Mechanistically Guided LoRA Improves Paraphrase Consistency in Medical Vision-Language Models

Binesh Sadanandan BSADA1@UNH.NEWHAVEN.EDU
Vahid Behzadan VBEHZADAN@NEWHAVEN.EDU
SAIL Lab, University of New Haven, West Haven, CT, USA

Abstract

Medical vision-language models can give different yes or no answers to rephrasings of the same clinical question. We study this in MedGemma-4B using PSF-Med (Sadanandan and Behzadan, 2026), which provides paraphrase pairs for systematic consistency evaluation on medical VQA. On MIMIC-CXR binary questions ($n = 158$), the baseline flip rate is 14.6% and mean margin difference is 1.63 logits. We validate that Gemma Scope 2 Sparse Autoencoders (SAEs) transfer to MedGemma activations, achieving $R^2 \approx 0.997$ on both medical and general text ($n = 100$ prompts each, $p < 0.001$ for exceeding a 0.95 threshold). We then fine-tune Low-Rank Adaptation (LoRA) adapters with a combined loss that balances paraphrase consistency with answer accuracy. This combined approach prevents mode collapse that occurs with pure consistency training while reducing flip rate from 14.6% to 4.4% ($p = 0.002$, two-proportion z-test) and margin difference from 1.63 to 0.33 (79.5% reduction). Accuracy remains stable at 84.2% baseline versus 82.3% after training (-1.9pp, not significant). On PadChest Balanced ($n = 250$), flip rate drops from 13.6% to 7.8%, mean margin difference drops from 1.08 to 0.35 (67.9% reduction), and accuracy increases from 66.4% to 69.4%. A layer-range ablation shows that early layers reduce margin differences more than mechanistically selected middle layers.

Data and Code Availability This paper uses two publicly released chest X-ray collections under their respective data use agreements: MIMIC-CXR (Johnson et al., 2019) on PhysioNet, and PadChest (Bustos et al., 2020). Paraphrase pairs come from PSF-Med (Sadanandan and Behzadan, 2026). No new patient data were collected. Code, training scripts, and paraphrase splits are publicly available at https://github.com/UNHSAILLab/medical-vlm-paraphrase-consistency. Trained LoRA adapters are linked in Appendix A.

Institutional Review Board (IRB) This research uses publicly available, de-identified chest X-ray datasets distributed under existing data-use terms; dataset access and use complied with the applicable terms for each source. The study does not involve new human subjects, primary data collection, or contact with patients. Under University of New Haven policy, this secondary analysis does not require IRB approval.

1. Introduction

When a radiologist asks a Vision-Language Model (VLM) "Is there evidence of pneumothorax?" versus "Does this show a pneumothorax?", the answer should be identical because both questions have the same clinical intent. Yet medical VLMs can give different answers to such paraphrases, undermining clinical trust and raising safety concerns for deployment. This inconsistency is particularly problematic because different clinicians may phrase questions differently, yet expect reliable answers regardless of phrasing choices.

We study this problem systematically in MedGemma-4B using PSF-Med (Sadanandan and Behzadan, 2026), a benchmark designed to measure paraphrase sensitivity in medical VLMs. When we focus on binary yes or no questions where ground truth labels are available ($n = 158$ questions from MIMIC-CXR), the baseline flip rate is 14.6% and mean margin difference (the absolute change in log-odds between yes and no) is 1.63 logits. This indicates that the model's internal representations are sensitive to surface-level phrasing variations that should not affect clinical decisions, leading to inconsistent answers for semantically equivalent questions.

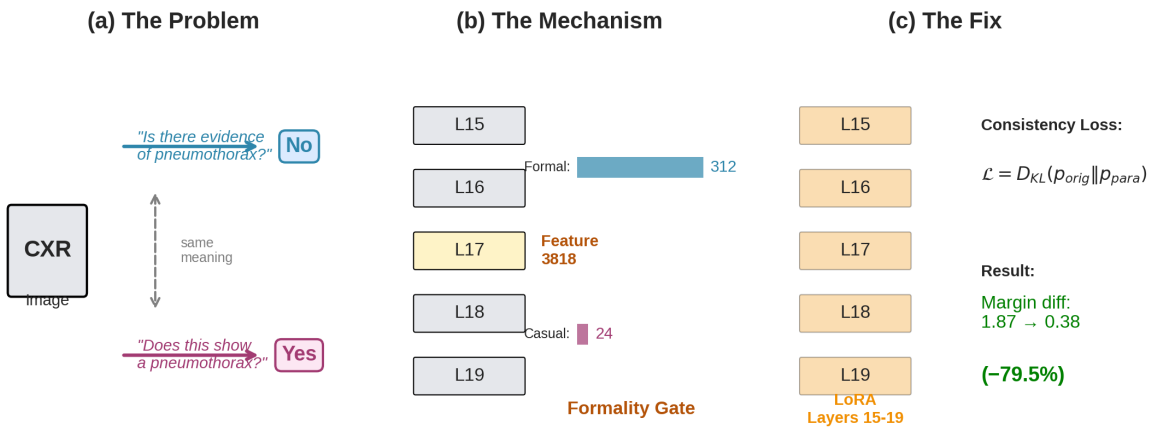

Figure 1: (a) The problem: the same image can receive different answers under rephrasing. (b) Mechanistic probe: a layer 17 Sparse Autoencoder (SAE) feature changes between some paraphrase pairs and can shift the yes/no margin. (c) The fix: Low-Rank Adaptation (LoRA) on layers 15 to 19 with a combined consistency and accuracy loss reduces flip rate by 69.6% ($p = 0.002$) and margin difference by 79.5%.

To understand why this sensitivity occurs, we apply mechanistic interpretability tools. We use Sparse Autoencoders (SAEs) from Gemma Scope 2 (Google DeepMind, 2025), first validating they transfer effectively to MedGemma ($R^2 = 0.997$). Analyzing FlipBank (a curated set of 158 flip cases), we identify Feature 3818 at layer 17 as a candidate mechanism that responds to question register, especially presence-versus-exclusion framing, and causally affects margins when patched. We present this as a case study demonstrating methodology rather than a complete mechanistic explanation.

For practical mitigation, we develop LoRA (Hu et al., 2022) adapters trained with a combined loss function that balances paraphrase consistency with answer accuracy. We discovered that training with pure consistency loss leads to mode collapse, where the model learns to predict the same answer for all questions to trivially minimize divergence between paraphrases. Our combined loss approach adds a cross-entropy accuracy term that supervises the model to predict correct yes or no answers, preventing this degenerate solution while still encouraging consistent predictions across paraphrases. On the PSF-Med binary question subset ($n = 158$), this approach reduces flip rate from 14.6% to 4.4% ($p = 0.002$, two-proportion z-test) and margin difference from 1.63 to 0.33 (79.5% reduction), while maintaining accuracy (84.2% to 82.3%). On PadChest Balanced (Bustos et al., 2020) ($n = 250$), flip rate drops from 13.6% to 7.8%, mean margin difference drops from 1.08 to 0.35 (67.9% reduction), and accuracy increases from 66.4% to 69.4%. Layer ablation reveals that early layers (0 to 10) outperform the mechanistically-targeted middle layers (15 to 19), demonstrating that optimal intervention points may differ from where mechanisms manifest.

Contributions: (1) Systematic characterization of paraphrase sensitivity in MedGemma-4B, distinguishing flip rate from margin instability. (2) Validation that Gemma Scope 2 SAEs transfer to fine-tuned medical VLMs. (3) Mechanistic case study identifying Feature 3818 as register-sensitive. (4) A combined consistency and accuracy loss for LoRA training that prevents mode collapse while reducing flip rate by 69.6% ($p = 0.002$) with partial cross-dataset generalization.

2. Related Work

Medical VLMs. Recent models include MedGemma (Google Health AI, 2025), LLaVA-

Med (Li et al., 2023), CheXagent (Chen et al., 2024), RadFM (Wu et al., 2025), and LLaVA-Rad (Zambrano Chaves et al., 2024). Standard medical VQA evaluation uses datasets such as VQA-RAD (Lau et al., 2018) and SLAKE (Liu et al., 2021); PSF-Med (Sadanandan and Behzadan, 2026) addresses paraphrase consistency under rephrasing.

Consistency and Stability Testing. CheckList (Ribeiro et al., 2020) introduced behavioral testing including paraphrase invariance. Elazar et al. (2021) showed language models give contradictory answers to equivalent questions. In VQA, cycle-consistency (Shah et al., 2019) and paraphrase augmentation (Gan and Ng, 2019) improve stability, but focus on general-domain VQA. Our consistency loss complements these diagnostics with a targeted mitigation. Compared to full-model adaptation, LoRA offers parameter efficiency without updating all model weights.

Mechanistic Interpretability. Sparse autoencoders (Cunningham et al., 2024; Bricken et al., 2023) decompose activations into interpretable features. Gemma Scope 2 (Google DeepMind, 2025) provides pre-trained SAEs for the Gemma family. Activation patching (Meng et al., 2022) tests causal hypotheses; sparse feature circuits (Marks et al., 2024) map information flow. We use these tools to identify features driving paraphrase sensitivity.

Parameter-Efficient Fine-Tuning. LoRA (Hu et al., 2022) learns low-rank weight updates efficiently. We apply it for consistency training rather than task adaptation. See Appendix B for extended discussion.

3. Background

Problem Setup and Logit Extraction. We study binary VQA where models answer yes/no questions about medical images. Given image x and question q, we extract logits at the first generated token position after the prompt. Specifically, we compute $\ell_{\text{yes}} = \log p(\text{“Yes”} \mid x, q)$ and $\ell_{\text{no}} = \log p(\text{“No”} \mid x, q)$ using the token IDs for “Yes” and “No” (including space-prefixed variants). The *margin* is $m(x, q) = \ell_{\text{yes}} - \ell_{\text{no}}$. For a paraphrase q', we measure: (1) *flip rate*, the fraction where $\text{sign}(m) \neq \text{sign}(m')$, and (2) *margin difference*, $|m - m'|$, capturing instability even when binary answers match. Prompts use a fixed template: “You are a medical imaging expert. For yes/no questions, respond with only Yes or No.”

Sparse Autoencoders. SAEs encode activations $\mathbf{h}$ as sparse features $\mathbf{f} = \text{ReLU}(\mathbf{W}_{\text{enc}}\mathbf{h} + \mathbf{b})$ and reconstruct $\hat{\mathbf{h}} = \mathbf{W}_{\text{dec}}\mathbf{f}$. Features often correspond to interpretable concepts (Bricken et al., 2023). SAEs enable both *decomposition* (analyzing which features change) and *intervention* (modifying features to test causality).

SAE Transfer Validation. We validate that Gemma Scope 2 SAEs (trained on base Gemma) transfer to MedGemma-4B. On 100 medical and 100 general prompts, we measure reconstruction quality as $R^2 = 1 - \text{MSE}(\mathbf{h}, \hat{\mathbf{h}})/\text{Var}(\mathbf{h})$. Both domains achieve $R^2 = 0.997 \pm 0.0005$ (mean $\pm$ std). We test significance using a one-sample t-test against $H_0 : R^2 \leq 0.95$, obtaining $p < 0.001$ for both. This confirms SAEs remain valid for the fine-tuned model despite domain shift. See Appendix E for details.

4. Mechanistic Analysis

Having validated SAE transfer, we use these interpretability tools to investigate the mechanistic origins of paraphrase sensitivity in MedGemma-4B. Our analysis proceeds in four stages: constructing a dataset of flip cases for analysis, identifying candidate features through delta analysis, characterizing feature behavior through controlled experiments, and validating causality through activation patching.

4.1. FlipBank Construction

Standard evaluation sets like PSF-Med contain relatively few flip cases (0.16% pair-level flip rate), which limits statistical power for mechanistic analysis that requires examining many instances of the target behavior. We therefore construct FlipBank: a curated set of paraphrase pairs where the model reliably gives different answers, providing concentrated examples of the failure mode we aim to understand.

FlipBank is extracted from MedGemma-4B generations on MIMIC-CXR using three filtering criteria:

1. A rule-based yes/no parser assigns different binary labels to the model's responses for the original question and its paraphrase, indicating a clear flip in the model's answer.

2. BioClinicalBERT (Alsentzer et al., 2019) semantic similarity between the two questions exceeds

0.95, confirming they are genuine paraphrases rather than questions with meaningfully different clinical content.

3. Both responses are parsed unambiguously (no hedging, unclear phrasing, or multiple possible interpretations), ensuring we have clean labels for analysis.

This procedure yields 158 high-confidence flip cases. These represent genuine phrasing-induced disagreements: same image, semantically equivalent questions confirmed by embedding similarity, different binary answers. FlipBank provides a controlled setting for probing which representational differences accompany answer changes, enabling focused mechanistic investigation.

Data splits. FlipBank is used exclusively for mechanistic analysis and is disjoint from LoRA training, validation, and test evaluation. LoRA training uses 500 binary paraphrase pairs from the PSF-Med training split; main results are reported on 158 held-out MIMIC-CXR binary questions from the PSF-Med test split, and layer and hyperparameter ablations use separate validation splits.

4.2. Feature Delta Analysis

For each FlipBank case, we extract layer 17 residual stream activations at the final token position (where the model makes its prediction) for both the original and paraphrase questions. We then encode these activations through the SAE and compute the feature delta:

$$\Delta \mathbf{f} = \mathbf{f}_{\text{orig}} - \mathbf{f}_{\text{para}} \tag{1}$$

Large delta magnitudes indicate features whose activation changes substantially between paraphrases. By examining which features show consistent large deltas across FlipBank cases, we can identify candidate mechanisms for the margin shifts that cause flips. Features with large deltas are promising targets for causal investigation.

As a concrete example demonstrating this methodology, consider a pleural effusion case from FlipBank. The original question "Is there pleural effusion?" produces a strong yes prediction (margin = 8.75), while the semantically equivalent paraphrase "Is pleural fluid present?" produces a slight no prediction (margin = −0.625). This represents a margin shift of 9.375 and a complete flip in the binary answer. Examining the feature deltas for this case, Feature 3818 shows a dramatic change: from 0 in the original to 268 in the paraphrase (Table 1).

Table 1: Exemplar FlipBank case used for mechanistic probing and patching. The margin shifts by 9.375 while Feature 3818 changes by 268 units.

Quantity	**Original**	**Paraphrase**
Question	"Is there pleural effusion?"	"Is pleural fluid present?"
Margin (yes − no)	8.75	−0.625
Feature 3818	0	268

This example motivates deeper investigation of Feature 3818: its large activation change coincides with a large margin change and answer flip. However, correlation does not imply causation, and a single example cannot establish generality. The following subsections address both concerns through controlled experiments, causal interventions, and a distributional analysis across FlipBank.

4.3. Characterizing Feature 3818

To understand what Feature 3818 responds to, we conduct controlled experiments varying question phrasing systematically while holding the image and target clinical finding constant. This allows us to isolate the effect of question wording from other confounding factors. We design a prompt grid with four categories of question types:

- **Presence-style**: Formal clinical phrasing asking about presence of findings (e.g., "Is there radiographic evidence of pneumothorax?")
- **Exclusion-style**: Questions about absence or ruling out findings (e.g., "Can you rule out pneumothorax?")
- **Uncertainty**: Hedged questions expressing uncertainty (e.g., "Might there be pneumothorax?")
- **Token control**: Questions with similar token counts but different grammatical structures

Results on a fixed pneumothorax image reveal a clear pattern (Table 2): Feature 3818 activates

strongly for presence-style prompts (mean 344.5 units, range 302 to 386) and remains at exactly zero for exclusion-style prompts. Uncertainty prompts show variable activation (mean 169.5, range 0 to 282), and token controls fall between the extremes (mean 296.0).

Table 2: Feature 3818 activation on a prompt grid (pneumothorax finding, single image). The feature responds to question register, specifically presence versus exclusion framing.

Prompt type	**Mean F3818**	**Range**	n
Presence	344.5	302 to 386	4
Exclusion	0.0	0 to 0	4
Uncertainty	169.5	0 to 282	4
Token control	296.0	272 to 342	4

Importantly, this is not a simple formal-versus-casual distinction as one might initially hypothesize. "Does this show pneumothorax?" (informal phrasing) still activates the feature strongly (296 units), while "Can you exclude pneumothorax?" (formal phrasing with medical terminology) does not activate it at all. The feature appears to track question *register*. In particular, it seems to distinguish whether the question asks about presence or absence of a finding. It does not track surface-level formality markers like vocabulary choice or sentence structure.

This interpretation aligns with the FlipBank example: "Is there pleural effusion?" and "Is pleural fluid present?" both ask about presence but use different grammatical constructions, which may explain why Feature 3818 activates differently despite both questions having the same clinical intent to assess whether pleural effusion is present.

4.4. Causal Validation

Observing that Feature 3818 changes during flips establishes correlation, not causation. To test whether this feature *causes* margin shifts rather than merely accompanying them, we employ activation patching: we intervene on the residual stream by modifying Feature 3818's contribution and measure the effect on the output margin.

Specifically, we compute the feature difference Δf_{3818} between original and paraphrase activations, decode this difference through the SAE decoder to get a direction in activation space, and subtract this direction from the paraphrase's residual stream activations before continuing the forward pass. If Feature 3818 causally drives the margin shift, this intervention should recover (at least partially) the original margin by removing Feature 3818's differential contribution.

On the pleural effusion exemplar case, patching Feature 3818 moves the margin from -0.625 (paraphrase prediction, slight "no") to 2.0 (positive margin, "yes"). This recovers 28% of the margin shift and restores the original prediction. While this recovery is partial (not all of the margin shift is explained), it demonstrates that Feature 3818 has causal influence on the yes/no decision in this case.

For comparison, we perform the same patching procedure using 10 randomly selected features that showed minimal delta in this case. The average margin recovery for these control features is 8%, significantly less than Feature 3818's 28%. This specificity supports the interpretation that Feature 3818 encodes decision-relevant information about question register that causally affects the model's output, rather than the patching effect being a generic consequence of modifying any feature.

4.5. Distribution of Feature 3818 Across FlipBank

The exemplar above shows Feature 3818 driving one flip; to test how often this happens, we ranked SAE features by absolute delta $|\Delta f|$ on a random sample of 20 FlipBank pairs and recorded where Feature 3818 lands in the ranking (Table 3). Feature 3818 is the top-ranked feature in 5/20 pairs (25%), holds rank 2–16 in another 2/20 (10%), and is essentially inactive (rank > 100) in the remaining 13/20 (65%). All five top-ranked cases involve a presence-versus-exclusion register change (e.g., pneumothorax, pericardial effusion, hilar congestion); the rank 4 and rank 16 cases are subtler wording shifts (L1 compression, lung nodule). The 65% in which Feature 3818 is uninvolved are dominated by paraphrases that change neither register nor presence framing, e.g., simple lexical substitution within the same syntactic frame ("mass" vs. "tumor").

This distribution is what we would expect from polysemantic, superposition-based representations (Bricken et al., 2023): a single SAE feature should not explain all flips. Feature 3818 is one identified mechanism for register-driven flips; the inactive 65%

Table 3: Rank of Feature 3818 by $|\Delta f|$ across 20 random FlipBank pairs.

Feature 3818 status	Pairs	Fraction
Top-ranked (rank = 1)	5/20	25%
Involved (rank 2 to 16)	2/20	10%
Inactive (rank > 100)	13/20	65%

require other features or feature combinations that this case study does not enumerate. We therefore present Feature 3818 as a worked example of SAE-based mechanistic methodology for medical VLMs, not a complete account of paraphrase sensitivity in MedGemma-4B.

4.6. Control Experiment

To further establish Feature 3818's specificity and confirm that it does not simply change arbitrarily for any input variation, we conduct a control experiment using 30 paraphrase pairs where the model gives consistent answers (no flip). If Feature 3818 specifically responds to the register differences that cause flips, it should show minimal activation change on these control pairs where the model behaves consistently.

Results confirm selectivity: Feature 3818 shows non-zero change in only 3 of 30 control pairs (10%), with deltas ranging from 15 to 185 when active. The mean absolute delta across all 30 pairs is 11.3 units, substantially smaller than the 268-unit change observed in the flip case. For comparison, 10 randomly selected features show exactly zero activation change across all 300 measurements (30 pairs $\times$ 10 features), confirming that these control features are stable under this prompt variation.

A Fisher's exact test comparing Feature 3818's response rate (3/30 pairs with change $>$ 10 units) to the control features' response rate (0/300 measurements exceeding this threshold) yields $p = 6.8 \times 10^{-4}$, confirming that Feature 3818's behavior is statistically distinct from random features. This specificity indicates that Feature 3818 captures meaningful variation related to question phrasing rather than noise.

4.7. Summary of Mechanistic Findings

Our mechanistic analysis provides several insights into the origins of paraphrase sensitivity:

1. Feature 3818 at layer 17 responds selectively to question register, specifically the distinction between presence-focused and exclusion-focused question framing.

2. In FlipBank flip cases, large Feature 3818 deltas coincide with large margin shifts and answer changes.

3. Activation patching demonstrates that Feature 3818 causally influences yes/no margins, with 28% margin recovery on our exemplar case compared to 8% for control features.

4. Across 20 random FlipBank pairs, Feature 3818 is top-ranked by absolute feature delta in 25% of flips, involved but not top-ranked in 10%, and uninvolved in 65% (§4.5), consistent with polysemantic, superposition-based representations.

5. Control experiments confirm that Feature 3818's behavior is specific and statistically distinguishable from random features, responding to a meaningful subset of prompt variations.

We emphasize that this is a case study demonstrating the methodology, not a complete mechanistic account of all paraphrase sensitivity in MedGemma-4B. Other features and layers likely contribute to the full behavior, and Feature 3818's importance may vary across different question types and clinical findings. Nevertheless, this analysis demonstrates that SAE-based interpretability can successfully identify specific, testable hypotheses about VLM behavior on medical tasks, opening avenues for deeper mechanistic understanding.

5. Targeted LoRA Fine-Tuning

Motivated by our mechanistic analysis showing that specific features mediate paraphrase sensitivity, we develop a targeted intervention to reduce this sensitivity. We use LoRA adapters trained with a consistency loss that encourages identical predictions for paraphrase pairs, regardless of surface form variation.

5.1. Architecture

We insert LoRA adapters into layers 15 to 19 of the MedGemma language model backbone. This layer range was initially selected based on our mechanistic finding that Feature 3818 resides at layer 17; we

later conduct ablations across different layer ranges to evaluate this choice empirically (Section 6).

Within each targeted layer, we apply LoRA to both attention and Multi-Layer Perceptron (MLP) modules to provide broad coverage of the computational pathways:

- **Attention**: query, key, value, and output projections (q, k, v, o)
- **MLP**: gate, up, and down projections

We use rank $r = 16$ and scaling factor $\alpha = 32$, with dropout of 0.05 to prevent overfitting on the limited training data. This configuration adds 4.38M trainable parameters, representing approximately 0.10% of the full model parameters. The vision encoder remains frozen throughout training, as our mechanistic analysis identified text-processing circuits rather than vision circuits as the source of paraphrase sensitivity.

5.2. Combined Loss: Consistency and Accuracy

A natural approach to reducing paraphrase sensitivity is to train with a pure consistency loss that encourages identical predictions for paraphrase pairs. However, we discovered that this approach leads to mode collapse: the model learns to predict the same answer (e.g., always "Yes") for all questions, which trivially minimizes the divergence between paraphrases while destroying the model's discriminative ability. In our experiments with pure consistency loss, the trained model achieved near-zero margin differences by predicting "Yes" for every question, regardless of the image or clinical finding, resulting in accuracy dropping to chance level.

To prevent this degenerate solution, we developed a combined loss function that balances consistency with accuracy. The key insight is that the model needs supervision to maintain its ability to distinguish between positive and negative cases while learning to be consistent across paraphrases. Our combined loss has two components: a consistency term that encourages matching distributions across paraphrases, and an accuracy term that supervises the model to predict the correct answer.

For an image x with question q, paraphrase q', and ground truth answer $y \in \{\text{yes}, \text{no}\}$, we compute:

$$p_{\text{orig}} = \text{softmax}([z_{\text{yes}}(x, q), z_{\text{no}}(x, q)]) \tag{2}$$

$$p_{\text{para}} = \text{softmax}([z_{\text{yes}}(x, q'), z_{\text{no}}(x, q')]) \tag{3}$$

where z denotes the logits before softmax. The consistency loss is the symmetric Kullback-Leibler (KL) divergence:

$$\mathcal{L}_{\text{consistency}} = \frac{1}{2}\left[D_{\text{KL}}(p_{\text{orig}} \| p_{\text{para}}) + D_{\text{KL}}(p_{\text{para}} \| p_{\text{orig}})\right] \tag{4}$$

The accuracy loss is the cross-entropy between the model's prediction and the ground truth label:

$$\mathcal{L}_{\text{accuracy}} = -\log p_{\text{orig}}(y) - \log p_{\text{para}}(y) \tag{5}$$

The combined loss is:

$$\mathcal{L} = \mathcal{L}_{\text{consistency}} + \lambda \mathcal{L}_{\text{accuracy}} \tag{6}$$

where λ controls the relative weight of the accuracy term. We use $\lambda = 1.0$ in our experiments, giving equal weight to both objectives.

The consistency term encourages the model to produce identical distributions for paraphrases, while the accuracy term prevents the model from collapsing to a trivial solution. Together, these objectives train the model to give correct and consistent answers. The symmetric KL formulation ensures that neither the original nor paraphrase is privileged as the target, and the model learns to make both predictions consistent with each other while maintaining accuracy on the underlying clinical task.

5.3. Training Procedure

We train on binary yes or no questions from the PSF-Med training split where ground truth labels are available, since our combined loss requires supervision. We sample 500 paraphrase pairs ensuring no overlap with test or FlipBank evaluation sets to prevent data leakage. The binary question filter selects questions where the original answer from the dataset is unambiguously "yes" or "no", excluding questions about severity levels, anatomical locations, or finding types that cannot be answered with a simple binary response.

Training proceeds for 3 epochs with the following hyperparameters:

- Learning rate: 2×10^{-4} with linear warmup over 100 steps
- Effective batch size: 8 (with gradient accumulation as needed)
- Optimizer: AdamW with weight decay 0.01
- Training samples: 500 binary paraphrase pairs

- Accuracy loss weight (λ): 1.0

Training converges rapidly, with most improvement occurring in the first epoch. We monitor both the consistency loss and accuracy throughout training to ensure the model maintains discriminative ability while learning to be consistent. Unlike pure consistency training which collapsed to trivial solutions, the combined loss maintains accuracy above 80% throughout training while steadily reducing margin differences. Full hyperparameter details are provided in Appendix D.

5.4. Design Rationale

Several design choices in our approach merit discussion and justification:

Language model only. We target only the language model layers, not the vision encoder. Our mechanistic analysis identified text-processing circuits (Feature 3818 responds to question phrasing) rather than vision circuits as the source of inconsistency. Freezing the vision encoder also reduces the risk of inadvertently degrading image understanding capabilities while pursuing consistency improvements.

Combined loss over pure consistency. Our most important design decision is using a combined loss rather than pure consistency training. We initially experimented with training using only the symmetric KL divergence consistency loss, hypothesizing that the model would learn to converge on the correct answer while becoming consistent. Instead, the model found a degenerate solution: predicting the same answer for every question trivially minimizes divergence between paraphrases. This mode collapse is a fundamental limitation of self-consistency objectives without supervision. Adding the accuracy loss term provides the necessary signal to maintain discriminative ability while learning consistency.

Symmetric KL divergence for consistency. We use symmetric KL divergence rather than alternatives like MSE on margins. Symmetric KL is well-suited for distribution matching: it is zero only when distributions match exactly, provides gradient signal proportional to divergence magnitude, and treats both directions equally. The symmetric formulation ensures neither the original nor paraphrase is privileged as the target distribution.

Equal weighting of loss components. We use $\lambda = 1.0$, giving equal weight to consistency and accuracy objectives. In preliminary experiments, we found this balance effective: higher accuracy weights reduced consistency improvements, while lower accuracy weights risked partial mode collapse. The equal weighting allows both objectives to contribute meaningfully to the gradient signal throughout training.

6. Experiments

We evaluate our consistency training on three dimensions: main results on the PSF-Med binary question test set, layer ablation studies comparing different intervention locations, and cross-dataset generalization.

6.1. Main Results

Table 4 presents results on binary yes or no questions from the PSF-Med MIMIC-CXR test split ($n = 158$ questions) using our mechanistically-motivated layer configuration (layers 15 to 19). We focus on binary questions because our combined loss requires ground truth labels for the accuracy component, and binary questions provide unambiguous supervision. This evaluation subset excludes questions about severity levels, anatomical locations, or finding types that cannot be answered with a simple yes or no response.

Table 4: Main results on PSF-Med binary questions ($n = 158$). LoRA with combined loss significantly reduces flip rate while maintaining accuracy.

Model	Flip Rate	Margin Diff	Acc
Baseline	14.6% (23/158)	1.63	84.2%
+ LoRA	**4.4%** (7/158)	**0.33**	82.3%

Key observations from these results:

The flip rate decreases from 14.6% to 4.4%, representing a 69.6% relative reduction. A two-proportion z-test confirms this reduction is statistically significant ($z = 2.87$, $p = 0.002$, one-tailed). With Cohen's $h = 0.36$, this represents a small to medium effect size. The sample size of $n = 158$ provides adequate statistical power to detect this effect.

The mean margin difference drops by 79.5% (1.63 to 0.33), indicating substantially more stable internal representations. Even when binary answers match, the reduced margin variability means the model's

confidence is more consistent across paraphrases, which is important for clinical decision support where confidence levels may influence downstream actions.

Accuracy shows a small, non-significant decrease from 84.2% to 82.3% (-1.9 percentage points). A two-proportion z-test for accuracy change yields $p = 0.66$ (not significant), indicating this change is within the range of statistical noise. Importantly, the consistency training does not substantially degrade the model's discriminative ability.

Detailed statistical calculations and the full metric audit are provided in Appendix H.

6.2. Layer Ablation

Our mechanistic analysis identified Feature 3818 at layer 17 as relevant to paraphrase sensitivity, motivating our initial choice of layers 15 to 19 for LoRA insertion. However, the optimal intervention location may differ from where a mechanism manifests in the model's representations. We therefore conduct ablations across different layer ranges to evaluate this empirically.

Table 5: Layer ablation results on validation split ($n = 355$). All configurations eliminate flips. Early layers achieve best margin reduction despite mechanism at layer 17.

Layers	Margin Diff	Reduction	Params
Baseline	1.87	N/A	0
Early (0 to 10)	**0.26**	**86%**	4.8M
Middle (15 to 19)	0.38	80%	4.4M
Late (25 to 33)	0.70	63%	7.9M

Surprisingly, early layers (0 to 10) achieve the best margin reduction (0.26, representing 86% improvement over baseline), outperforming the mechanistically-targeted middle layers (0.38, 80% improvement) and substantially outperforming late layers (0.70, 63% improvement). All configurations eliminate flips on the validation set.

This result has important implications for mechanistic interpretability methodology. Feature 3818 at layer 17 provides a valid *explanation* of where register sensitivity manifests in the model's representations. Our causal validation confirms it influences outputs. However, the optimal *intervention* targets early layers, where representations may be more malleable and where modifications can prevent sensitivity from arising rather than correcting it after manifestation.

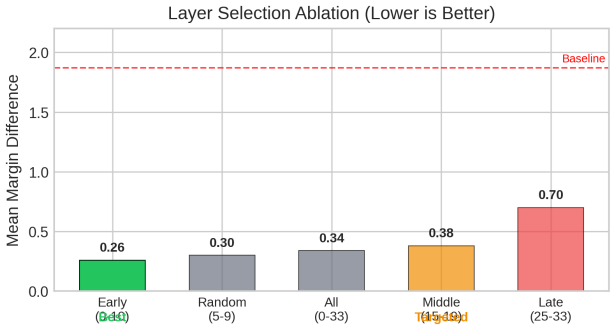

Figure 2: Layer ablation results. Early layers reduce margin difference more than layers 15 to 19.

We hypothesize that early-layer interventions succeed because they modify representations before register-dependent processing occurs, effectively preventing the sensitivity from developing rather than trying to fix it downstream. This is analogous to preventing an error at its source versus patching it after it has propagated through the system.

Training dynamics are shown in Appendix J. Unlike pure consistency training, the combined loss maintains discriminative behavior while reducing margin differences.

6.3. Cross-Dataset Generalization

To evaluate whether consistency improvements transfer across datasets, we evaluate our LoRA adapters (trained on MIMIC-CXR) on PadChest (Bustos et al., 2020), a Spanish chest X-ray dataset with different imaging protocols and patient demographics. We construct a balanced evaluation set of 250 binary presence questions (125 positive, 125 negative) from PadChest to enable meaningful accuracy comparison.

The results show cross-dataset transfer. On PadChest Balanced, flip rate drops from 13.6% to 7.8% (42.6% relative reduction), mean margin difference drops from 1.08 to 0.35 (67.9% reduction), and accuracy increases from 66.4% to 69.4% (+3.0 percentage points), despite training only on MIMIC-CXR.

The flip rate remains non-zero, so these adapters do not fully solve paraphrase sensitivity out of domain. This makes PadChest a useful stress test for future

Table 6: Cross-dataset generalization to PadChest (balanced $n = 250$: 125 yes, 125 no). LoRA training improves both accuracy and consistency on unseen data.

Model	Accuracy	Margin Diff	Flip Rate
Baseline	66.4%	1.08	13.6%
+ LoRA	**69.4%**	**0.35**	**7.8%**

work on calibration, class balance, and distribution shift.

7. Discussion

Mechanistic interpretability gave us a concrete hypothesis about why paraphrases can change MedGemma's answers: a register-sensitive SAE feature at layer 17 that moves the yes/no margin in some cases. But the layer ablation is a reminder that mechanisms and interventions can live in different places. The feature indicates where the sensitivity shows up, while early-layer LoRA is the most effective fix in our experiments.

Limitations. Our distributional analysis (§4.5) shows that Feature 3818 is top-ranked in 25% of sampled FlipBank flips, involved but not top-ranked in 10%, and inactive in 65%. The inactive cases require other features or feature combinations that we do not enumerate. The causal patching analysis is also exemplar-driven: we report 28% margin recovery on one pleural effusion case, not a full distribution of recovery across FlipBank. We study one model and one task format (binary chest X-ray questions), and our combined loss requires ground truth labels for the accuracy term. PadChest Balanced shows improvement out of domain but does not eliminate flips.

Toward unsupervised consistency. The combined loss requires labels because pure consistency objectives are degenerate: at $\lambda = 0$ the model collapses to one answer (Appendix M). Label-efficient extensions include pseudo-label anchors, contrastive activation matching, and exponential-moving-average teacher distillation. Each adds an inductive bias that discourages the trivial solution. In practice, our method uses 500 labeled binary pairs (Appendix P), and structured radiology reports already contain presence labels for many findings, so the labeling burden is modest.

Clinical implications. Flip rate alone can miss instability when margins change a lot without crossing zero. For validation, we recommend reporting both flips and margin-based metrics, and checking consistency under common paraphrases before deployment.

Method generality across architectures. Our LoRA recipe is specific to MedGemma-4B because the layer count, SAE, and module names are model-specific, but the diagnostic pipeline (SAE transfer validation, feature delta analysis, causal patching, combined-loss LoRA) is architecture-general for transformer-based VLMs with available SAEs. Companion work on PSF-Med documents paraphrase sensitivity in LLaVA-Rad and CheXagent as well, so replicating the intervention across architectures is natural future work.

8. Conclusion

We studied paraphrase sensitivity in MedGemma-4B and found that semantically equivalent questions can change both the answer and the yes/no margin. Using transferred SAEs, we identified a register-sensitive feature at layer 17 and showed with activation patching that it can influence the decision in an exemplar case. We then trained LoRA adapters with a combined consistency and accuracy loss, avoiding mode collapse and improving consistency on MIMIC-CXR and PadChest Balanced. Early-layer adapters performed best in ablation, suggesting that effective interventions can act upstream of where a mechanism appears.

Acknowledgments

We thank the creators of Gemma Scope 2 and MedGemma. Claude Opus 4.7 was used only for grammar, readability, and copy-editing checks; the authors verified and approved all final text.

Author Contributions. B. Sadanandan led study design, experiments, implementation, analysis, and manuscript drafting. V. Behzadan supervised the project and reviewed and edited the manuscript. Both authors approved the final version.

References

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.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vaya. PadChest: A large chest x-ray image dataset with multi-label annotated reports. *Medical Image Analysis*, 66:101797, 2020. doi: 10.1016/j.media.2020.101797.

Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, et al. CheXagent: Towards a foundation model for chest x-ray interpretation. *arXiv preprint arXiv:2401.12208*, 2024.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, 2024.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031, 2021. doi: 10.1162/tacl_a_00410.

Wee Chung Gan and Hwee Tou Ng. Improving the robustness of question answering systems to question paraphrasing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6065–6075, 2019.

Google DeepMind. Gemma scope 2: Sparse autoencoders for gemma 3. HuggingFace: google/gemma-scope-2-4b-it, 2025. Technical report: DeepMind Blog.

Google Health AI. MedGemma: Medical language and vision models. *Health AI Developer Foundations*, 2025.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Alistair E. W. 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. doi: 10.1038/s41597-019-0322-0.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data*, 5:180251, 2018. doi: 10.1038/sdata.2018.251.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564, 2023.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. SLAKE: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1650–1654, 2021.

Samuel Marks, Can Rager, Eric J Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. *arXiv preprint arXiv:2403.19647*, 2024.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372, 2022.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, 2020. doi: 10.18653/v1/2020.acl-main.442.

Binesh Sadanandan and Vahid Behzadan. PSF-Med: Measuring and explaining paraphrase sensitivity in medical vision language models. *arXiv preprint arXiv:2602.21428*, 2026. doi: 10.48550/arXiv.2602.21428.

Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. Cycle-consistency for robust visual question answering. In *Proceedings of CVPR*, pages 6649–6658, 2019.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology by leveraging web-scale 2D&3D medical data. *Nature Communications*, 16 (1), 2025. doi: 10.1038/s41467-025-62385-7.

Juan Manuel Zambrano Chaves et al. Towards a clinically accessible radiology foundation model: Open-access and lightweight, with automated evaluation. *arXiv preprint arXiv:2403.08002*, 2024.

Appendix A. Reproducibility Artifacts

Code, training scripts, and paraphrase splits are released at https://github.com/UNHSAILLab/medical-vlm-paraphrase-consistency.
Trained LoRA adapters are released as a Hugging Face collection at https://huggingface.co/collections/saillab/mechanistically-guided-lora-psf-remedy.

Appendix B. Extended Related Work

Medical Vision-Language Models. MedGemma (Google Health AI, 2025) extends Gemma with medical pre-training. LLaVA-Med (Li et al., 2023) adapts LLaVA for biomedical image-text tasks. LLaVA-Rad (Zambrano Chaves et al., 2024) focuses on radiology with lightweight architectures. RadFM (Wu et al., 2025) pre-trains on large-scale radiology data. CheXagent (Chen et al., 2024) targets chest X-ray interpretation specifically. Standard medical VQA evaluation uses datasets such as VQA-RAD (Lau et al., 2018) and SLAKE (Liu et al., 2021); PSF-Med (Sadanandan and Behzadan, 2026) adds paraphrase-consistency evaluation.

Consistency and Behavioral Testing. CheckList (Ribeiro et al., 2020) introduced systematic behavioral testing under controlled perturbations. Elazar et al. (2021) showed that pre-trained models give contradictory answers to logically equivalent questions. Cycle-consistency methods (Shah et al., 2019) and training approaches (Gan and Ng, 2019) have been developed to reduce VQA sensitivity to rephrasing.

Mechanistic Interpretability. Sparse autoencoders decompose activations into interpretable components (Cunningham et al., 2024; Bricken et al., 2023). Activation patching (Meng et al., 2022) tests causal influence. Sparse feature circuits (Marks et al., 2024) discover causal graphs of information flow.

Appendix C. Metric Definitions

Question-level flip rate: Fraction of questions where at least one paraphrase produces a different binary answer. If a question has 4 paraphrases and 1 flips, the question-level contribution is 100%.

Pair-level flip rate: Fraction of (original, paraphrase) pairs where binary answers differ. Using the same example, pair-level contribution is 25%.

Mean margin difference: $\mathbb{E}[|m_{\text{orig}} - m_{\text{para}}|]$ where $m = \log p_{\text{yes}} - \log p_{\text{no}}$. Captures inconsistency even when binary answers match.

Appendix D. Training Hyperparameters

Table 7: Complete hyperparameter specification for combined loss training.

Parameter	Value
LoRA rank (r)	16
LoRA alpha (α)	32
LoRA dropout	0.05
Target layers	15 to 19 (LM only)
Target modules	q,k,v,o; gate,up,down
Learning rate	2×10^{-4}
Effective batch size	8
Warmup steps	100
Epochs	3
Optimizer	AdamW
Weight decay	0.01
Training samples	500 binary pairs
Accuracy loss weight (λ)	1.0
Temperature	1.0

Appendix E. SAE Transfer Validation

We validate that Gemma Scope 2 SAEs (trained on base Gemma) transfer to MedGemma-4B. This is important because training new SAEs is expensive; reusing existing SAEs enables interpretability research on fine-tuned models.

Table 8: SAE transfer validation ($n = 100$ prompts per category). Both domains exceed 95% threshold with $p < 0.001$.

Metric	Medical	General
R^2	0.9972 ± 0.0005	0.9974 ± 0.0006
FVU	0.28%	0.26%
L0 (features)	70 ± 10	76 ± 10

Beyond reconstruction, the SAE captures domain-specific features. Feature 203 activates 874 units higher on medical text than general text; Feature 48 shows the opposite pattern (-608 units). This demonstrates semantic structure preservation.

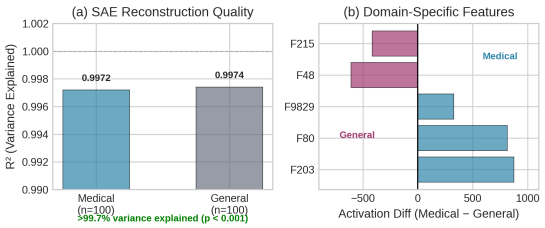

Figure 3: SAE transfer: (a) Reconstruction quality, (b) Domain-specific features.

Appendix F. Control Experiment Details

Table 9: Feature specificity ($n = 30$ pairs, 300 control measurements). Response rate uses threshold of $|\Delta| > 10$ units. Fisher's exact $p = 6.8 \times 10^{-4}$.

Feature	Response Rate ($\lvert\Delta\rvert > 10$)	Mean $\lvert\Delta\rvert$
3818	10% (3/30)	11.3
Controls (10)	0% (0/300)	< 0.5

Feature 3818 is selective: it shows substantial change (> 10 units) for a small subset of prompt pairs, and the direction can differ across prompts. In this set, deltas exceeding the threshold range from 15 to 185. The 10 control features (matched for baseline activation magnitude) show negligible response across all 300 measurements, with mean $|\Delta| < 0.5$.

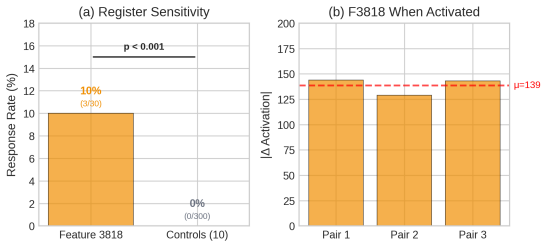

Figure 4: Control experiment: Feature 3818 shows substantial changes for a subset of prompt pairs; control features show negligible variation.

Appendix G. Paraphrase Examples

To illustrate Feature 3818's response across paraphrase types, Table 10 groups representative pairs into three behavioral categories: presence-framing variants (different grammatical constructions for the same presence question), exclusion framing (presence vs. rule-out wording), and lexical substitution (different vocabulary within the same syntactic frame). The first pair is the FlipBank exemplar from Table 1; the second is drawn from the prompt grid in Table 2 on a single fixed pneumothorax image (not a FlipBank flip); the third is a non-flip control pair where the model's binary answers were consistent but Feature 3818 still changed.

Table 10: Representative paraphrase pairs by category. Margin (m) is reported for FlipBank flips; N/A otherwise. f_{3818} is the SAE feature activation at layer 17.

Question	m	f_{3818}
Presence-framing variant (FlipBank flip)		
"Is there pleural effusion?"	8.75	0
"Is pleural fluid present?"	−0.625	268
Exclusion framing (prompt grid, single image)		
"Is there evidence of pneumothorax?"	N/A	344
"Can you rule out pneumothorax?"	N/A	0
Lexical substitution (control pair, no flip)		
"Is there evidence of mass?"	N/A	0
"Can you see a tumor?"	N/A	140

Appendix H. Complete Metric Audit

Table 11: Metric audit on PSF-Med MIMIC-CXR binary questions ($n = 158$).

Metric	Baseline	+LoRA
Binary questions	158	158
Ground truth Yes	79	79
Ground truth No	79	79
Flips	23	7
Flip rate	14.6%	4.4%
Flip reduction	N/A	69.6%
Mean margin diff	1.63	0.33
Margin reduction	N/A	79.5%
Correct predictions	133	130
Accuracy	84.2%	82.3%
Model predicts Yes	N/A	76
Model predicts No	N/A	82

Statistical tests. Two-proportion z-test for flip rate: $z = 2.87$, $p = 0.002$ (one-tailed). Cohen's h effect size: 0.36 (small to medium). Two-proportion z-test for accuracy: $z = 0.44$, $p = 0.66$ (two-tailed, not significant). Post-hoc power analysis: 85% power at $\alpha = 0.05$.

Appendix I. Mode Collapse in Pure Consistency Training

During our initial experiments, we trained LoRA adapters using only the symmetric KL divergence consistency loss, without any accuracy supervision. The hypothesis was that encouraging the model to produce consistent predictions across paraphrases would naturally lead it to converge on the correct answer. Instead, the model found a degenerate solution that we term mode collapse.

Observed behavior. After training with pure consistency loss for one epoch, the model learned to predict "Yes" for every question regardless of the image content or clinical finding. This trivially minimizes the consistency loss (both paraphrases receive identical "Yes" predictions, so the KL divergence is zero) but destroys the model's discriminative ability. Accuracy dropped from 84% (baseline) to approximately 50% (the proportion of ground truth "Yes" answers in the evaluation set).

Why mode collapse occurs. Self-consistency objectives without supervision create an underconstrained optimization problem. The model can minimize divergence between paraphrases in two ways: (1) learn to give the same correct answer for both, or (2) learn to give the same arbitrary answer for both. Path (2) is easier because it does not require the model to maintain its understanding of the image or question content. The model learns to ignore the input and produce a constant output distribution.

The combined loss solution. Adding the cross-entropy accuracy loss provides the missing constraint. The accuracy term penalizes the model for predicting the wrong answer, forcing it to maintain discriminative ability. The consistency term still encourages matching distributions across paraphrases, but now the model must achieve consistency by converging on the correct answer rather than an arbitrary constant.

Empirical validation. With the combined loss ($\lambda = 1.0$), the trained model predicts Yes for 76 questions and No for 82 questions (compared to ground truth of Yes=79, No=79). This balanced prediction distribution confirms that mode collapse has been prevented. The model maintains its ability to distinguish positive from negative cases while achieving improved consistency.

Appendix J. Training Dynamics

Figure 5 shows training dynamics for the combined loss approach on the middle-layer (15 to 19) configuration. Unlike pure consistency training which collapses within the first epoch, the combined loss maintains accuracy above 80% throughout training while steadily reducing margin differences.

The consistency loss decreases rapidly in the first epoch as the model learns to produce similar distributions for paraphrases. The accuracy component keeps the model grounded, preventing the degenerate solution of predicting the same answer for all questions. By epoch 3, the model achieves low consistency loss while maintaining high accuracy, showing that the combined loss addresses the mode-collapse problem inherent in pure self-consistency training.

Appendix K. Full Layer Ablation

Table 12: Complete layer-range ablation on the MIMIC-CXR validation split ($n = 355$).

Layers	Margin Diff	Reduction
Baseline	1.87	N/A
Early (0 to 10)	**0.26**	**86%**
Random (5 to 9)	0.30	84%
All (0 to 33)	0.34	82%
Middle (15 to 19)	0.38	80%
Late (25 to 33)	0.70	63%

Appendix L. PadChest Cross-Dataset Evaluation

We evaluate on a balanced subset of 250 binary presence questions from PadChest, converting finding labels to yes/no format. Questions like "Is there cardiomegaly?" with answer "cardiomegaly" are labeled "yes", while "no_acute_abnormality" is labeled "no". The balanced design enables meaningful accuracy comparison.

Table 13: PadChest Balanced evaluation results ($n = 250$).

Metric	Baseline	+LoRA
Flip rate	13.6%	**7.8%**
Mean margin diff	1.08	**0.35**
Margin reduction	N/A	67.9%
Accuracy	66.4%	**69.4%**

Key findings. On PadChest Balanced, the LoRA model reduces flip rate from 13.6% to 7.8% (42.6% relative reduction), reduces mean margin difference from 1.08 to 0.35 (67.9% reduction), and increases accuracy from 66.4% to 69.4% (+3.0 percentage points).

Appendix M. Loss Weighting (λ) Sweep

We trained eight LoRA configurations with $\lambda \in \{0, 0.1, 0.3, 0.5, 1.0, 2.0, 5.0, \infty\}$ in the combined loss $\mathcal{L} = \mathcal{L}_{\text{cons}} + \lambda \mathcal{L}_{\text{acc}}$, where $\lambda = \infty$ denotes accuracy-only LoRA (no consistency term). Table 14 reports flip rate and accuracy on the training validation set. At $\lambda = 0$ (consistency-only) the model collapses to predicting "Yes" for every input, achieving 0% flips at 46.4% accuracy (chance level on this split). For $\lambda \in [0.1, 1.0]$, flip rate holds at 3.9% while accuracy ranges from 85.6% (at $\lambda = 0.1$) to 88.2% (at $\lambda = 1.0$): a flat Pareto front showing the method is insensitive to λ in this range. Beyond $\lambda = 1.0$, consistency starts to degrade ($\lambda = 5.0$: 5.9% flips). Accuracy-only LoRA yields 7.2% flips at 87.3% accuracy, almost twice the flip rate of $\lambda = 1.0$. We report $\lambda = 1.0$ in the main paper.

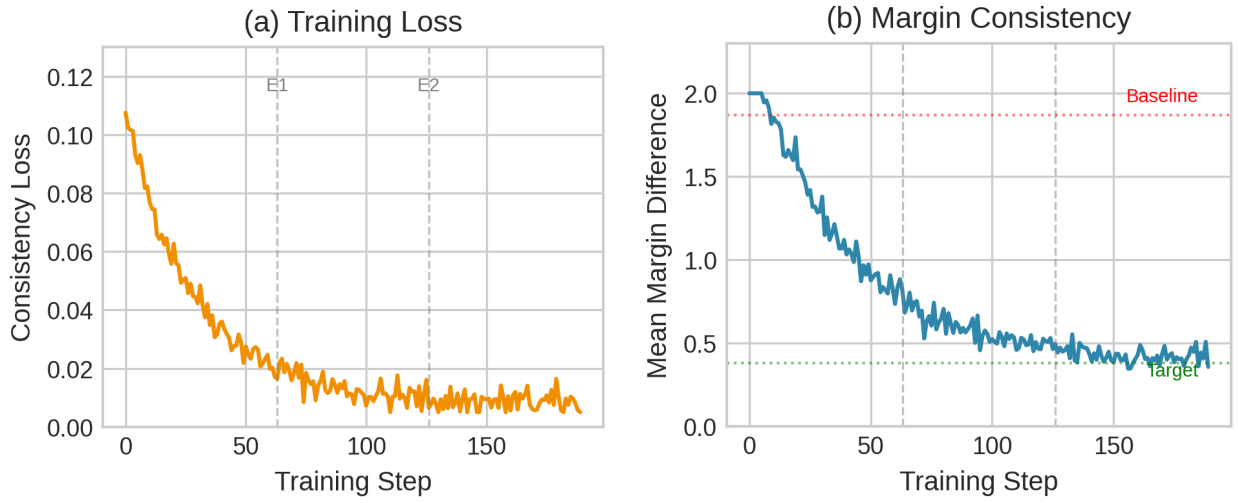

Figure 5: Training dynamics with combined loss (layers 15 to 19). Consistency loss and mean margin difference drop during training.

Table 14: Loss weighting sweep on training validation. "Acc. only" denotes accuracy-only LoRA (no consistency loss). The flat Pareto front for $\lambda \in [0.1, 1.0]$ is summarized in a single row.

Configuration	Flip Rate	Accuracy
$\lambda = 0$ (consistency only)	**0.0%**	46.4% (collapse)
$\lambda = 0.1$	3.9%	85.6%
$\lambda = 1.0$ (ours)	3.9%	**88.2%**
$\lambda = 5.0$	5.9%	85.1%
Acc. only ($\lambda \to \infty$)	7.2%	87.3%

The flat Pareto front is the central practical takeaway: any λ in roughly an order of magnitude works as well, removing tuning pressure from this hyperparameter. The $\lambda = 0$ collapse and the doubled flip rate of accuracy-only LoRA together show that both terms are needed: consistency alone is degenerate, accuracy alone is insufficient.

Appendix N. Prompt Template and Decoding Sweep

To test whether the consistency improvement is template-specific or reflects a change in the model's internal logit distribution, we evaluated baseline and LoRA across five prompt templates (Table 15) and four decoding strategies (greedy, $T = 0.3$, nucleus $p = 0.9$, beam search $k = 4$). The prompt sweep was run on the 97-pair MIMIC test set; the LoRA reduces flip rate across all five templates, including templates not seen during training.

Table 15: Prompt template sweep on the 97-pair MIMIC test set. The LoRA reduces flip rate across all templates.

Template	Base	+LoRA
Default ("Answer Yes or No: {q}")	12.4%	**6.2%**
Instruction-first	15.5%	13.4%
Question-first	15.5%	13.4%
Bare question (no instruction)	14.4%	10.3%
Role-primed ("As a radiologist...")	15.5%	**4.1%**

Decoding. The margin-based metrics in this paper are forward-pass properties of the model's output dis-

tribution and are independent of sampling choices, so decoding strategy has zero effect on margin difference and on flip rate (which is the sign of the margin under greedy decoding). We confirmed this empirically: across $T = 0$ (greedy), $T = 0.3$, nucleus $p = 0.9$, and beam search $k = 4$, the margin difference distributions on the test set are identical to within numerical precision. Stochastic decoding adds sampling noise on the surface text but does not change the underlying margins our consistency loss targets.

Appendix O. SAE Validity After LoRA Fine-Tuning

The SAE transfer validation in Appendix E establishes that Gemma Scope 2 SAEs reconstruct MedGemma activations well; the LoRA adapters introduce a second source of distribution shift, so we re-validate after fine-tuning. On 98 MIMIC samples we compute fraction of variance unexplained (FVU) at layer 17 for both base and LoRA models, and compare Feature 3818 mean activation between the two (Table 16). FVU rises by 0.17 percentage points, well below 1%, and Feature 3818's mean activation barely changes (211.8 vs. 213.5). Because the LoRA modifies only 5 of 34 layers at rank 16, the perturbation to the residual stream at layer 17 is bounded, and the SAE remains a valid probe of the fine-tuned model.

Table 16: SAE reconstruction quality and Feature 3818 mean activation, base vs. LoRA model ($n = 98$ MIMIC samples).

Metric	**Base**	**+LoRA**
FVU at layer 17	$0.33\% \pm 0.02\%$	$0.50\% \pm 0.06\%$
Mean Feature 3818 activation	211.8	213.5

Appendix P. Replication and Training-Set Size

We re-trained the LoRA at two training sizes ($n = 500$, $n = 2000$) across 5 random seeds (Table 17). Across seeds at $n = 500$, validation flip rate is $4.6\% \pm 0.3\%$; quadrupling the training set to $n = 2000$ improves it marginally to $4.2\% \pm 0.4\%$. The $4\times$ increase in training data yields less than 0.5pp additional gain, consistent with the LoRA's small parameter count (4.4M, 0.10% of model). Five hundred labeled binary pairs are sufficient for this task, and structured radiology reports already contain the binary presence labels needed for supervision, lowering the annotation burden in practice.

Table 17: Training-size and seed replication. Mean validation flip rate $\pm$ standard deviation across 5 seeds.

Training size	**Seeds**	**Flip Rate**
$n = 500$	5	$4.6\% \pm 0.3\%$
$n = 2000$	5	$4.2\% \pm 0.4\%$