication further reveals severe disagreement with an LLM equivalence judge, while polarity-consistent evaluation rules out answer collapse on the flagship MedGemma result. The mechanistic evidence remains exploratory. We release the corpus, labels, analysis code, raw outputs, and adapter artefacts to support reproducible robustness evaluation.

Limitations

Single clinical sub-domain. All experiments use VQA-RAD (Lau et al., 2018), which is radiology *yes/no* VQA only. We do not test free-text generation, non-radiology specialties, human-written rephrasings, or a second dataset. Train and test questions are split-disjoint, but their paraphrases come from the same GPT-5.5 pipeline, so generator-style transfer remains untested.

Architecture scope. We evaluate eight checkpoints from five architecture stacks at 2.2B–27.4B scale: MedGemma (4B/27B), Lingshu, Qwen2.5-VL (3B/7B), SmolVLM2, Granite Vision, and LLaVA-OneVision. The evaluation provides cross-family coverage and within-family capacity comparisons for MedGemma and Qwen, but remains limited to English VLMs. Architecture, scale, instruction tuning, and medical adaptation are confounded, so we do not attribute differences to medical pretraining alone. Moreover, Lingshu’s published training mixture includes VQA-RAD (Xu et al., 2025); its result is a robustness replication, not evidence of unseen-dataset generalization.

Phase 1 / Phase 2 Qwen baseline confound. The paraphrase corpus was generated and judged once on Qwen-7B’s answer distribution and then re-used to evaluate Qwen-3B. The Phase 1 Qwen-3B baseline was computed only *after* the Phase 2 LoRA evaluation, so Phase 2 Qwen-3B was trained against paraphrases that targeted Qwen-7B behaviour. We re-verified that the headline Cat 7 number remains within bootstrap noise of the re-baseline, but smaller category-level deltas may carry residual size confounds.

Calibration sample is small for per-category numbers. The 500-pair calibration set contains only 50 Cat 7 pairs, dual-labelled. The filter-robustness numbers in Section 7 therefore have Wilson intervals on the order of ± 15 pp. The headline κ result is statistically sharp (CI [0.841, 0.955]), but per-category κ stratification (Figure 14) leans on ~ 55 pairs per category and should be read as descriptive.

LLM-judge biases are characterised, not eliminated. We quantify the Claude Sonnet 4 judge’s $\kappa \approx 0.14$ disagreement with clinicians on 500 pairs and demonstrate filter-robustness on 50 Cat 7 pairs. We did not re-train the LoRA on a clinician-validated equivalence subset (which would have

required clinician labelling of the full $\sim 1,150$ -pair Cat 7 *training* corpus, well beyond the resources of this submission). It remains an open question whether clinician-validated training would push the headline reduction past the -79.48% we report under clinician-gold evaluation.

Accuracy and calibration are retrospective. The corrected Cat 7 analysis improves ECE and FNR while modestly worsening FPR, and paraphrase accuracy improves substantially. These estimates are retrospective on a small benchmark, not clinical-safety evidence; confidence intervals and prospective, site-specific evaluation would be required for deployment.

Activation patching is base→paraphrase only. We patch the base-side residual into the paraphrase-side forward pass at the last token only, single-layer at a time. We did not probe paraphrase→base, multi-layer paths, or attention-head-level resolution. The layer-31–33 crystallisation result holds in our protocol; finer-grained protocols may localise sub-components within those layers.

Single generator, judge, and limited training repeats. All paraphrases were generated once with GPT-5.5 using its default sampling configuration and judged with Claude Sonnet 4. The 500-pair sample has dual-clinician labels, but both clinicians are co-authors and no external annotator was available. We repeated Lingshu Cat 7 LoRA across three seeds, but the remaining adapter results and every layer-window condition are single-run.

Ethics Statement

This work uses VQA-RAD (Lau et al., 2018), a public, IRB-cleared, de-identified dataset of radiology images and human-curated *yes/no* questions. No private patient data, PHI, or clinical reports were used. The 500-pair calibration set was labelled by two clinician co-authors (a board-certified radiologist and an M.B.B.S. physician) on public VQA-RAD material, with no patient interaction. We release the per-labeller Y/N labels under the dataset’s applicable terms.

Intended use and dual-use considerations. The adapters are research artefacts, not clinical devices. Although Cat 7 paraphrase accuracy and calibration improve in our corrected analysis, FPR rises from 0.19 to 0.27 on MEDGEMMA-4B, and none of these retrospective benchmark measurements

establishes patient safety. Any downstream study must use clinician-labelled, site-specific validation, explicit abstention and monitoring, and a clinician in the loop; direct clinical use is out of scope.

Compute and environmental cost. The original experiment estimate was ~ 120 A100 80GB GPU-hours; the three added Lingshu LoRA seeds took 5.28 measured A100 80GB hours, and MedGemma-27B LoRA took 1.98 measured H100 hours. Added inference and the other extension runs used the GPUs listed in Appendix Table 7. Total paraphrase-generation cost was $\sim \$58$ OpenAI plus approximately INR 400 Anthropic. We did not train any model from scratch; all adaptation is LoRA on publicly released checkpoints.

References

- Rahul Bhagat and Eduard Hovy. 2013. [What is a paraphrase?](#) *Computational Linguistics*, 39(3):463–472.
- Olivier Bodenreider. 2004. [The unified medical language system \(UMLS\): Integrating biomedical terminology.](#) *Nucleic Acids Research*, 32(Database issue):D267–D270.
- Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermy, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, et al. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits Thread.
- Wendy W. Chapman, Will Bridewell, Paul Hanbury, Gregory F. Cooper, and Bruce G. Buchanan. 2001. [A simple algorithm for identifying negated findings and diseases in discharge summaries.](#) *Journal of Biomedical Informatics*, 34(5):301–310.
- Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics (TACL)*, 9:1012–1031.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, pages 1321–1330.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.
- Benjamin Kompa, Jasper Snoek, and Andrew L. Beam. 2021. Second opinion needed: Communicating uncertainty in medical machine learning. *npj Digital Medicine*, 4(1):4.
- Venelin Kovatchev, M. Antònia Martí, and Maria Salamó. 2018. ETPC: A paraphrase identification corpus annotated with extended paraphrase typology and negation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association.
- J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.
- Jason J. Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data*, 5:180251.
- Callum McDougall, Arthur Conmy, János Kramár, Tom Lieberum, Senthoooran Rajamanoharan, and Neel Nanda. 2025. Gemma Scope 2: Technical paper. Technical report, Google DeepMind.
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Sungrim Moon, Serguei Pakhomov, Nathan Liu, James O. Ryan, and Genevieve B. Melton. 2014. [A sense inventory for clinical abbreviations and acronyms created using clinical notes and medical dictionary resources.](#) *Journal of the American Medical Informatics Association*, 21(2):299–307.
- Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using Bayesian binning. In *AAAI Conference on Artificial Intelligence*, pages 2901–2907.
- Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. 2019. Can you trust your model’s uncertainty? Evaluating predictive uncertainty under dataset shift. *Advances in Neural Information Processing Systems (NeurIPS)*.
- David M. Panicek and Hedvig Hricak. 2016. [How sure are you, doctor? a standardized lexicon to describe the radiologist’s level of certainty.](#) *AJR American Journal of Roentgenology*, 207(1):2–3.
- Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4902–4912.
- Binesh Sadanandan and Vahid Behzadan. 2026a. [Mechanistically guided LoRA improves paraphrase consistency in medical vision-language models.](#) *Preprint*, arXiv:2603.00148.
- Binesh Sadanandan and Vahid Behzadan. 2026b. [PSF-Med: Measuring and explaining paraphrase sensitivity in medical vision language models.](#) *Preprint*, arXiv:2602.21428.

- Atul B. Shinagare, Ronilda Lacson, Giles W. Boland, Aijia Wang, Stuart G. Silverman, William W. Mayo-Smith, and Ramin Khorasani. 2019. [Radiologist preferences, agreement, and variability in phrases used to convey diagnostic certainty in radiology reports](#). *Journal of the American College of Radiology*, 16(4):458–464.
- Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Adam Jermy, et al. 2024. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. Transformer Circuits Thread.
- Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. 2008. [The BioScope corpus: Biomedical texts annotated for uncertainty, negation and their scopes](#). *BMC Bioinformatics*, 9(Suppl 11):S9.
- Jan Philip Wahle, Bela Gipp, and Terry Ruas. 2023. [Paraphrase types for generation and detection](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12148–12164.
- Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, et al. 2025. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *arXiv preprint arXiv:2506.07044*.
- Qing T. Zeng and Tony Tse. 2006. [Exploring and developing consumer health vocabularies](#). *Journal of the American Medical Informatics Association*, 13(1):24–29.

A Taxonomy Specification with Exemplars

Table 6 reproduces the operationalisation used by Script 05’s GPT-5.5 prompt.

B Checkpoint and Compute Details

Table 7 gives the exact checkpoint identifiers, loader classes, and accelerators used by the frozen audit. The same deterministic prompt and parser were used throughout; the full added-model run-time is pinned in Script 32.

Table 8 reports the complete eight-checkpoint baseline grid rather than only the headline Cat 7 column.

Lingshu training-seed variance. Table 9 reports the strict paired result for each independent seed. The adapter weight hashes are distinct and the training metadata records each seed and run duration.

All-model Cat 7 safety audit. Table 10 gives the complete paired accuracy, calibration, and polarity-specific error analysis promised for the added checkpoints. It exposes heterogeneous trade-offs that flip rate alone conceals; for example, several adapters reduce one polarity-specific error while increasing the other or worsening ECE.

C Supplementary Phase 1 Figures

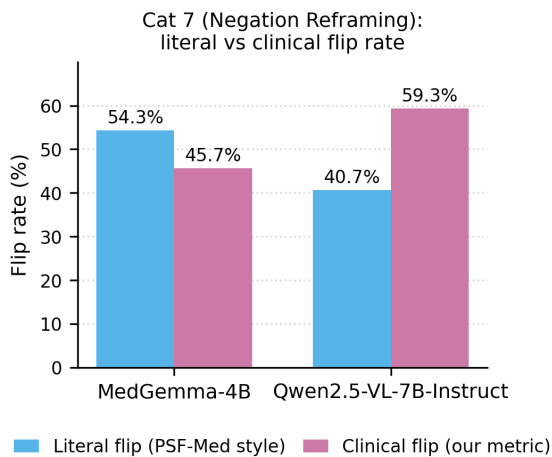


Figure 8: Literal vs. clinical flip rate on Cat 7. On Cats 1–6, 8–9 the paraphrase preserves polarity and the two metrics coincide; on Cat 7 they diverge by ~10 pp because the literal yes/no token swaps but the clinical answer should not.

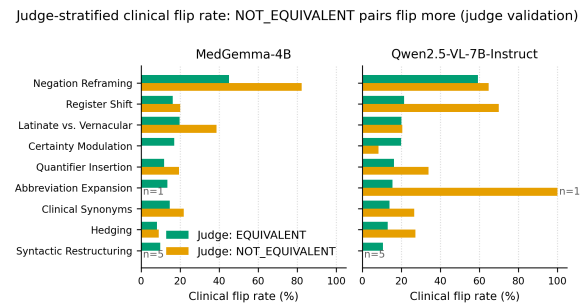


Figure 9: Flip rate stratified by Claude Sonnet 4’s equivalence verdict. Judge-equivalent pairs flip less, confirming partial population-level signal; per-pair reliability is far below clinicians (see Section 7).

D Phase 1: Per-category Flip Rates

E Phase 2: Per-category LoRA results

F Recipe Ablation Hyperparameters

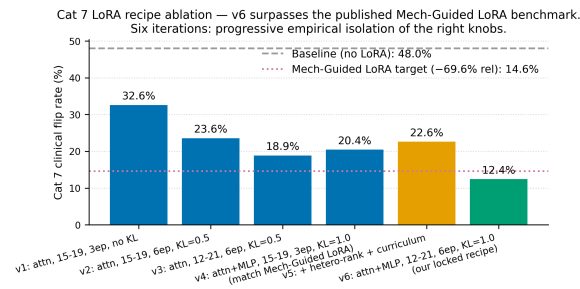


Figure 10: Recipe-ablation progression on MEDGEMMA-4B Cat 7. v6 reduces flips by 74.09% versus 55.3% for the matched layers-15–19 reimplementations; on the strict paired evaluation, the reduction is 73.85%.

G Transfer Matrix

H Supplementary SAE Figures

I Supplementary IAA Figures

J Supplementary Calibration Figures

K Post-hoc v6 vs v6lite Comparison

L Activation-Patching Reversion Rate

M Bug Log / Reproducibility Notes

The following issues were encountered and resolved during the experimental campaign; they are documented here to aid reproduction.

- MOCK_MODE clobbered on git pull.** Script 16’s `MOCK_MODE = True` default was overwritten on the GPU pod by `git pull`, causing

Table 6: Nine-category paraphrase taxonomy. Each category corresponds to a single linguistic operation type; paraphrases combine operations across categories only via the polarity-preserving union (Cat 7 is the sole polarity-flipping category).

Cat	Definition	Example
1	Clinical synonyms (same-register)	<i>cancer</i> → <i>malignancy</i>
2	Latinate ↔ vernacular	<i>cardiomegaly</i> → <i>enlarged heart</i>
3	Syntactic restructuring	<i>Is there fluid?</i> → <i>Fluid – is it present?</i>
4	Hedging	<i>Is there X?</i> → <i>Could there be X?</i>
5	Register shift	<i>Is there X?</i> → <i>Spot any X?</i>
6	Quantifier insertion	<i>Is there a mass?</i> → <i>Is there any mass?</i>
7	Negation reframing (polarity flip)	<i>Is there X?</i> → <i>Has X been ruled out?</i>
8	Abbreviation expansion	<i>Is there CHF?</i> → <i>Is there congestive heart failure?</i>
9	Certainty modulation	<i>Is there X?</i> → <i>Is it definite that X is present?</i>

Table 7: Exact public checkpoints and execution paths. All processors are loaded with `AutoProcessor`; added 7B checkpoints use an A100 80GB, MedGemma-27B uses an H100, and the two added compact checkpoints use an L40S. Original MedGemma/Qwen runs used an A100 80GB.

Display name	Hugging Face checkpoint ID	Transformers model class	GPU
MedGemma-4B	google/medgemma-4b-it	<code>AutoModelForImageTextToText</code>	A100 80GB
MedGemma-27B	google/medgemma-27b-it	<code>AutoModelForImageTextToText</code>	H100
Qwen2.5-VL-3B-Instruct	Qwen/Qwen2.5-VL-3B-Instruct	<code>Qwen2_5_VLForConditionalGeneration</code>	A100 80GB
SmolVLM2-2.2B-Instruct	HuggingFaceTB/SmolVLM2-2.2B-Instruct	<code>AutoModelForImageTextToText</code>	L40S
Granite-Vision-3.3-2B	ibm-granite/granite-vision-3.3-2b	<code>AutoModelForVision2Seq</code>	L40S
LLaVA-OneVision-7B	llava-hf/llava-onevision-qwen2-7b-ov-hf	<code>LlavaOnevisionForConditionalGeneration</code>	A100 80GB
Qwen2.5-VL-7B-Instruct	Qwen/Qwen2.5-VL-7B-Instruct	<code>Qwen2_5_VLForConditionalGeneration</code>	A100 80GB
Lingshu-7B	lingshu-medical-mlm/Lingshu-7B	<code>Qwen2_5_VLForConditionalGeneration</code>	A100 80GB

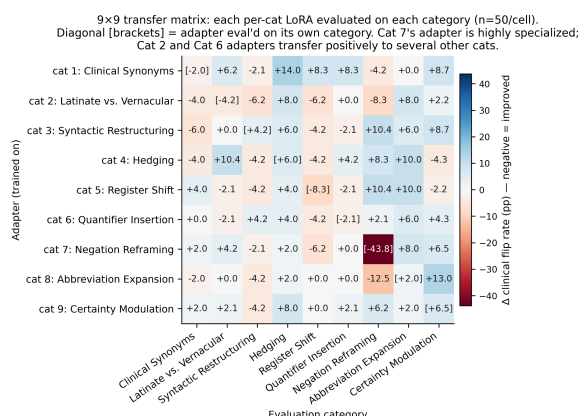


Figure 11: Exploratory transfer matrix: row = adapter, column = evaluation category, cell = post-LoRA flip rate (n=50/cell). Cat 7 shows strong own-category specialization; the matrix does not show universal diagonal dominance.

the first autonomous queue to no-op. Fixed by an explicit env-var override; the queue script now exports `MOCK_MODE=false`.

- Script 21 adapter prefix hardcoded.** The prefix `medgemma_cat{N}_v6` was hardcoded; now `MODEL_FAMILY-aware` (qwen25v13b vs. medgemma).
- Script 17 baseline hardcoded.** The Qwen Cat 7 baseline was originally hardcoded

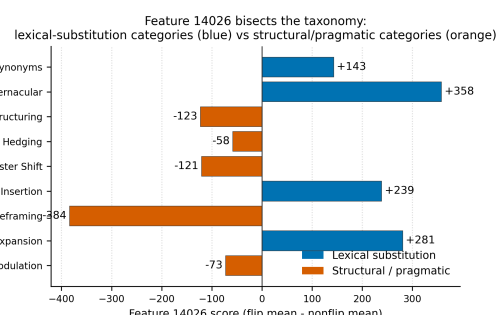


Figure 12: Per-category mean Δ activation of feature 14026. Positive on Cats 1, 2, 6, 8 (lexical); negative on Cats 3, 4, 5, 7, 9 (structural/pragmatic).

to MEDGEMMA-4B's 0.457; the reported reduction was an erroneous -71.84% . Now loads per-VLM baselines from `flip_rates_summary.json`; the correct value is -70.88% .

- Phase 1 Qwen-3B baseline computed post-LoRA.** Phase 1 originally ran on Qwen-7B; Phase 2 LoRA used Qwen-3B due to a 24 GB GPU constraint. The Qwen-3B Phase 1 baseline was computed only after the LoRA evaluation finished. This is the residual confound documented in Section Limitations.
- SAE loader.** We did not install `sae-lens` (the

Table 8: Clinical flip rate (%) for every checkpoint and canonical category under the full frozen audit. Each cell uses the model-retained yes/no pairs for that category; bold marks the row maximum. Cat 7 is highest for every checkpoint.

Model	C1	C2	C3	C4	C5	C6	C7	C8	C9
MedGemma-4B	13.9	17.7	7.7	6.0	13.4	10.4	48.0	12.4	14.5
MedGemma-27B	10.7	19.8	8.8	10.8	19.1	14.8	30.7	10.1	12.7
Qwen2.5-VL-3B-Instruct	16.4	28.4	16.0	21.7	27.1	25.0	44.2	22.3	20.2
SmolVLM2-2.2B-Instruct	16.1	18.2	13.8	21.0	20.6	19.6	47.3	20.5	21.7
Granite-Vision-3.3-2B	3.7	3.6	3.0	3.5	12.7	3.1	95.3	3.2	3.4
LLaVA-OneVision-7B	10.7	17.4	5.6	10.3	10.5	14.4	32.9	12.8	17.9
Qwen2.5-VL-7B-Instruct	14.3	20.3	10.9	12.7	20.5	17.6	59.7	15.0	20.4
Lingshu-7B	11.9	16.8	9.6	16.9	15.0	16.4	40.3	10.9	13.5

Table 9: Independent Lingshu-7B Cat 7 LoRA runs on the strict paired yes/no slice. Values are percentages; SD is the sample standard deviation across three seeds.

Seed	Post flip	Rel. red.	Base acc.	Para acc.
8675309	10.43	74.08	76.09	75.74
314159	13.10	67.53	73.80	76.17
271828	9.65	76.03	73.04	72.78
Mean $\pm$ SD	11.06 $\pm$ 1.81	72.55 $\pm$ 4.45	74.31 $\pm$ 1.58	74.90 $\pm$ 1.85

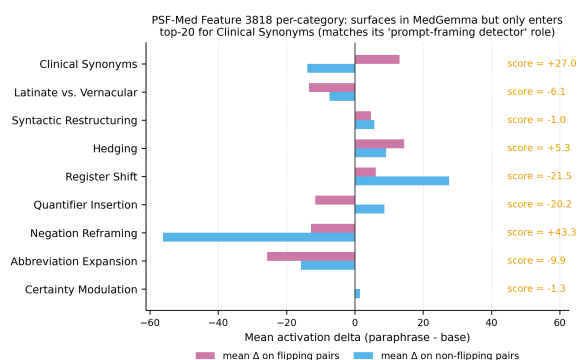


Figure 13: Partial replication of PSF-Med’s feature 3818 on the 9-cat setup. The feature is active on some categories but does not bisect the taxonomy as feature 14026 does.

package forces a torch upgrade that conflicts with our transformers version); a ~ 30 -line hand-rolled JumpReLU loader on the Gemma Scope 2 safetensors suffices.

- Activation patching runtime.** The $N=300$ patching sweep over 34 layers ran in 16 min, not the projected 3 hr; PyTorch’s hook overhead is small.

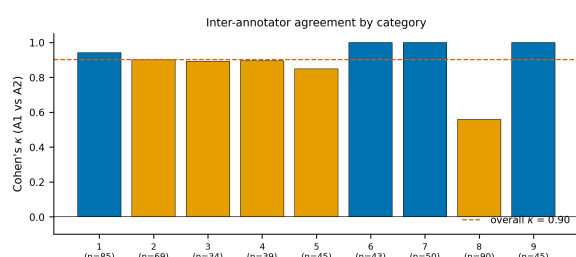


Figure 14: Per-category Cohen’s κ between A1 and A2 on the calibration subset. Agreement is high across all nine categories.

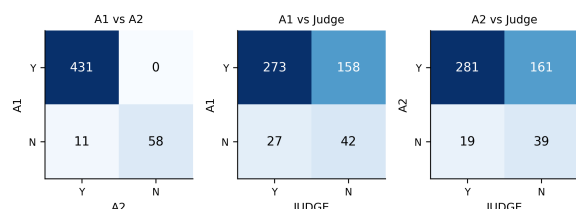


Figure 15: Pairwise confusion matrices. Clinicians disagree on 11/500 = 2.2% of pairs; clinician–judge disagreements concentrate in the Y/N off-diagonal where the judge says *not-equivalent* and the clinician says *equivalent*.

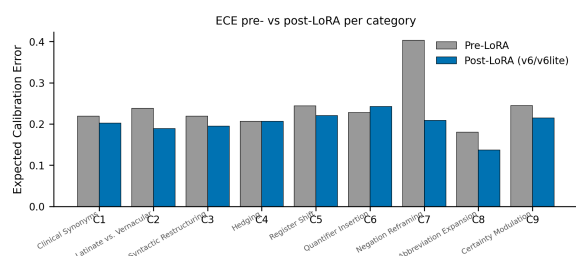


Figure 16: Per-category ECE before and after v6 LoRA on MEDGEMMA-4B under the corrected category mapping and clinical-polarity convention. Cat 7 improves most; Cat 6 rises modestly, while the remaining categories are flat or improve.

Table 10: Strict paired Cat 7 performance and clinical-polarity safety metrics for all eight checkpoints. Arrows show pre- to post-LoRA values. Accuracy is scored against the displayed question’s literal token; ECE, Brier, FPR, and FNR are computed after mapping predictions, probabilities, and labels to the underlying clinical state. Uncertain outputs count as incorrect for accuracy and are excluded from FPR/FNR.

Model	n	Flip (%)	Rel. red. (%)	Base acc. (%)	Para. acc. (%)
MedGemma-4B	1116	48.0→12.5	73.9	79.6→77.8	55.1→75.9
MedGemma-27B	1149	30.7→10.5	65.7	79.6→73.9	68.4→75.5
Qwen2.5-VL-3B-Instruct	1150	44.2→12.9	70.9	71.7→75.2	57.7→71.7
SmolVLM2-2.2B-Instruct	1150	47.3→19.4	59.0	67.8→73.9	53.4→68.8
Granite-Vision-3.3-2B	1150	95.3→23.4	75.5	56.5→70.9	48.2→68.2
LLaVA-OneVision-7B	1150	32.9→11.8	64.0	83.9→84.3	63.2→76.0
Qwen2.5-VL-7B-Instruct	1150	59.7→14.4	75.8	71.3→70.9	53.0→71.2
Lingshu-7B	1150	40.3→10.4	74.1	76.5→76.1	65.0→75.7

Clinical-polarity calibration and error rates				
Model	ECE	Brier	FPR	FNR
MedGemma-4B	0.401→0.210	0.411→0.223	0.190→0.269	0.728→0.210
MedGemma-27B	0.221→0.240	0.253→0.242	0.247→0.259	0.390→0.231
Qwen2.5-VL-3B-Instruct	0.058→0.229	0.244→0.248	0.529→0.212	0.310→0.359
SmolVLM2-2.2B-Instruct	0.109→0.288	0.258→0.293	0.402→0.282	0.535→0.344
Granite-Vision-3.3-2B	0.516→0.298	0.516→0.303	1.000→0.333	0.002→0.303
LLaVA-OneVision-7B	0.044→0.202	0.226→0.210	0.128→0.166	0.625→0.319
Qwen2.5-VL-7B-Instruct	0.256→0.270	0.316→0.275	0.822→0.249	0.092→0.330
Lingshu-7B	0.059→0.230	0.228→0.234	0.266→0.215	0.441→0.272

Table 11: Full Phase 1 clinical flip rates with question-cluster bootstrap 95% CIs. $n_{\text{para}} \approx 1,150$ per category except Cat 8 Abbreviation Expansion (533 admitted paraphrases).

Cat	Name	MEDGEMMA-4B		Qwen2.5-VL-7B		QWEN2.5-VL-3B	
		Clin-flip	95% CI	Clin-flip	95% CI	Clin-flip	95% CI
1	Clinical Synonyms	15.11%	[11.6, 18.8]	14.59%	[11.3, 18.1]	16.4%	[13.0, 20.0]
2	Latinate vs. Vernacular	20.43%	[16.5, 24.4]	20.00%	[16.1, 24.0]	28.4%	[24.2, 32.8]
3	Syntactic Restructuring	9.83%	[7.3, 12.6]	10.61%	[8.0, 13.3]	16.0%	[12.5, 19.7]
4	Hedging	8.09%	[5.8, 10.6]	13.13%	[9.8, 16.7]	21.7%	[17.5, 26.1]
5	Register Shift	16.17%	[12.7, 19.8]	21.83%	[17.8, 26.0]	27.1%	[22.7, 31.7]
6	Quantifier Insertion	12.23%	[9.3, 15.3]	17.21%	[14.2, 20.3]	25.0%	[21.3, 28.7]
7	Negation Reframing	45.65%	[41.0, 50.4]	59.30%	[54.1, 64.4]	44.2%	[40.9, 47.5]
8	Abbreviation Expansion	13.51%	[8.5, 18.9]	15.57%	[9.9, 21.5]	22.3%	[16.5, 28.7]
9	Certainty Modulation	16.78%	[13.0, 20.8]	19.83%	[16.4, 23.4]	20.2%	[17.0, 23.7]
OVERALL		17.79%	[15.5, 20.2]	21.7%	[19.6, 23.9]	24.7%	[22.4, 27.1]

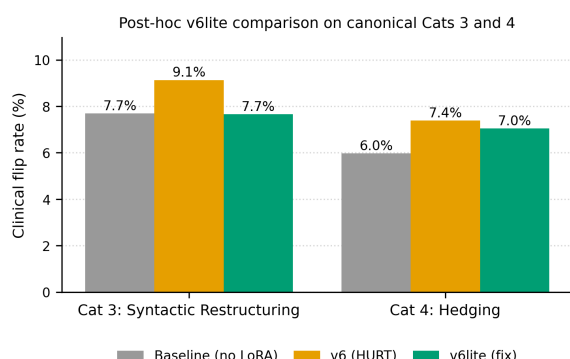


Figure 17: Post-hoc comparison of v6 and v6lite on canonical Cat 3 (Syntactic Restructuring) and Cat 4 (Hedging). The recipe choice is exploratory.

Table 12: Per-category LoRA on MEDGEMMA-4B. “Baseline” is recomputed on the paired slice where both pre/post runs emit yes/no, and therefore differs from the full Phase 1 denominator in Table 11. v6lite is a post-hoc diagnostic.

Cat	Name	Baseline	Post-LoRA	Δ abs (pp)	Δ rel	Recipe
1	Clinical Synonyms	13.91%	12.12%	−1.8	−12.9%	v6
2	Linate vs. Vernacular	17.75%	16.85%	−0.9	−5.1%	v6
3	Syntactic Restructuring	7.69%	7.51%	−0.2	−2.3%	v6lite
4	Hedging	5.97%	6.95%	+1.0	+16.4%	v6lite
5	Register Shift	13.39%	10.96%	−2.4	−18.1%	v6
6	Quantifier Insertion	10.36%	6.70%	−3.7	−35.3%	v6
7	Negation Reframing	47.99%	12.55%	−35.4	−73.9%	v6
8	Abbreviation Expansion	12.38%	11.24%	−1.1	−9.2%	v6
9	Certainty Modulation	14.49%	7.25%	−7.2	−50.0%	v6

Table 13: Canonical per-category v6 LoRA results on QWEN2.5-VL-3B. All nine improve descriptively after remapping source artifacts by category name; corrected significance testing was not repeated for this exploratory table.

Cat	Name	Baseline	Post-LoRA	Δ abs (pp)	Δ rel	Recipe
1	Clinical Synonyms	16.4%	10.83%	−5.6	−34.0%	v6
2	Linate vs. Vernacular	28.4%	14.78%	−13.7	−48.0%	v6
3	Syntactic Restructuring	16.0%	7.30%	−8.7	−54.3%	v6
4	Hedging	21.7%	5.39%	−16.3	−75.2%	v6
5	Register Shift	27.1%	9.75%	−17.3	−64.0%	v6
6	Quantifier Insertion	25.0%	7.25%	−17.7	−71.0%	v6
7	Negation Reframing	44.2%	12.87%	−31.3	−70.9%	v6
8	Abbreviation Expansion	22.3%	16.14%	−6.2	−27.7%	v6
9	Certainty Modulation	20.2%	6.43%	−13.7	−68.1%	v6

Table 14: Pre/post accuracy by canonical category on the paired yes/no slice. Base accuracy is measured on the base-question records retained for that category; paraphrase accuracy uses the literal token appropriate to the displayed question (inverted for Cat 7).

Cat	Name	MedGemma-4B				Qwen2.5-VL-3B-Instruct			
		Base pre	Base post	Para pre	Para post	Base pre	Base post	Para pre	Para post
1	Clinical Synonyms	79.9	77.2	77.1	76.7	71.6	75.5	71.3	78.7
2	Linate vs. Vernacular	79.6	74.2	74.1	76.5	71.7	74.8	67.3	71.5
3	Syntactic Restructuring	79.6	76.4	76.1	76.9	71.7	76.1	70.0	72.8
4	Hedging	79.6	76.0	78.2	75.4	71.7	76.5	70.3	75.8
5	Register Shift	79.6	75.6	73.6	74.7	71.7	73.5	63.3	71.9
6	Quantifier Insertion	79.5	72.8	75.3	71.8	71.6	66.8	69.7	67.8
7	Negation Reframing	79.6	77.8	55.1	75.9	71.7	75.2	57.7	71.7
8	Abbreviation Expansion	85.0	82.2	80.5	83.6	74.1	84.3	65.9	78.0
9	Certainty Modulation	79.6	76.4	73.7	77.0	71.7	72.6	70.5	71.2

Table 15: Recipe ablation on MEDGEMMA-4B Cat 7. Each recipe is trained from scratch. v4 is the meaningful within-setup comparator; the published 69.6% used a different dataset and severity regime.

Recipe	Layers	Modules	Epochs	λ_{KL}	Cat 7 flip	Δ rel	Notes
Baseline (no LoRA)	—	—	—	—	47.99%	—	—
SAE clamp {14026, 12744, 14479} $\rightarrow$ 0	—	—	—	—	45.21%	−5.8%	training-free
v1	15–19	attn	3	0	32.61%	−28.6%	PSF-Med defaults
v2	15–19	attn	6	0.5	23.57%	−48.4%	+ KL + 6ep
v3	12–21	attn	6	0.5	18.87%	−58.7%	broaden layers
v4	15–19	attn+MLP	3	1.0	20.43%	−55.3%	Mech-Guided match
v5	15–19	attn+MLP	3	1.0+het+curr	22.61%	−50.5%	het-rank hurts
v6	12–21	attn+MLP	6	1.0	12.43%	−74.09%	locked
Mech-Guided LoRA (Sadanandan and Behzadan, 2026a)	15–19	attn+MLP	3	1.0	—	−69.6%	benchmark

Table 16: Per-layer reversion rate from base→paraphrase residual patching at the last token on MEDGEMMA-4B. $N=300$ stratified subsample; 216 pairs admitted (both sides yes/no-answered).

Layer	Reversion rate	$n_{\text{reverted}}/n_{\text{admit}}$
0–14	0.0%	0/216
15	12.5%	27/216
16	12.5%	27/216
17	25.0%	54/216
18	62.5%	135/216
19	62.5%	135/216
20–23	62–75%	varies
24–26	86–87%	varies
27–30	87.5%	189/216
31	97.7%	211/216
32	99.1%	214/216
33	100.0%	216/216