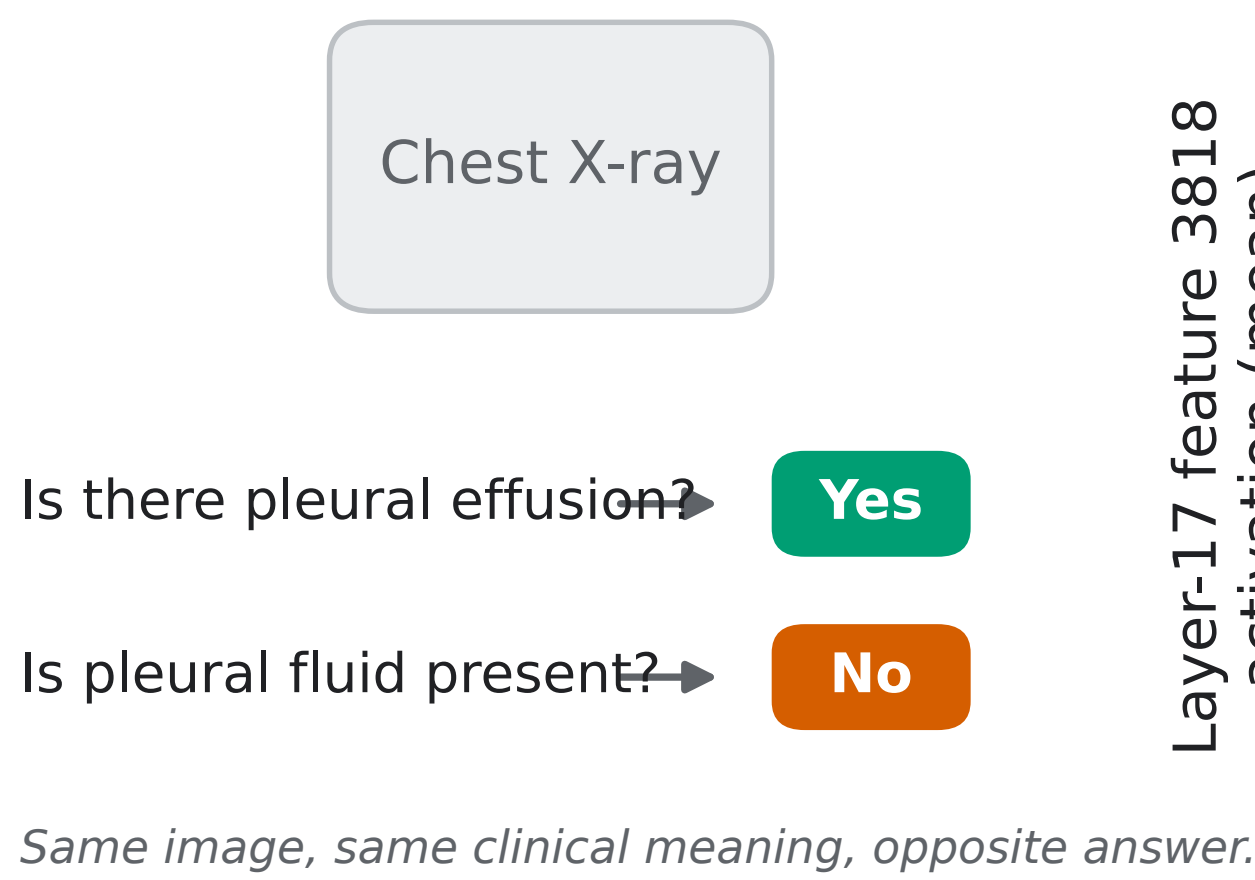


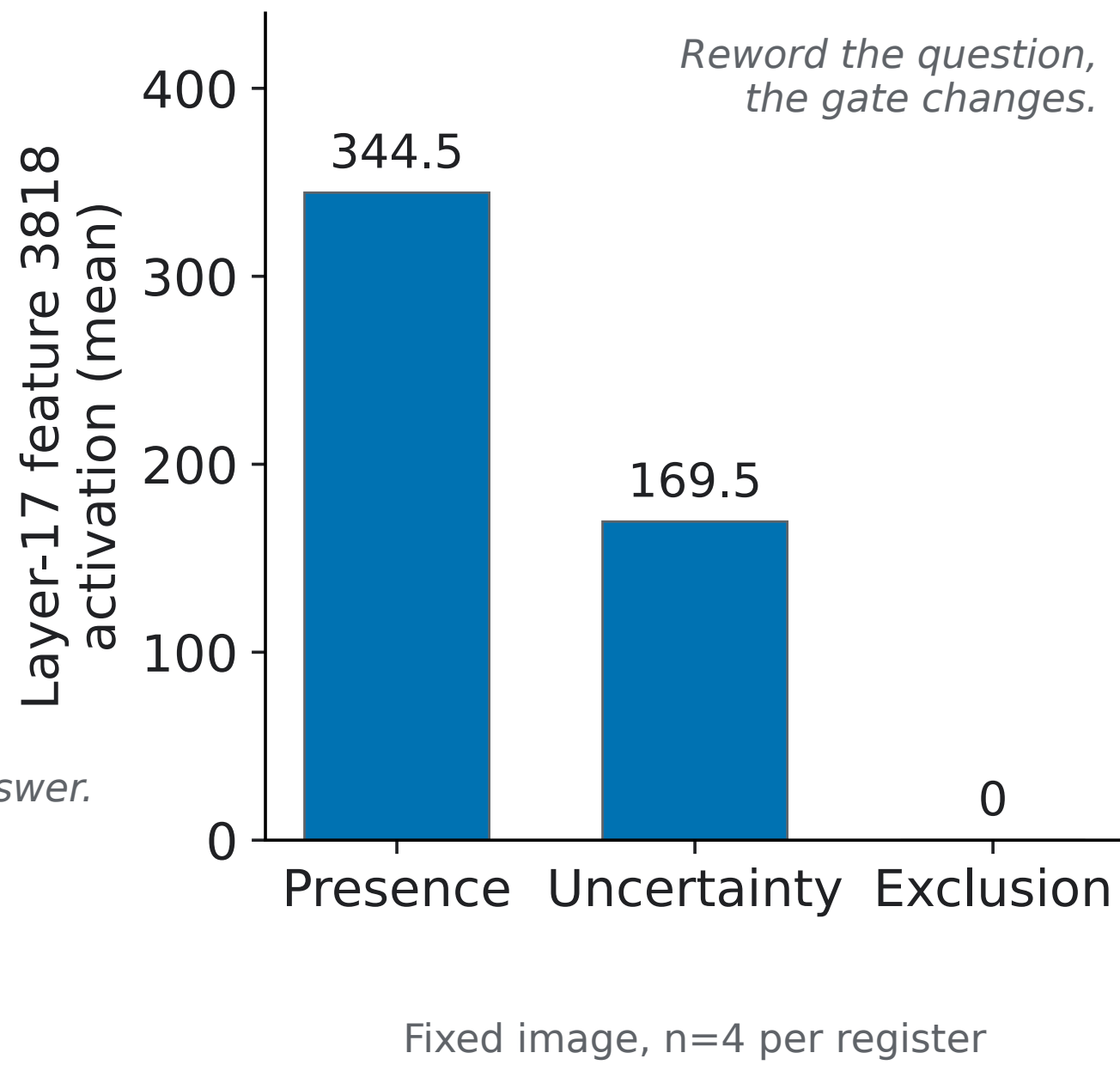
Same scan. Reworded question. Opposite answer.

Binesh Sadanadan Vahid Behzadan
SAIL Lab, University of New Haven · CHIL 2026

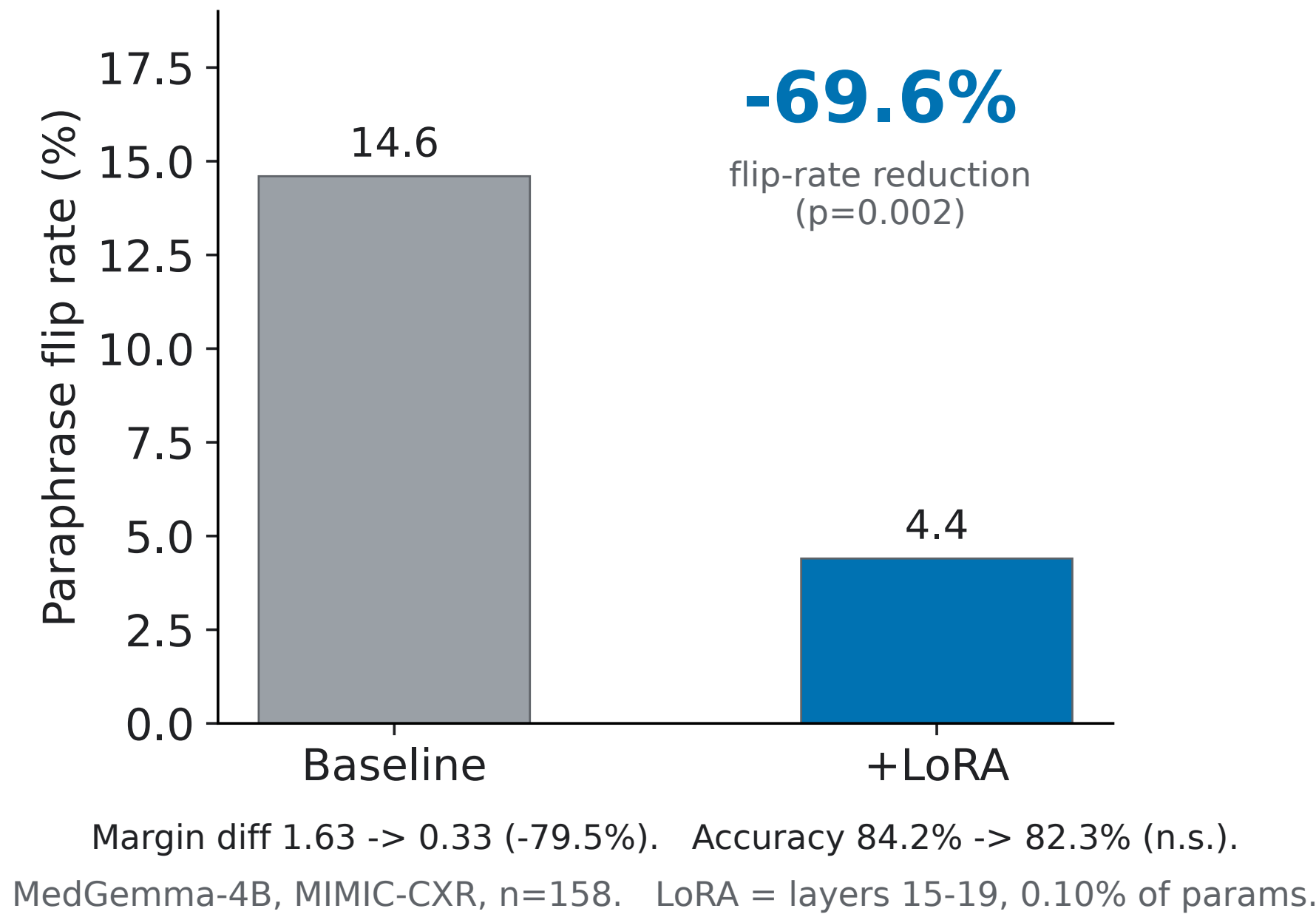
1. The problem



2. The mechanism



3. The fix



A layer-17 feature gates the flip; a **0.10% LoRA** closes it.
On MIMIC-CXR ($n=158$) the paraphrase flip rate falls **14.6% → 4.4%** (-69.6% , $p=0.002$), with accuracy unchanged.

Consistency alone collapses the model

Training only for agreement makes MedGemma answer “Yes” to everything (accuracy 46.4%, chance level). Adding an accuracy term prevents it.
Flip rate and accuracy stay flat for $\lambda \in [0.1, 1.0]$, so the result is not sensitive to how the two objectives are weighted.

Trained on MIMIC, it still helps PadChest

No PadChest data is used in training.

PadChest ($n=250$)	Baseline	+LoRA
Flip rate	13.6%	7.8% (-42.6%)
Accuracy	66.4%	69.4% ($+3.0$ pp)

The mechanism sits at layer 17. The best fix is earlier.

We targeted layers 15 to 19 because feature 3818 fires there. An ablation shows intervening at layers 0 to 10 cuts the gap even more.

LoRA layers	Margin diff	Reduction
Early (0 to 10)	0.26	86%
Middle (15 to 19), targeted	0.38	80%
Late (25 to 33)	0.70	63%

Baseline margin diff 1.87 (MIMIC validation, $n=355$).

The best place to fix a failure can sit upstream of where it appears. We report this, not hide it.

A specific, causal gate, not keyword matching

Patching feature 3818 recovers **28%** of the yes-minus-no logit margin, versus **8%** for control features (Fisher $p = 6.8 \times 10^{-4}$).

It is the top flip driver in 25% of sampled pairs. It drives a substantial share of flips, not all of them.

Takeaway

A **0.10%-parameter**, mechanistically located LoRA makes a medical VLM answer rephrasings consistently without losing accuracy or image reliance. The best intervention point need not match the mechanism’s location.

MedGemma-4B · GemmaScope-2 SAEs · LoRA $r=16$, $\alpha=32$ · combined KL + cross-entropy loss · 500 paraphrase pairs · MIMIC-CXR ($n=158$) and PadChest ($n=250$). No VinDr.



Paper + weights