When First-Token Readouts Misread Feature Steering in a Medical Vision–Language Model

Kevin Jin
Duke University
kevin.jin@duke.edu

Abstract

Interventions on vision–language models are commonly scored on yes/no questions by the model's first output token, often as the probability of "yes" normalized over "yes" and "no", the probability form of the logit difference that interpretability work recommends. We show that this score can miss most of an intervention's effect on the model's written answer. We train a transcoder for MedGemma 1.5 (4B) and use it to suppress the 80 most active features on 40 chest radiographs that the model calls positive for cardiomegaly although they are labelled negative. The normalized score barely moves (0.976 to 0.933), but the share of sampled answers saying yes falls from 0.95 to 0.73. Of the greedy answers the edit turns to "no" at three strengths, the score still says yes on 88%. When we instead cover the part of the image the model relies on, the score still says yes on only 6% of the answers that flip. The generated text shows why. After suppression the model more often opens with "Based on the image, ..." as opposed to "Yes", and that reasoning more often ends in "No", but its first token almost never becomes "No". Occlusion often does, so the score sees that edit. The gap replicates on radiographs selected by their written answer alone and on correct effusion and pneumothorax calls (correct cardiomegaly calls barely respond to the edit). We hypothesize that it comes from how circuit-tracing tools apply edits, which constrain the first token but not the rest of the answer. An edit that looks inert by this score can still change what the model tells a clinician, so such interventions should be scored on sampled answers. We release the transcoder and code.

1 Introduction

A simple way to ask a vision–language model (VLM) about an image is a closed question: "Does this chest X-ray show cardiomegaly? Answer yes or no." The answer is usually scored from the first output token, and interpretability studies score an intervention by how much it changes the probability of "yes". In medical VLMs, examples include activation steering [Farina et al., 2026] and sparse-autoencoder feature patching [Sadanandan and Behzadan, 2026].

The probability of "yes" has a known weakness, as a model can put next-token probability on tokens that are not answers [Wiegreffe et al., 2023]. An instruction-tuned model may open with "Sure" or a refusal, and its first-token probabilities can disagree with the answer it ends up writing [Wang et al., 2024b]. Interpretability work therefore recommends the logit difference between the two answers [Zhang and Nanda, 2024, Heimersheim and Nanda, 2024], or its probability form, normalized over the answers [Rimsky et al., 2024]. Both assume that the model decides at the answer position. A medical model that opens with "Based on the image, ..." decides only after reasoning, and that answer can differ from a direct one [Wang and Zhou, 2024]. A first-token readout cannot see how such reasoning ends. In a clinical tool, the user reads the written answer, so a readout that misses changes to the answer misreports the behavior that matters.

GenAI for Health Workshop @ NeurIPS 2026, Sydney.

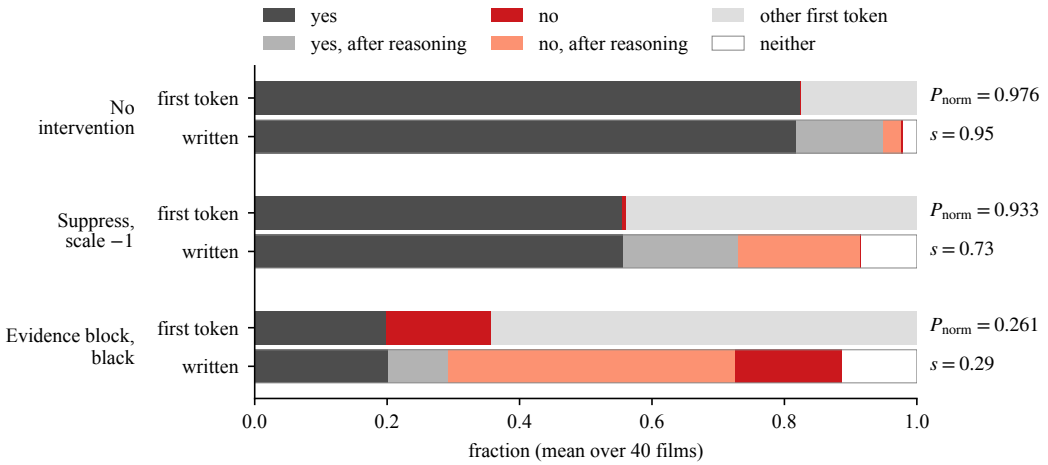


Figure 1: **What the first token says, and what the model then writes.** For each intervention, the upper bar splits the probability of the first output token into "yes" (dark grey, P_{raw}), "no" (red) and every other token (light grey, mostly "Based"); P_{norm} is the "yes" share of "yes" plus "no". The lower bar splits sampled answers by what they conclude and how: dark segments answer directly, light segments reason first, and white segments give neither answer; s is the share that say yes. Suppression moves the first token from "yes" to reasoning, and the reasoning now often concludes "no"; P_{norm} misses both. Blacking out the evidence block, the image region whose occlusion most lowers the first-token readout (§3), does both too, and also moves the first token to "no", so P_{norm} falls with s. Means over 40 cardiomegaly false positives (labelled negative, called positive), 50 samples each; the printed P_{norm} is a mean of per-film values, so it can differ from the ratio of the mean bars. Table 1 gives every intervention.

We test this in MedGemma 1.5. We steer it by suppressing features of a transcoder, a learned dictionary that rewrites the model's internal activations as a sparse set of more interpretable features, and compare three first-token readouts with the answers the model writes. On one radiograph after suppression, the normalized probability of "yes" is 0.982, yet the model writes "*Based on the image, the chest X-ray shows a normal heart size*" and answers no. As a reference, we occlude the image region that the call depends on, an edit the readouts should register. Figure 1 shows both.

We apply the edit as circuit-tracing tools do [Ameisen et al., 2025]. As the model reads the prompt, each feed-forward (MLP) block keeps the output it computed on the unedited input, plus the edit's decoded change, instead of recomputing from its changed input. We call this the *clamp*. The first answer token is predicted at the last prompt position, where the MLP blocks are held fixed by the clamp. Every later token is predicted at a new position, whose MLP blocks respond to the edit. We hypothesize that this mismatch lets the first token miss what the rest of the answer says, and test it by also editing the *live model*, in which every MLP block recomputes from its input. Each test's decision rule was committed before its run; post hoc comparisons are marked †.

Contributions.

1. **The readout misses the edit's effect on the answers.** Across three strengths of suppressing the 80 transcoder features most active on the image, 32 of 120 greedy answers turn to "no", and the normalized readout still says yes on 28 of them†. Across three ways of occluding the image region the call depends on, 53 of 120 turn to "no", and the readout still says yes on only 3†. Pre-registered tests confirm the gap, including on radiographs that no first-token readout selected (§4).
2. **Two routes, and a candidate cause.** Suppression shifts answers from an immediate "Yes" to first reasoning, and makes the reasoning end in "no" more often. The normalized readout registers neither, and the raw probability of "yes" matches the answers only by coincidence† (§4.1). We hypothesize that the clamp is why. Without it, the original strengths mostly stop the model answering; at the only weaker strength that met the pre-registered criteria, the first token moves with the answers, but the test is not conclusive (§4.2).

3. **Scope.** The gap recurs on correct effusion and pneumothorax calls. Correct cardiomegaly calls barely respond to the edit, so the pre-registered test of generalisation is inconclusive. The same selection rule picks almost the same features for every finding, so the edit is not specific to the finding (§5).

These results come from one model, one transcoder, one dataset and tens of radiographs per group. They show that a recommended readout can miss an intervention's effect in a realistic medical setting, and by what route, but not how often this affects published results. We release the transcoder, the code, and a result file behind every number.[1]

2 Setup

Model, data and prompt. We steer MedGemma 1.5-4b-it [Sellergren et al., 2026, 2025] in bfloat16 on a 4,999-image subset of NIH ChestX-ray14 [Wang et al., 2017] (1,335 patients). Each radiograph is presented in a single turn with the question "Does this chest X-ray show *finding*? Answer yes or no." ChestX-ray14 labels are mined from radiology reports and are noisy. We use them only to select radiographs, and every steering comparison compares the model with itself.

First-token readouts. Let ℓ be the logits at the position where the answer would begin and $p = \mathrm{softmax}(\ell)$. For each answer we use the token with the largest logit among the first tokens of its surface forms ("`yes`", "`␣yes`", "`Yes`", "`␣Yes`", "`YES`", and likewise for "no"), and define

$$P_{\text{raw}} = p_{\text{yes}}, \qquad m = p_{\text{yes}} + p_{\text{no}}, \qquad P_{\text{norm}} = \frac{p_{\text{yes}}}{p_{\text{yes}} + p_{\text{no}}} = \sigma(\Delta), \quad \Delta = \ell_{\text{yes}} - \ell_{\text{no}}.$$

P_{raw} is the probability that the next token is "yes", and the *answer mass* m approximates the probability that it is an answer at all, so $P_{\text{raw}} = P_{\text{norm}} \cdot m$. P_{norm} is a sigmoid of the yes-minus-no logit margin Δ: the probability form of the logit difference recommended for activation patching [Zhang and Nanda, 2024, Heimersheim and Nanda, 2024]. On our radiographs P_{norm} starts near 1 (median Δ of 12.2 nats, odds of about 200,000 to 1), so we also report Δ and how many films each readout puts below 0.5.

Written answers. For each radiograph and intervention, we decode greedily and draw 50 samples at temperature 1 (no top-k or nucleus truncation), up to 128 new tokens. A fixed rule removes markdown symbols, lowercases, and takes the first whole word "yes" or "no", and a text with neither gives neither answer. The *sampled yes rate* s is the fraction of samples that say yes. A sample *answers directly* if its first word is "yes" or "no", and *reasons first* otherwise. A blind hand reading agreed with the rule on 94 of 100 random samples. The rule errs toward "neither", and counting such answers as yes or as no changes only one verdict (Appendix C).

The edit. A cross-layer transcoder (CLT) reads the input to each MLP block and reconstructs the outputs of that block and all later ones from a sparse set of features [Lindsey et al., 2024, Ameisen et al., 2025]. Ours covers all 34 layers with 2,048 features per layer and was trained on activations from radiographs and, separately, report texts (Appendix D). On each radiograph we pick the 80 features with the highest mean activation over the image tokens and edit them there. `scale` λ multiplies a feature's activation by λ ($\lambda = -1$ reverses its sign, $\lambda = -2$ reverses and doubles it), and `ablate` sets it to zero. To apply an edit, we run the unedited prompt and record, at every position and layer, the transcoder's features and its reconstruction error (the error term of Ameisen et al., 2025). We then rerun the prompt with each MLP output replaced by the reconstruction from the recorded features, edited at image positions, plus the recorded error. With no edit this reproduces the model exactly. With an edit, each MLP output equals its unedited value plus the decoded change; it is not recomputed from its changed input, while attention is recomputed throughout. This is the clamp. The readouts are taken at the last prompt position, a text position whose MLP outputs keep their unedited values, so the edit reaches them only through attention. Constrained patching [Ameisen et al., 2025] additionally freezes attention over a range of layers, but likewise leaves the MLP blocks at the readout position unable to respond to the edit, the property our hypothesis concerns. The circuit-tracer

[1] Code and results: https://github.com/kevinjin918/first-token-readouts. Transcoder: https://doi.org/10.5281/zenodo.22105454.

library [Hanna et al., 2025] applies its constraints only while predicting the first new token.[2] The first generated token has the readouts' distribution (Appendix D), so the readouts and the text can part ways only after the first token.

Radiographs. The main comparison uses 40 radiographs (*films*) that the model calls positive for cardiomegaly although ChestX-ray14 labels them negative: false positives with respect to the label, which no radiologist checked. All are anteroposterior (AP), a projection that magnifies the heart relative to the standard posteroanterior view, so the heart can look enlarged when it is not. Chest-radiograph classifiers are known to pick up how a film was acquired, such as a portable unit [Zech et al., 2018] and, probably, the projection [DeGrave et al., 2021]. The call then plausibly rests on the heart's apparent size, so we can check where the model's evidence lies; nothing below requires the call to be right or wrong. The films come from 39 patients, in two sets of 20 with no patient in both. The first set, from our pilot study (the version of this work first submitted, which scored first-token readouts only), was selected by $P_{\text{raw}} \geq 0.5$. The second was selected by the greedy answer being yes, so that no readout under test chose it (Appendix C). For correct calls, we drew 20 radiographs each for cardiomegaly, pleural effusion (fluid around the lung) and pneumothorax (air around the lung), labelled positive and answered yes, one film per patient and no patient in two sets.

Pre-registration. Each test we call pre-registered had its decision rule committed in its script before the run. We state outcomes in plain words with the rule's label in small capitals; Appendix A gives each rule, outcome and amendment.

3 A reference effect: occluding the evidence

To know what a readout should register, we need an edit that changes the call by removing its evidence. MedGemma encodes each image as a 16×16 grid of tokens. We occlude one 2×2 block of that grid (one sixty-fourth of the image) at a time and record the change in the first-token readout. The block with the largest effect is the film's *evidence block*.[3]

What do these blocks cover? On the false positives, anatomy whose apparent size depends on how the film was taken. On 20 of the first 20 films the block lies in the central 4×4 blocks, and on 16 in the two columns just right of the image midline, the patient's left, where most of the cardiac silhouette lies (Figure 4a,b). This is consistent with a call that rests on the apparent size of the heart, which the AP projection inflates. On other findings the blocks lie where those findings usually appear: effusion blocks at the lung bases and pneumothorax blocks in the upper zones (Figure 4c). On 3 of the 6 pneumothorax films, all from one patient, the block lies on the curled tip of a chest tube, the treatment-device shortcut documented by Oakden-Rayner et al. [2020].

Occluding the evidence block changes the answers, and the first-token readouts follow (Table 1, lower rows). Occlusion changes only the image, so it runs on the unedited model, with no clamp. Blacking it out turns the greedy answer to "no" on 28 of 40 films and lowers the sampled yes rate from 0.95 to 0.29; P_{norm} says no on 31 films. Three controls tie the effect to what the block covers. An identical black square at a random border block leaves the answers in place, so the effect is not the size of the perturbation. Mean fill, which leaves no sharp edge, still changes the answers, so it is not an edge artifact. And on the 8 films of the first set also called positive for effusion, occluding the cardiac block leaves that call at $P_{\text{norm}} \geq 0.999$, so it is not general damage to the model. By the rule we pre-registered for suppression (Table 2), applied here as a reference†, no first-token readout falls short of the answers when the evidence is removed.

4 Feature suppression changes the written answer

What the first token says. Read through P_{norm}, suppression leaves the model's call nearly unchanged.[4] We report `scale` -1; `ablate` and `scale` -2 agree on P_{norm} throughout (Tables 1 and 2).

[2] The documentation of `feature_intervention_generate` states that the constraints "will only apply to the very first token generated". Later tokens attend to the cached states of the constrained pass.

[3] For the first 20 films, the block that most lowers P_{raw}, as chosen in the pilot study; for the films drawn for this paper, the block that most lowers Δ.

[4] Our pilot study scored first-token readouts only and drew this conclusion. The written answers below show that the call does change.

Table 1: **Suppression and occlusion on the same false-positive radiographs.** Means over 40 films. Δ is the yes-minus-no logit margin; "below 0.5" counts films on which the readout says no. s is the fraction of sampled answers that say yes; "greedy no" counts films whose greedy answer is no and, in parentheses, those on which P_{norm} still says yes.

Intervention	first token				written answer	
	P_{raw}	P_{norm}	median Δ change (nats)	below 0.5 (P_{raw} / P_{norm})	s	greedy no (P_{norm} yes)
None	0.824	0.976	–	5 / 1	0.95	0 (0)
Suppress, `scale` -1	0.555	0.933	-2.9	16 / 2	0.73	11 (9)
Suppress, `ablate`	0.604	0.952	-2.1	14 / 1	0.74	12 (11)
Suppress, `scale` -2	0.523	0.916	-3.1	17 / 3	0.65	9 (8)
Evidence block, black	0.198	0.261	-14.3	31 / 31	0.29	28 (0)
Evidence block, mean fill	0.296	0.469	-9.6	28 / 21	0.44	21 (3)
Evidence block, blur	0.677	0.880	-1.1	12 / 5	0.86	4 (0)
Border block, black	0.822	0.964	$+0.0$	6 / 1	0.94	1 (0)

Table 2: **Does a first-token readout fall by as much as the written answers?** Difference: for each film, the readout's drop minus the drop in the sampled yes rate s; negative when the readout falls by less. Mean over 40 films, 95% film-bootstrap interval in brackets. Pre-registered verdicts: MISREADS if the interval excludes 0 and the mean exceeds 0.05 in size, TRACKS if the interval lies within ± 0.05, UNRESOLVED otherwise. The rule was registered for the suppression rows; the occlusion rows apply it as a reference, and their two MISREADS cells are readouts that fall slightly further than the answers, as expected of a block chosen by a first-token readout.

Intervention	drop in s	P_{raw}		P_{norm}	
		difference	verdict	difference	verdict
Suppression					
`scale` -1	0.22 [0.16, 0.28]	$+0.05$ [0.00, $+0.09$]	UNRESOLVED	-0.18 [-0.24, -0.12]	MISREADS
`ablate`	0.21 [0.14, 0.28]	$+0.01$ [-0.05, $+0.06$]	UNRESOLVED	-0.19 [-0.25, -0.13]	MISREADS
`scale` -2	0.30 [0.22, 0.37]	0.00 [-0.04, $+0.04$]	TRACKS	-0.24 [-0.32, -0.16]	MISREADS
Occlusion (reference)					
evidence, black	0.66 [0.54, 0.76]	-0.03 [-0.09, $+0.03$]	UNRESOLVED	$+0.06$ [$+0.03$, $+0.09$]	MISREADS
evidence, mean	0.51 [0.39, 0.62]	$+0.02$ [-0.05, $+0.10$]	UNRESOLVED	0.00 [-0.05, $+0.05$]	UNRESOLVED
evidence, blur	0.09 [0.03, 0.17]	$+0.06$ [0.00, $+0.11$]	MISREADS	$+0.01$ [-0.02, $+0.04$]	TRACKS
border, black	0.01 [-0.01, 0.02]	0.00 [-0.02, $+0.01$]	TRACKS	$+0.01$ [0.00, $+0.02$]	TRACKS

Suppression lowers P_{raw} from 0.82 to 0.56, and the lost probability goes mostly to "Based" (Figure 1); mean p_{no} stays at most 0.007 under every variant. P_{norm} therefore moves only from 0.976 to 0.933, and says no on 2 of 40 films.

What the model writes. On one film under `scale` -1, P_{norm} is 0.982 and the model writes:

```
Based on the image, the chest X-ray shows a normal heart size.
Therefore, the answer is **No**.
```

Unedited, the same film gets "`Yes, the chest X-ray shows cardiomegaly. The heart appears enlarged.`" Over the three variants and 40 films (120 greedy answers, all "yes" unedited), suppression turns 32 to "no" (on 15 distinct films), and P_{norm} still says yes on 28 of them, on 14 of the 15 films† (Figure 2), each time at $P_{\text{norm}} \geq 0.53$. The three occlusions of the evidence block (Table 1) turn 53 of 120 to "no", and P_{norm} still says yes on only 3†, a comparison that favours the readout, since a first-token readout chose the evidence blocks. Since occlusion runs without the clamp and is a different kind of edit, it shows only that the readout can follow the answers on these films.

The sampled answers show the same gap, and the pre-registered tests confirm it. Under `scale` -1, the sampled yes rate falls from 0.95 to 0.73. Under each variant, P_{norm} falls short of the drop in s by 0.18 to 0.24, a miss by the rule under all three (MISREADS, Table 2). These verdicts also hold with non-answers set aside. An earlier pre-registered check on the first 20 films (10 samples each) had a simpler rule, met under all three variants: suppression turns the greedy answer to "no" on at

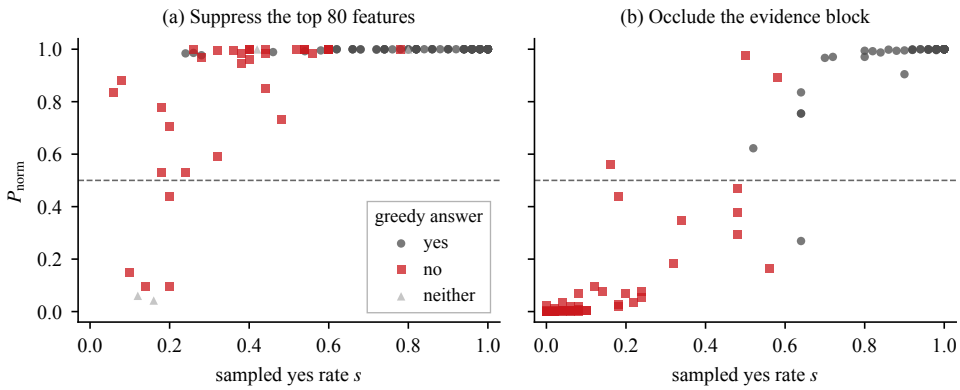


Figure 2: **Film by film, P_{norm} misses the answers that suppression turns to "no".** Each point is one of the 40 false-positive films under one intervention: the sampled yes rate s against P_{norm} after the edit, coloured by the greedy answer. (a) The three suppression variants. (b) The three occlusions of the evidence block (black, mean fill, blur). A red point above the dashed line is an answer the edit turned to "no" on which P_{norm} still says yes: 28 of 32 under suppression, 3 of 53 under occlusion. Unedited, every film's greedy answer is yes.

Table 3: **Sampled answers by their first word,** pooled over 40 films and 50 samples each. Answers that open with "Yes" or "No" give that answer by construction. Suppression moves answers to the reasoning route and changes where that route ends. Black occlusion does both, and also turns direct answers to "no".

	answers directly		reasons first	
Intervention	share	says yes	share	says yes / no / neither
None	0.82	1.00	0.18	0.73 / 0.16 / 0.12
Suppress, `scale` -1	0.56	0.99	0.44	0.39 / 0.42 / 0.19
Suppress, `ablate`	0.61	0.99	0.39	0.33 / 0.59 / 0.08
Suppress, `scale` -2	0.53	0.99	0.47	0.26 / 0.41 / 0.33
Evidence block, black	0.36	0.56	0.64	0.14 / 0.68 / 0.18

least 3 films where P_{norm} still says yes. On the 20 films selected by greedy answer alone, which no first-token readout chose, the rule is met again (REPLICATES).

4.1 Two routes to the answer

Why do the written answers move when P_{norm} barely does? Sorting the sampled answers by their first word shows two routes (Table 3)†. On the direct route, the model opens with "Yes" or "No", and almost always says yes. On the reasoning route, it opens (almost always) with "Based on the image" and concludes later. A sample opens with "Yes" with probability about P_{raw}. If r is the share of samples that reason first and y the share of those that conclude yes, then

$$s \approx P_{\text{raw}} + r\,y.$$

Suppression changes both which route the model takes and where the reasoning route ends. At the first token it moves probability from "Yes" to "Based" ("Based" rises from 0.15 to 0.40), so P_{raw} falls and more samples reason first (r rises from 0.18 to 0.44). The samples that reason also conclude yes less often (y falls from 0.73 to 0.39). P_{norm} sees neither change, because it compares "yes" with "no" at the first token, and "no" gets almost no probability there. Black occlusion sends more samples to the reasoning route ($r = 0.64$) and fewer of them conclude yes ($y = 0.14$), but it also moves the first token to "No" (mean p_{no} 0.159), so P_{norm} falls with the answers. Depending on the variant, P_{norm} falls by less than 0.01 on 30 to 34 of the 40 films, while s on those same films falls by 0.16 to 0.21 on average.

Before the sigmoid, the margin Δ does respond: its median change ranks all seven interventions in the same order as the drop in s (Table 1)†. It starts from a median of 12.2 nats, though, and

rarely crosses zero, because the "yes" probability that suppression removes goes to "Based", which Δ ignores.

P_{raw} falls by about as much as s (0.27 against 0.22), but only by coincidence†. Since $s \approx P_{\text{raw}} + r\,y$, P_{raw} and s fall together only if $r\,y$ barely moves, and it barely moves because the two changes largely offset: more samples reason, but fewer of them conclude yes ($r\,y$ is 0.13 unedited, and 0.17, 0.13 and 0.12 under the three variants). By the pre-registered rule, P_{raw} tracks the answers under `scale` -2 and is unresolved under the other two variants (Table 2). With non-answers set aside, it overstates the effect under `scale` -1 and -2 (by $+0.08$ $[+0.03,\ +0.13]$ and $+0.07$ $[+0.01,\ +0.12]$).

4.2 Is the clamp the cause?

Under the clamp, the MLP blocks where the first token is predicted cannot respond to the edit, while those at every later token can (§1). If this is why the first token misses how the reasoning ends, the first token should move with the answers once every block responds. A stronger edit under the clamp does not close the gap: under `scale` -2, s falls by 0.30, yet P_{norm} falls only to 0.916. We therefore ran two pre-registered tests on the live model, with the same films, feature sets and seeds, adding each edit's change in the transcoder's output to the model's own MLP outputs (Appendix E).

Same strengths. At the strengths of §4, the live edit is far stronger than the clamp and mostly stops the model from answering: 66% to 87% of samples give neither answer. The result then turns on how non-answers are counted. Counted as not yes, P_{norm} falls by less than s under two variants (PARTIAL); with non-answers set aside, it falls by more than the answers do under all three. With most samples not answering, this test cannot decide†.

Weaker strengths. The second test scaled each feature by λ, from 0.9 (mildest) down to 0.15; at $\lambda = 0.6$, each activation changes a fifth as much as under `scale` -1. A strength qualified only if s still fell (interval above 0.05) and at most 17% of samples failed to answer (gave neither answer or said they could not see an image), the most under any intervention of §4. Only $\lambda = 0.6$ qualified. The response is a cliff: s falls by 0.06 $[-0.02,\ 0.15]$ at $\lambda = 0.75$, by 0.75 $[0.65,\ 0.84]$ at $\lambda = 0.6$, and every stronger edit leaves too many samples without an answer. At $\lambda = 0.6$ the first token moves with the answers: it puts 0.26 on "no" (at most 0.007 under the clamp), and P_{norm} and s fall by about the same amount (difference -0.005 $[-0.070,\ +0.056]$).

This is what the hypothesis predicts, but it does not show that the clamp is the cause, for three reasons. (i) Dose: at $\lambda = 0.6$ the answers move 3.4 times as much as under the clamp, and no milder live edit moved them enough to qualify, so the two are never compared at the same effect size. (ii) Precision: the interval of that difference is wider than the ± 0.05 the rule needs to call agreement (UNRESOLVED). (iii) Confound: the live edit frees the MLP blocks at the image positions as well as at the answer position. Freeing only the answer position's blocks would isolate the hypothesis, and is the natural next test.

5 Scope

Correct calls and other findings. We repeated `scale` -1 and the evidence occlusion on 20 correctly called positive films for each of three findings (Table 4). On effusion and pneumothorax, suppression lowers s by 0.17 and 0.21, and P_{norm} misses the effect by the pre-registered rule (MISREADS), as on the false positives. Correct cardiomegaly calls respond only slightly, as s falls by 0.04. The pre-registered test of generalisation asks P_{norm} to miss on every finding whose answers the edit moves at all, and fails if it tracks on any. This small drop counts, but on it P_{norm}'s shortfall is too small to call a miss and too uncertain to call agreement, so the verdict is INCONCLUSIVE. With non-answers set aside, P_{norm} tracks the answers on cardiomegaly, and the verdict becomes DOES NOT GENERALISE. We read this† as an edit too weak on these calls for any readout to miss: blacking out the evidence block moves them (s falls by 0.28), but this suppression barely does.

What the edit removes. The suppressed features are not specific to the finding. If the 80 most active features were cardiomegaly evidence, the same rule would pick different features for other findings. We built top-activation sets for cardiomegaly, effusion and pneumothorax, each from 8 *donor* radiographs used only to choose features (one per patient, with no patient shared across sets or with the first 20 films). The three sets nearly coincide: their mean pairwise Jaccard index is 0.944,

Table 4: **Correctly called radiographs, three findings.** 20 per finding, labelled positive and answered yes. Difference: the readout's drop minus the drop in s, as in Table 2; mean with 95% film-bootstrap interval. Pre-registered verdicts for P_{norm}: UNRESOLVED for cardiomegaly, MISREADS for effusion and MISREADS for pneumothorax; for P_{raw}, UNRESOLVED on all three.

	no intervention			suppression, `scale` -1			evidence black
Finding	P_{raw}	m	s	drop in s	difference, P_{raw}	difference, P_{norm}	drop in s
Cardiomegaly	0.97	0.97	0.99	0.04 [0.02, 0.07]	+0.04 [+0.02, +0.06]	−0.04 [−0.07, −0.01]	0.28 [0.11, 0.48]
Effusion	0.92	0.92	0.96	0.17 [0.07, 0.28]	+0.02 [−0.01, +0.05]	−0.15 [−0.26, −0.05]	0.22 [0.07, 0.37]
Pneumothorax	0.91	0.93	0.96	0.21 [0.11, 0.32]	+0.02 [−0.01, +0.05]	−0.16 [−0.24, −0.08]	0.29 [0.13, 0.47]

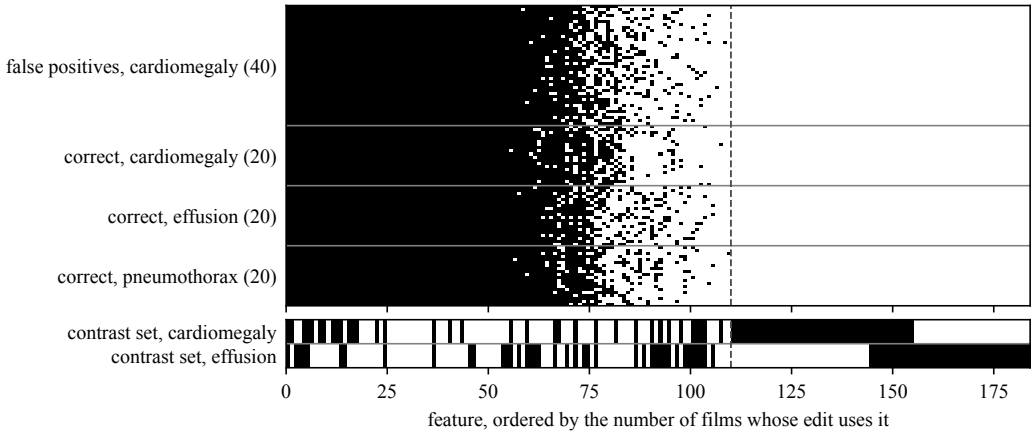


Figure 3: **The edit suppresses nearly the same features on every film.** Each row of the upper panel is one film, and a dark cell is one of the 80 features suppressed on it (its 80 most active). Over 100 films in four groups, the edits draw on 110 features, 55 of them suppressed on every film, and any two films share 67 to 79 of their 80. The lower panel shows the contrast sets of Table 5, which differ between findings and take 40 to 45 of their 80 features from outside the films' sets (right of the dashed line).

against 0.284 for sets chosen by contrast between positive and negative donors, and the gap holds for sets of 20 and 320 features (Appendix F). Top activation picks what is active on any chest radiograph (Figure 3).

Contrast sets built from larger groups (16 positive against 16 negative donors) are finding-specific (Jaccard 0.231 between the cardiomegaly and effusion sets), but suppressing them barely changes the answers. On correct calls of its own finding, neither set beats 80 random active features by the pre-registered margin, and neither does the cardiomegaly set on the false positives (CONTRASTIVE SETS INERT, CONTRAST INERT ON FP; Table 5). Nearly all their features are active on the false positives, but we did not record how strongly they fire, so weak firing could explain the small effect. Either way, the readout failure of §4 does not depend on the edit being specific.

6 Related work

First-token readouts of interventions. Farina et al. [2026] steer the live MedGemma and score it from the next-token "Yes" and "No" logits without parsing the text, noting that steering changes their combined mass. Sadanandan and Behzadan [2026] score sparse-autoencoder patching of MedGemma by the yes-minus-no margin.

The choice of metric. Zhang and Nanda [2024] show that probability and logit difference can lead to different conclusions in activation patching, and recommend the logit difference. Heimersheim and Nanda [2024] note that it is insensitive to components that move all compared logits together, the blind spot we measure. Steering work on multiple-choice questions uses the probability normalized over the options (in the released code of Rimsky et al., 2024) or the logit difference, as in Tan

et al. [2024], who argue that the model decides at the answer token; ours is a case where it does not. Steering vectors built from multiple-choice questions can fail on open-ended answers [Pres et al., 2024], a first-token proxy for refusal is less accurate than full completions [Arditi et al., 2024], and a circuit-tracing intervention can reverse when attention is allowed to respond [Lindsey et al., 2025]. Concurrently, Wang and Xu [2026] find, in general-domain VLMs prompted to reason first, that a direction patch that barely changes the first-position choice still lowers the accuracy of the generated answers by 13 to 16 points. We add a medical case under a direct yes/no prompt, where readout and answer disagree film by film; we show the routes by which they part, and identify a candidate condition: a first token computed under constraints that do not apply to the rest of the answer.

First tokens and generated text. Instruction-tuned models shown the options place nearly all first-token probability on them [Wiegreffe et al., 2023]; MedGemma, told to answer yes or no, often does not. Under image swaps in MedGemma-27B, first-token and generated answers agree on 91 to 99% of cases [Cajas Ordóñez et al., 2026], while generated answers are the more robust of the two to question perturbations [Wang et al., 2024a].

7 Discussion

Recommendation. When an intervention is evaluated on a closed question, score sampled answers, and report the first-token readouts and the shares of answers that reason first or give no answer alongside them. Scoring the answers would have caught both failures we saw: under the clamp, the first token missed how the reasoning ended, and under strong live edits, the readout reported a decision while the model mostly stopped answering. The answer mass m alone is not enough. A fall in m shows that the first token no longer carries most answers, but not which way they moved: it fell further under black occlusion (0.36), where P_{norm} followed the answers, than under suppression (0.56). A greedy answer alone can hide a shift in the distribution [Pres et al., 2024]; here, 50 samples per film sufficed. Finally, our clamp, like circuit-tracer's generation function, constrains only the pass that predicts the first token. Whether or not that caused the gap here (§4.2), the first token and the rest of the answer should come from one configuration, or the difference should be stated.

When first-token readouts suffice. When the model still answers at the first token under the intervention (m near 1 after the edit), the logit difference describes the decision and remains the right tool for the patching studies it was recommended for. In our setting, too, the margin ranked the seven interventions as the answers did†, though it could not tell what the model answered. We do not judge whether the model's cardiomegaly calls are clinically right, or treat its written reasoning as an explanation of its computation [Lanham et al., 2023].

Limitations. We steer one model, with one transcoder, on one dataset. The steering comparisons use 40 false-positive and 60 true-positive radiographs, and the top-activation sets of §5 use 8 donors each. No radiologist reviewed the radiographs, labels, occlusion maps or generated text, and the anatomical descriptions are our own. The parse rule reads the first "yes" or "no" within 128 tokens, and a model could revise its answer later; the author did the hand reading, blind but not independent. Per-film effects depend on near-ties at the top-80 cutoff (Appendix D), so we report means and pooled counts. We sample at temperature 1 without truncation, unlike a deployed model; the greedy answers tell the same story.

Acknowledgments and Disclosure of Funding

Compute for this work was provided by the YC Summer Fellowship. The author declares no competing interests.

References

Emmanuel Ameisen, Jack Lindsey, Adam Pearce, Wes Gurnee, Nicholas L. Turner, Brian Chen, Craig Citro, et al. Circuit tracing: Revealing computational graphs in Language Models. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/attribution-graphs/methods.html.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, volume 37, pages 136037–136083, 2024. doi: 10.52202/079017-4322.

Sebastián Andrés Cajas Ordóñez, Maximin Lange, Quang Bui, Anqi Peter Li, Felipe Ocampo Osorio, Rafi Al Attrach, Kushul Reddy Palakala, Sahil Kapadia, Zakaria Laouabdia Sellami, Xinyue Zhang, Ashley Zhang, and Leo Anthony Celi. ModaLens: Measuring image sensitivity in report-conditioned medical VLMs. *arXiv preprint arXiv:2609.15635*, 2026.

Pierre Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven QH Truong, Chu The Chuong, and Curtis P. Langlotz. CheXpert Plus: Augmenting a large chest x-ray dataset with text radiology reports, patient demographics and additional image formats. *arXiv preprint arXiv:2405.19538*, 2024. URL https://arxiv.org/abs/2405.19538.

Alex J. DeGrave, Joseph D. Janizek, and Su-In Lee. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619, 2021. doi: 10.1038/s42256-021-00338-7.

Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. Transcoders find interpretable LLM feature circuits. In *Advances in Neural Information Processing Systems*, volume 37, pages 24375–24410, 2024. doi: 10.52202/079017-0768.

Eduardo Moreno Judice de Mattos Farina, Mateus A. Esmeraldo, Felipe Akio Matsuoka, Paulo Eduardo de Aguiar Kuriki, and Felipe Campos Kitamura. Translating inference-time control to radiology vision-language models: Activation steering for pneumonia classification on chest X-rays, 2026. URL https://arxiv.org/abs/2606.20852.

Michael Hanna, Mateusz Piotrowski, Jack Lindsey, and Emmanuel Ameisen. Circuit-tracer: A new library for finding feature circuits. In *Proceedings of the 8th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 239–249, 2025. doi: 10.18653/v1/2025.blackboxnlp-1.14.

Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*, 2024.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, et al. Measuring faithfulness in chain-of-thought reasoning, 2023. URL https://arxiv.org/abs/2307.13702.

Jack Lindsey, Adly Templeton, Jonathan Marcus, Thomas Conerly, Joshua Batson, and Christopher Olah. Sparse crosscoders for cross-layer features and model diffing. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/crosscoders/index.html.

Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, et al. On the biology of a large language model. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/attribution-graphs/biology.html.

Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM Conference on Health, Inference, and Learning (CHIL)*, pages 151–159, 2020. doi: 10.1145/3368555.3384468.

Itamar Pres, Laura Ruis, Ekdeep Singh Lubana, and David Krueger. Towards reliable evaluation of behavior steering interventions in LLMs. In *NeurIPS 2024 Workshop on Foundation Model Interventions (MINT)*, 2024. URL https://arxiv.org/abs/2410.17245. arXiv:2410.17245.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.828.

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

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, et al. MedGemma technical report. *arXiv preprint arXiv:2507.05201*, 2025. URL https://arxiv.org/abs/2507.05201.

Andrew Sellergren, Chufan Gao, et al. MedGemma 1.5 technical report. *arXiv preprint arXiv:2604.05081*, 2026. URL https://arxiv.org/abs/2604.05081.

Daniel Tan, David Chanin, Aengus Lynch, Brooks Paige, Dimitrios Kanoulas, Adrià Garriga-Alonso, and Robert Kirk. Analysing the generalisation and reliability of steering vectors. In *Advances in Neural Information Processing Systems*, volume 37, pages 139179–139212, 2024. doi: 10.52202/079017-4417.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. ChestX-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Xinpeng Wang, Chengzhi Hu, Bolei Ma, Paul Röttger, and Barbara Plank. Look at the text: Instruction-tuned language models are more robust multiple choice selectors than you think. In *First Conference on Language Modeling (COLM)*, 2024a. URL https://arxiv.org/abs/2404.08382. arXiv:2404.08382.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. "My answer is C": First-token probabilities do not match text answers in instruction-tuned language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7407–7416. Association for Computational Linguistics, 2024b. doi: 10.18653/v1/2024.findings-acl.441.

Xuezhi Wang and Denny Zhou. Chain-of-thought reasoning without prompting. In *Advances in Neural Information Processing Systems*, volume 37, pages 66383–66409, 2024. doi: 10.52202/079017-2123.

Zeyu Wang and Xinming Xu. Knowing isn't always saying: When do spatial encodings reach answers in vision-language models? *arXiv preprint arXiv:2608.22916*, 2026.

Sarah Wiegreffe, Matthew Finlayson, Oyvind Tafjord, Peter Clark, and Ashish Sabharwal. Increasing probability mass on answer choices does not always improve accuracy. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8392–8417. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.522.

John R. Zech, Marcus A. Badgeley, Manway Liu, Anthony B. Costa, Joseph J. Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS Medicine*, 15(11):e1002683, 2018. doi: 10.1371/journal.pmed.1002683.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations (ICLR)*, pages 1651–1678, 2024.

A Pre-registered decision rules

Each rule was committed in its experiment's script before the run, and the commit hashes are in the result files. Verdicts are stated in the rule's own terms.

- **Generation check** (first check, 20 films, 10 samples). Per suppression variant: DISAGREES if the greedy answer is "no" on at least 3 films where P_{norm} calls yes; otherwise HEDGE-DOMINATED if it gives neither answer on at least 5; otherwise VALIDATED if it matches the P_{norm} call on at least 18; otherwise MIXED. Result: DISAGREES under all three variants (4, 4 and 3 films under `scale` -1, `ablate` and `scale` -2). The script also required the unedited model and `scale` -1 to reproduce the pilot study's P_{raw} within 0.02 on every film. It was amended twice, without changing the films, arms, seed or verdict rules. When `scale` -1 failed on two of the first five films (a feature swap at the top-80 boundary, Appendix D), the requirement was kept for the unedited model only, before any generated text was read. When the unedited model's P_{raw} then drifted by 0.032 on the seventh film, the requirement was dropped and the run restarted. This second amendment was not blind: the greedy answers for the first six films had been seen.
- **Replication with sampled answers** (40 false-positive films, 50 samples). Per variant, the per-film difference between a readout's change and the change in s is bootstrapped (10,000 resamples). MISREADS if the 95% interval excludes 0 and the mean exceeds 0.05 in size; TRACKS if the interval lies within ± 0.05; otherwise UNRESOLVED. P_{norm} is SUPPORTED as missing the effect if all three variants misread, REFUTED if all three track, and PARTIAL otherwise; likewise P_{raw}. Fresh films: the first check's rule, REPLICATES if at least 2 of 3 variants disagree. True positives, under `scale` -1: a finding shows an effect if the interval of its drop in s excludes 0; GENERALISES if at least two findings show an effect and P_{norm} misreads on all of them, DOES NOT GENERALISE if P_{norm} tracks on any of them, otherwise INCONCLUSIVE. Every verdict is repeated with s over answered samples only. Results: P_{norm} SUPPORTED and P_{raw} PARTIAL as missing the effect (answered only: SUPPORTED, PARTIAL); fresh films REPLICATES; true positives INCONCLUSIVE (answered only: DOES NOT GENERALISE).
- **Contrastive selection** (40 false positives, and 20 correct calls each for cardiomegaly and effusion; 80 features per set, `scale` -1, 10,000-resample film bootstrap). SELECTION RESCUES SPECIFICITY if, on correct cardiomegaly calls, the cardiomegaly contrast set lowers s by more than the effusion set (mean above 0.10, interval above 0) and the converse holds on correct effusion calls; CONTRASTIVE SETS INERT if neither contrast set beats random active features by 0.10 (upper bound) on its own finding; otherwise MIXED. On the false positives, the cardiomegaly contrast set is compared with random active features, and evidence-block with border-block anchoring, at the same threshold. Results: specificity CONTRASTIVE SETS INERT; on the false positives CONTRAST INERT ON FP, and UNRESOLVED for anchoring.
- **Live MLPs** (the replication's films, feature sets, seeds and samples). Each suppression's change to the MLP outputs, the reconstruction from the edited features minus that from the unedited ones, is added to the live MLP outputs. INCONCLUSIVE if under no variant the interval of the drop in s excludes 0; HOLDS WITH LIVE MLPs if P_{norm} misreads with a negative mean under all three variants; FROZEN-MLP ARTIFACT if it tracks under all three; otherwise PARTIAL. Result: PARTIAL (MISREADS, MISREADS and UNRESOLVED under `scale` -1, `ablate` and `scale` -2).
- **Live MLPs, weaker suppression.** As above, with each film's features scaled by $\lambda \in \{0.9, 0.75, 0.6, 0.45, 0.3, 0.15\}$. A strength is eligible if at most 17% of samples give neither answer or say they cannot see an image (the largest share under any arm of the replication), and the interval of its drop in s lies above 0.05. At the eligible strength whose drop in s is closest to that of clamped `scale` -1, by the replication's rule: HOLDS LIVE if P_{norm} misreads with a negative mean, REVERSES LIVE if with a positive mean, TRACKS LIVE if it tracks, otherwise UNRESOLVED; NO MATCHED DOSE if no strength is eligible. Result: UNRESOLVED.
- **Patient-disjoint donor sets.** Mean pairwise Jaccard at $N{=}80$ above 0.70: STILL FINDING-AGNOSTIC; below 0.50: PATIENT-DRIVEN. Result: STILL FINDING-AGNOSTIC (0.944). Amplification, on 12 cardiomegaly-negative films with $0.05 < P_{\text{raw}} < 0.5$, each set ampli-

fied ten times at image positions and scored by first-token readouts only: NO EFFECT if the cardiomegaly set raises the probability of cardiomegaly by at most 0.10 more than random features. Result: NO EFFECT (0.045 above random on P_{norm}, 0.093 on P_{raw}).

- **Pilot-study tests,** scored by first-token readouts only. Readout audit: ARTIFACT (normalized effect under half the raw one), the pilot study's conclusion, which §4 revisits. Occlusion audit: SURVIVES. Operator sweep: no operator reaches selectivity $\geq 2\times$. Evaluation sweep: AUCs under P_{raw} and P_{norm} within 0.03. Prompt sweep: no phrasing shifts P_{raw} by more than 0.15 while shifting P_{norm} by less than 0.05.

A separate pre-registered test, of how the readout affects an evaluation with no intervention (MedGemma and CheXagent-2), asks a different question; its rules and results are in the code repository.

B Where the evidence sits

Figure 4 shows the evidence blocks described in §3.

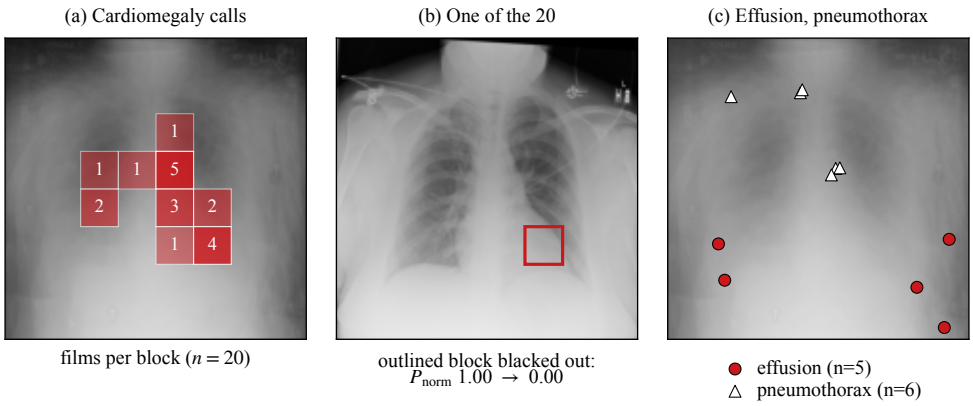


Figure 4: **Where the evidence for each call sits. (a)** For each of the first 20 false-positive radiographs, the 2×2-token block whose black occlusion most lowers P_{raw}, counted on the 8×8 block grid and drawn over the mean of the 20 images. All 20 fall over the cardiac silhouette and the mediastinum. **(b)** One of the 20: blacking out the outlined block, at the left heart border, takes P_{norm} from 1.00 to 0.00. **(c)** The most damaging block on a separate set of 20 effusion- and 20 pneumothorax-positive radiographs from the pilot study, which may include several films per patient, for films where one block moves P_{norm} by more than 0.15. Effusion blocks lie at the lung bases; pneumothorax blocks lie in the upper zones or, on three films from one patient, over a chest tube.

C Radiographs and parsing

Radiograph sets. The first false-positive set comes from the pilot study, was selected by $P_{\text{raw}} \geq 0.5$, and contains two films of one patient. The second was drawn for this paper with one film per patient. Eligible AP films labelled negative for cardiomegaly were visited in a seeded random order and accepted when the greedy answer was yes (20 accepted of 99 eligible). No patient is in both sets.

Parse rule and hand reading. The rule strips the characters `* _ # ‘ >`, lowercases, and returns the first whole word "yes" or "no". The hand-reading protocol was committed before the sheet was drawn. The sheet mixed 100 samples drawn uniformly from the replication with the 45 distinct greedy texts of the first check. The reader, the author, saw only the question and the reply, and wrote yes, no or neither. The rule agreed on 94 of 100 samples and 43 of 45 greedy texts. Of the 8 disagreements, 5 answer without the word yes or no ("`it does not appear to show cardiomegaly`"), which the rule calls neither; one calls the heart enlarged and then says "`The answer is **No**`"; and two say yes before reaching the length limit, which the sheet's instructions counted as neither. Each is quoted in the result file.

Sensitivity (not pre-registered). Counting every answer that gives neither as yes, and then as no, P_{norm} missing the effect stays SUPPORTED and SUPPORTED, P_{raw} stays PARTIAL and PARTIAL, and the fresh-film replication stays REPLICATES and REPLICATES. The true-positive verdict, INCONCLUSIVE as parsed, becomes DOES NOT GENERALISE when neither counts as yes and stays INCONCLUSIVE when it counts as no. Under three other parse rules (the last yes or no; the first after deleting mentions of the format such as "yes or no"; and that rule counting replies that cannot see an image or hit the length limit as neither), P_{norm} misreads every suppression variant, falling short of s by 0.19 to 0.28. Degenerate text is rare except under `scale` -2, where 20.5% of samples reach the length limit (3.6% under `scale` -1, 0.1% unedited). At most 1.9% of samples in any variant say they cannot see an image.

D The transcoder and reproducibility

The transcoder. Ours, the cross-layer form [Lindsey et al., 2024] of a transcoder [Dunefsky et al., 2024] used for attribution graphs [Ameisen et al., 2025, Lindsey et al., 2025], has 3.30B trained parameters, $2{,}048$ features per layer across all 34 layers, and Top-K sparsity with $K{=}32$. The checkpoint also stores 2.94B decoder weights that would write past the last layer and are never used. It was trained on activations captured while MedGemma processed ChestX-ray14 radiographs (prompt "Interpret this chest X-ray.") and, as separate inputs, CheXpert Plus report texts [Chambon et al., 2024]. Of the 80 features edited on each false-positive film, a mean of 21 belong to the last layer; at image positions no later layer reads them, so editing them cannot change the answer.

Fidelity. Substituting the reconstruction for every MLP, with no error term, reproduces the model's top-1 next token on 100% of 60 radiographs under an open-ended prompt (KL 0.044). Under our yes/no prompt it does not (0 of 10, KL 13.5), so every intervention adds back the error term recorded on the unedited input. On text positions the dictionary is prompt-bound (fraction of variance unexplained 0.049 on the training prompt, 1.404 on held-out yes/no prompts). Image-token activations precede the text under causal masking and are unaffected (0.355 on every prompt), and every edit here is at image positions. Checkpoint and configuration: `https://doi.org/10.5281/zenodo.22105454` (CC BY 4.0; `clt_ckpt.pt`, 37,630,140,293 bytes, MD5 `4c2747c75e4ce368bd48ccd194f9b8ef`).

Reproducing the suppression. The first check reran the pilot study's suppression on its 20 films with newer software (PyTorch 2.14.1, Transformers 5.14.1). The unedited model's P_{raw} stayed within 0.032 of the pilot study's; the suppressed arms differed by up to 0.108, 0.077 and 0.107, all on the film where the unedited model drifted most. On two films the selected feature set changed at the top-80 boundary. On one, the 80th and 81st mean activations are 1.0880 and 1.0812; on the other, replacing any one feature of ranks 45 to 77 with rank 81 moves P_{raw} anywhere from 0.591 to 0.814. The pilot study did not save its feature sets; every later run does. The replication reuses the first check's sets on those 20 films and reproduces its first-token readouts to within 0.000. As a check on the generation code, the first generated token's distribution must match the scored forward pass to within 0.02 in P_{raw} and P_{norm} on every radiograph and arm; the largest difference was 0.000.

Compute. One H100 (80 GB). Per their logs in the repository, the generation runs took 8 min (first check), 47 min (replication) and 48 min (contrastive selection), some sharing the GPU.

E Applying the edit to the live model

At the strengths of §4, the live edit mostly stops the model from answering. Between 66% and 87% of samples give neither answer (3% to 15% under the clamp), up to 85% run to the 128-token limit, and the answer mass is 0.03–0.04. In the weaker test, every λ below 0.6 leaves 62% to 83% of samples without an answer, so $\lambda = 0.6$ is the one eligible strength (16% of samples fail to answer). At that strength, the first token's answer mass is 0.36, P_{norm} falls by 0.75 $[0.62, 0.86]$, and the answers that open with "Based" conclude yes less often (0.16, against 0.41 under clamped `scale` -1).

F Feature selection

All donors here are patient-disjoint (§5). With 20, 80 and 320 features per set, the mean pairwise Jaccard overlap between the cardiomegaly, effusion and pneumothorax top-activation sets is 0.770, 0.944 and 0.939, against 0.194, 0.284 and 0.331 for contrast sets from the same donors. On correct calls, the larger contrast sets of Table 5 lower s by 0.00 [0.00, 0.01] (cardiomegaly set, cardiomegaly calls) and 0.02 [−0.01, 0.06] (effusion set, effusion calls); on the false positives, 80 random active features lower it by 0.04 [0.01, 0.07]. Anchoring the selection at the evidence block rather than a border block leans toward larger drops: +0.08 [+0.01, +0.16] on the false positives (UNRESOLVED, short of the pre-registered 0.10) and +0.27 [+0.11, +0.43] on correct effusion calls (no verdict).

Table 5: **Suppressing 80 features chosen in different ways** (`scale −1`). Drop in s, mean with 95% film-bootstrap interval, from a separate run with 30 samples per film (so the top-80 row differs slightly from Table 2). Contrast sets: the 80 features whose mean activation differs most between 16 positive and 16 negative donors. Anchored sets: the 80 most active features over the film's evidence block or a border block. "Active": how many of the 80 are active on the film; "in top 80": overlap with the film's own set (both on the false positives).

			drop in s		
				correct calls	
Feature set	active	in top 80	false positives	cardiomegaly	effusion
Own top 80 (the paper's)	80	80	+0.23 [+0.16, +0.30]	+0.04 [+0.02, +0.07]	+0.17 [+0.06, +0.30]
Cardiomegaly contrast	80	25	+0.07 [+0.03, +0.11]	+0.00 [+0.00, +0.01]	+0.03 [+0.01, +0.06]
Effusion contrast	80	25	+0.04 [+0.02, +0.06]	+0.00 [−0.01, +0.01]	+0.02 [−0.01, +0.06]
Random active 80	80	4	+0.04 [+0.01, +0.07]	−0.00 [−0.01, +0.00]	+0.03 [−0.00, +0.08]
Anchored, evidence block	80	68	+0.27 [+0.19, +0.36]	+0.09 [+0.02, +0.19]	+0.37 [+0.21, +0.53]
Anchored, border block	80	68	+0.19 [+0.13, +0.26]	+0.02 [+0.01, +0.04]	+0.10 [+0.04, +0.19]