# MODALENS: MEASURING IMAGE SENSITIVITY IN REPORT-CONDITIONED MEDICAL VLMS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

A radiology report can already answer a clinical question, so it is hard to tell whether a vision-language model also uses the image. MODALENS, a paired image-swap audit, measures how report availability changes image sensitivity: MedGemma-27B on 3,199 paired MIMIC-CXR cases from 293 patients, the 13 finding-specific questions per case, each image replaced by one from another study, usually of the same patient, with question and report fixed. Under an explicit answer instruction, the model's generated answer changes on 4.43% of trials with the report and 19.93% without it, a paired increase of 15.5 points (patient-clustered 95% CI [14.5, 16.5]), so report availability reduces image-swap sensitivity under this protocol, and substitutions also move continuous answer scores where the binary prediction does not change. The direction replicates in two further general-domain lineages and on a second institution's chest X-ray collection; two medical models read the substituted image at chance with or without the report. A cumulative attention block on the report tokens restores image sensitivity when it begins at or below layer 22 and not from layer 28 upward, a causal depth bound rather than a locus. The labels are derived from reports, which limits conclusions about visual correctness. Code, exact prompts and a committed run record behind every number are provided as anonymized supplementary material and will be released.

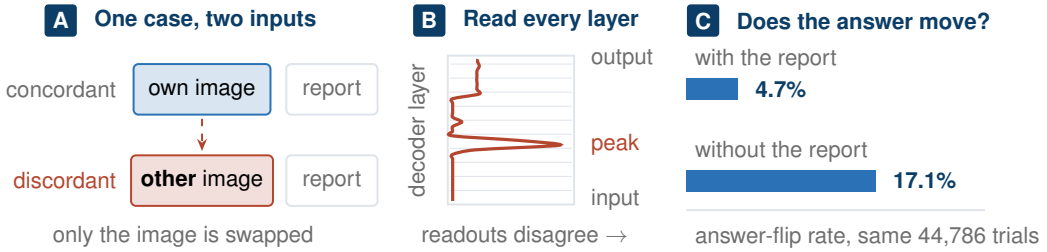



Figure 1: **(A)** One case rendered twice: own image (concordant) or another study's (discordant), report and question fixed. **(B)** Report-token attention knockout, cumulative from the plotted layer upward, 400 one-question cases: the flip rate rises toward the no-report rate when the block starts shallow, overshoots it at layer 22, and is at baseline from layer 32. Blocking from early layers raises image-swap sensitivity and blocking from later layers does not; this measures when direct access to report-token positions matters, not where report information stops being used (Appendix Q, Figure 7). **(C)** All-14 flip rate, report present and absent.

## 1 INTRODUCTION

A radiology report can state an answer the image does not support: it may describe a prior study, be wrong, or name a finding that has since resolved. A model that follows the report is then correct only when the two agree, and a forced-choice output gives no signal when they do not. Whether that is a failure depends on the task; what can be measured is how much the output moves when the image changes and the report stays fixed. That is the line a per-modality failure framework for clinical multimodal models draws between loud failures, which the system flags, and silent ones (Bui et al., 2026). Audits of this kind are one component of the broader engineering agenda for humble,



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

accountable clinical decision support that Arslan et al. (2026) set out, and a model-plane probe of the kind DOJO (Xiang et al., 2026), a community platform for adversarial evaluation of health AI, is designed to run so that shortcut learning and deployment failure modes surface before clinical use. The problem for clinical use is specific: when a model receives an image together with a report that already contains the answer, a correct response does not show that it verified the visual evidence. MODALENS quantifies how report availability changes sensitivity to the image, the quantity at stake when a model is asked to verify a report or to give a second reading rather than to restate it.

Answering from text without looking is well documented in general VQA (Goyal et al., 2017; Agrawal et al., 2018), persists in multimodal LLMs that miss visually obvious detail (Tong et al., 2024) and hallucinate objects (Rohrbach et al., 2018; Li et al., 2023c; Leng et al., 2024), and is an explicit debiasing target in medical VQA (Wan et al., 2025). The closest work is Lotfinia et al. (2026), which scores image dependence apart from accuracy with target and control occlusions and same-label image swaps across nine systems, prompting with the finding question and no report. MODALENS instead makes report availability the manipulated factor, runs the same case twice on the same substitution, and reports the paired flip rate and margin change (Figure 1, the visual summary of the study).

This paper makes three contributions. First, a paired audit that isolates report availability: the same case is scored twice on the same image substitution, with and without its report, so the difference is the report's effect on image sensitivity. Second, its result on MedGemma-27B over 3,199 MIMIC-CXR cases and the 13 finding-specific questions: the generated answer changes with the image on 4.43% of trials with the report and 19.93% without, a paired difference of 15.5 points [14.5, 16.5] (Table 2), replicated in direction in MedGemma-4B, Qwen3.5 at 9B and 27B, and LLaVA-NeXT on Mistral-7B, and traced by an attention knockout to report-token access at layers 0 to 20. Third, the clinically relevant splits that the audit makes possible: the report's certainty and the direction of a disagreement, and a direct test of whether the model will say that a report does not describe the image, which it almost never does. Every number carries a committed run record; code and records are supplementary material. The labels are report-derived throughout, so the audit measures image sensitivity rather than diagnostic correctness (Section 5).

## 2 METHODS

MODALENS measures *counterfactual answer sensitivity*: every case is rendered twice, identical in every token except the image, the *concordant* input with the case's own image and the *discordant* input with another study's. The whole image is substituted, so the quantity measured is sensitivity to the image intervention, not grounding of the queried finding.

**Cohort and units.** MIMIC-CXR test split (Appendix Table 8); a case is one frontal image of a source study with its own substitute. The estimand is the average over trials; weighting studies or patients equally, dropping the cross-patient pairs or those with a blank view field, or keeping only pairs whose known views match gives paired differences between 10.6 and 14.3 points (Appendix I).

**Two question designs.** *All 14 questions* (primary): on every case, 13 finding-specific questions, one per CheXpert finding, plus one composite question asking whether any acute cardiopulmonary finding is shown (Table 1). The composite question is scored against an any-finding target, `yes` iff any of the 13 finding-specific targets is `yes`; the dataset's own No Finding label never enters (Appendix Table 6). *One question per case* (secondary): only the case's own primary labelled finding, or the composite question with target `no` on cases with no positive finding (Section 4.2, Appendix Table 43); a label-conditioned sample, not a finding-detection task.

**Sampling the discordant image.** The substitute is chosen once per case: uniformly among the patient's other frontal test-split studies whose CheXpert positive set differs from the source, failing that any other study of the patient, and for the 19 single-study patients a differently labelled image from another patient. View and acquisition time are not matched; no pair exceeded the specified SSIM threshold of 0.9 (Appendix C and Table 5). Trials on which the substitution changes the asked finding's CheXpert label are *label-changed* and reported separately. That label is not visual ground truth: CheXpert labels come from report text (Irvin et al., 2019), which the model reads in the concordant condition, and the substitute's label comes from a report it never sees. Uncertain and

| Questions per case | Report | Order | Trials | Flip rate | Paired Δ, points | Ratio |
|---|---|---|---|---|---|---|
| All 14 | present | image first | 44,786 | 4.70% [4.35, 5.06] | reference | reference |
| All 14 | absent | image first | 44,786 | 17.07% [15.78, 18.39] | +12.4 [11.2, 13.6] | 3.6 [3.4, 3.9] |
| All 14 | present | text first | 44,786 | 4.91% [4.62, 5.20] | +0.2 [−0.1, 0.5] | 1.0 [1.0, 1.1] |
| All 14 | absent | text first | 44,786 | 12.75% [11.76, 13.76] | −4.3 [−5.4, −3.3] | 0.7 [0.7, 0.8] |
| All 14 | absent minus present | text first | 44,786 | report effect | +7.8 [6.9, 8.8] | 2.6 [2.4, 2.9] |
| One | present | image first | 3,199 | 1.41% [0.99, 1.90] | reference | reference |
| One | absent | image first | 3,199 | 10.28% [8.72, 12.02] | +8.9 [7.5, 10.6] | 7.3 [5.3, 10.5] |
| One | present | text first | 3,199 | 2.84% [2.12, 3.71] | +1.4 [0.6, 2.3] | 2.0 [1.4, 3.1] |

Table 1: MIMIC-CXR report ablation under the original uninstructed protocol, all 14 questions, with the secondary prompt-order analysis; the primary estimate is Table 2 (MedGemma-27B, 3,199 cases, 293 patients). Rows within a block share cases, substitutes and template; paired differences and ratios are against the block's reference row, and the two blocks are different units. The one-question report-present cell is 1.41% here, 45 of 3,199 cases, and 1.38%, 44 cases, in Tables 39, 35 and 37, which rerun that arm; the two runs differ on one case. The two all-14 text-first rate rows are paired against the all-14 image-first row of the same report condition, on identical case ids; the third gives the report effect within text first, so the two orders' report effects can be read side by side. Brackets: 95% patient-clustered percentile bootstrap, 10,000 draws for the all-14 rows and 2,000 for the rest.

absent labels count as negative; relabelling cannot change the flip rate, and under alternative policies the label-changed against label-unchanged difference never separates from zero (Appendix I).

**Prompt and readout.** The model's own chat template, one user turn, no system prompt: an image block, then a text block reading `Report: <report>`, a blank line, and `Question: <question>`; text first swaps the two blocks. Nothing tells the model the report is historical or that it may be unreliable; instructions that treat it as historical context and that warn it may be wrong are tested in Appendix E. The question is a fixed template per finding with no answer-format instruction (Appendix Table 6). The *lowercase first-token readout* is the argmax over the logits of the lowercase `yes` and `no` tokens at the first generated position under greedy decoding; its change under substitution is the *flip rate*. Those tokens (ids in Appendix L) carry negligible probability mass, which sits on the capitalised variants, so the readout is a proxy. A validation under an explicit answer instruction compares it, on both designs, with the token-family and generated-answer readouts under an explicit instruction to answer Yes or No (Section 4.1, Appendix L); the headline prompt carries no such instruction. In the ablation arms of Appendix Table 43, deleting the image builds no image block, so none reaches the model, and deleting the report removes the "Report:" line; nothing else changes. One image per case, so a report-level label may refer to a view the model was not shown.

**Statistics.** Intervals for the MIMIC-CXR analyses on the full cohort are patient-clustered percentile bootstraps over the 293 patients: each bootstrap sample retains all trials belonging to each sampled patient, and paired differences and ratios are computed inside each draw; 10,000 draws for the primary all-14 rates and for Appendix Table 43, 2,000 elsewhere. Subsets and other datasets state their resampling unit and method where they appear. The label-changed against label-unchanged comparison is a two-sided test that can reject equality but not establish it; no equivalence margin was pre-specified. Per-finding, per-category and per-layer analyses are exploratory and uncorrected for multiplicity.

## 3 Experimental setup

MedGemma-27B, the instruction-tuned release (Sellergren et al., 2025), built on Gemma 3 (Gemma Team et al., 2025): 62 decoder layers, residual width 5376, bfloat16, one NVIDIA H200 or H100 per run under Python 3.11, PyTorch 2.10 and Transformers 5.3, with the exact versions of every package in each run record's environment file. Experimental chronology, for reproducibility: the lowercase first-token readout under the plain prompt was run first; the explicit answer instruction, generated-answer scoring, the additional model families and the order rerun followed, and each is dated in its run record; images enter at the processor's default preprocessing (896 by 896 pixels, 256 tokens, pan-and-scan off).

MIMIC-CXR (Johnson et al., 2019a) is the primary dataset; its free-text reports pair with the 14 CheXpert labels (Irvin et al., 2019) from which the questions are built. VQA-RAD (Lau et al., 2018), SLAKE (Liu et al., 2021), OmniMedVQA (Hu et al., 2024) and ProbMed (Yan et al., 2025) have no report (Appendix Table 4).

**Analysis hierarchy.** The primary estimate is the 13 finding-specific questions, 41,587 trials, report present against absent, scored by the generated answer under the explicit answer instruction. The composite question is excluded from it: its wording does not match its target, and on its own it carries a report effect twice that of the finding-specific questions (Section 4.1); it is reported with the 14-question totals as a robustness check. The lowercase first-token readout under the original prompt is reported beside the primary as the original protocol; the one-question design, the other readouts and the other models are robustness checks. The abstract, Figure 1C and the conclusion use the primary estimate. The chronology above records when the instruction and generated-answer scoring were added; the 13-question primary was adopted after external review of the 14-question version, and both are in the run records.

# 4 RESULTS

## 4.1 REPORT AVAILABILITY AND IMAGE SENSITIVITY

**The report effect.** Under the explicit answer instruction, on the 13 finding-specific questions, the model's generated answer changes with the image on 19.93% of trials without the report against 4.43% with it, a paired difference of 15.5 points [14.5, 16.5] and a ratio of 4.5 (Table 2; Figure 1C); this is the primary estimate, and the token-family and lowercase readouts under the same instruction agree to within 0.2 points. The original prompt with the lowercase first-token readout, the paper's first measurement, gives 18.14% against 4.78%, 13.4 points [12.2, 14.6] and a ratio of 3.8 (Appendix Table 25), and is kept as a sensitivity analysis: adding the instruction changes the prompt, so the instructed estimate establishes the effect under that protocol rather than validating every output of the original one. The one-question design agrees in direction. The composite question's wording, whether the image shows any acute cardiopulmonary finding, does not match its target, which is `yes` whenever any of the 13 finding-specific targets is, support devices and fracture included, and on its own it carries a report effect of 32.0 points [29.0, 34.8] (2.13% against 34.10%), twice that of the finding-specific questions. It is therefore excluded from the primary estimate. With it included, all 14 questions and 44,786 trials, the generated answer gives 4.26% against 20.94%, 16.7 points [15.6, 17.7], and the lowercase headline prompt 4.70% against 17.07%, 12.4 points [11.2, 13.6] (Appendix I).

The report-absent rate is not uniform over findings, and the pooled 17.07% hides which ones carry it: five of the 14 sit at or below 4.1% because without the report the model answers yes on nearly every trial, at a yes rate of 97.8% to 100% (atelectasis and support devices flip on no trial at all), while the other nine run from 16.8% to 32.0%. The five are the same findings whose image-only specificity is at or near zero in Appendix Table 41, so the pooled effect varies substantially across findings: several report-absent readouts are nearly constant, and atelectasis and support devices show lower flip rates without the report than with it. The pooled increase should not be read as a uniform finding-level effect; the same pattern holds under the generated answer (Appendix Table 42), where atelectasis is 6.81% with the report against 4.63% without and support devices 11.03% against 1.44%. No pooled figure is a measure of reliability: a finding on which the model answers yes whatever the image needs its own evaluation, and Appendix Table 42 is the per-finding view. On VQA-RAD ($n = 52$), which has no report and questions the key marks image-decisive, the flip rate under a final-token yes/no readout is 61.5% [48.1, 75.0] with accuracy 0.731 on both images.

| Readout | Report present | Report absent | Paired difference, points | Ratio |
|---|---|---|---|---|
| Lowercase first token | 4.77% [4.40, 5.16] | 20.09% [18.81, 21.27] | +15.3 [14.2, 16.4] | 4.2 |
| Token families | 4.46% [4.14, 4.82] | 20.00% [18.83, 21.16] | +15.5 [14.5, 16.5] | 4.5 |
| Generated answer | 4.43% [4.10, 4.80] | 19.93% [18.76, 21.07] | +15.5 [14.5, 16.5] | 4.5 |
| Lowercase first token, headline prompt | 4.78% [4.43, 5.16] | 18.14% [16.83, 19.46] | +13.4 [12.2, 14.6] | 3.8 |

Table 2: The 13 finding-specific questions under the instruction to answer Yes or No (41,587 trials, 3,199 cases, 293 patients): flip rate under three readouts, paired difference absent minus present in points, ratio without interval; last row, the headline prompt without the instruction. Brackets: 95% patient-clustered percentile bootstrap, 2,000 draws.

| | MedGemma-27B | MedGemma-4B |
|---|---|---|
| Concordant accuracy | 0.649 [0.635, 0.665] | 0.692 [0.679, 0.707] |
| Balanced accuracy | 0.767 [0.758, 0.776] | 0.791 [0.783, 0.798] |
| Discordant accuracy | 0.593 [0.579, 0.608] | 0.624 [0.610, 0.640] |
| Flip rate, label changed | 4.37% [3.80, 4.96] | 7.24% [6.53, 7.94] |
| Flip rate, label unchanged | 4.89% [4.51, 5.30] | 7.51% [6.97, 8.04] |
| Flip rate, all trials | 4.78% [4.43, 5.16] | 7.46% [6.96, 7.94] |

Table 3: The 13 finding-specific questions, report present, image first, MIMIC-CXR test (41,587 trials, 3,199 cases, 293 patients). Each image is scored against its own study's report-derived label for the asked question; balanced accuracy is the mean of the yes and no recalls (majority 0.846). Brackets: 95% patient-clustered percentile bootstrap, 2,000 draws.

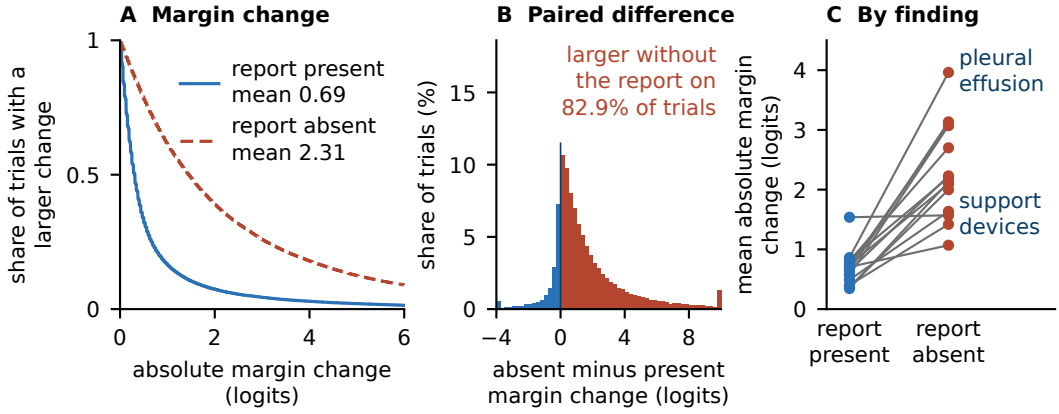



Figure 2: **(A)** Share of the 44,786 all-14 trials whose absolute margin change $|m_d - m_c|$ exceeds $x$, with and without the report. **(B)** Paired difference on the same trial, report absent minus report present; tails fold into the end bins. **(C)** Mean absolute margin change per finding, report present to report absent.

**Continuous margins.** We measure changes in the continuous margin $\log p(\texttt{yes}) - \log p(\texttt{no})$ across all trials, including those without a prediction flip. It is read from the same two lowercase tokens as the flip rate, and over a two-way answer space the log-softmax difference equals the difference of their logits exactly, so the margin is unaffected by how little absolute mass those tokens carry. With the report present the mean absolute paired change is 0.690 [0.657, 0.726]; 32.2% of trials move by more than 0.5, against 4.7% that flip, and so do 33.0% of label-unchanged trials. Without the report the mean change is 2.307, which exceeds the change seen with the report on 82.9% of trials (Figure 2A,B). The rise holds on every finding except support devices, whose mean change is 1.54 with the report and 1.57 without; pleural effusion moves most, 0.86 against 3.96 (Figure 2C). On label-changed trials the target-aligned change $y_d(m_d - m_c)$, with $y_d \in \{-1, +1\}$ the substitute's label, is 0.35 with the report and 1.44 without, positive on 57.9% and 65.4% of trials (Appendix I, Figure 3).

**Label alignment and transitions.** With the report present the flip rate is about the same on label-changed and label-unchanged trials (Table 3), a clustered difference of $-0.50$ points [$-1.03$, $+0.02$]; no equivalence is claimed. Scored against each study's own report-derived label for the asked finding, accuracy is 0.649 on the original image and 0.593 on the substitute [0.579, 0.608] (Table 3). On 60.8% of label-changed trials the answer is correct on the original image and wrong on the substitute, and on 34.8% the reverse: for a text-anchored model this is mostly the answer staying put while the visual evidence changes, not the model changing its mind (Appendix Table 28).

**Report certainty and direction of disagreement.** A hedged report does not send the model to the image. Grouped by the CheXpert status of the asked finding, the 27B flip rate with the report is 1.68% where the report is certain, 2.50% where it is hedged and 5.88% where the finding is not mentioned; the same hedged trials flip 18.53% without the report. The model answers yes on 82% of hedged trials with the report and 86% without, so the affirmative answer is a response tendency present in both conditions rather than a reading of the hedge: uncertainty stated in the report, uncertainty expressed by the model and sensitivity to the image are three separate quantities. A split read from the report text with a clinician-supplied lexicon of 49 hedging terms gives the same picture (Appendix J).

Direction matters more. Where the asked finding's label goes from positive in the report's study to negative in the substitute's (4,710 trials), the answer follows the substitute on 12.7% of trials with the report and 17.2% without. Where it goes from negative to positive (4,637 trials), it follows on 62.8% with the report and 90.5% without, a loss of 27.6 points [25.3, 30.1]; an explicit negative in the report takes that rate to 18.9%, no mention to 68.8%, and a hedge leaves it at 91.1%. Following the substitute means the answer moved with the image, not that it was right. The 4B model repeats every pattern (Appendix J).

**Asking the model to flag the mismatch.** The questions above force a yes or no about a finding, so we also asked one that names none: "Does the report correctly describe this image?", report present, on the report's own image and on the substitute. The 27B answers no on 1.25% [0.80, 1.75] of substituted images against 0.41% [0.20, 0.64] of the report's own, a paired difference of 0.84 points [0.43, 1.34]; under the generated answer it is 3.22% against 1.34%, 1.88 points [1.30, 2.51]. The flag lands where it should, on pairs whose label change runs negative to positive in the substitute (3.18%) rather than positive to negative (0.34%), a difference of 2.84 points [1.45, 4.48], and it never fires on the 55 pairs whose two studies carry the same 13 labels. The detection is therefore real and directional, and far too rare to act on: the substitute comes from a different study on 3,144 of 3,199 pairs and the model objects to at most 3% of them. The 4B model flags 0.19% [0.03, 0.40] and never once on a positive-to-negative pair (Appendix J).

**Robustness.** The 4B model flips 2.7 points more (Table 3). Under the answer instruction the paired report-presence difference is 13.3 points [11.2, 15.6] in Qwen3.5-9B, 19.6 [17.0, 22.4] in Qwen3.5-27B and 12.8 [11.1, 14.4] in LLaVA-NeXT on Mistral-7B, a third lineage, with image reading on that question set at chance without the report in every one of them (Appendix G). Two medical models diverge from this pattern. CheXagent-2-3b, tuned on chest X-rays, flips more with the report than without, 23.7% against 20.7% ($-3.0$ points [$-5.0$, $-1.0$]), while reading the substituted image at chance either way; LLaVA-Med v1.5 keeps the direction, 0.09% against 2.13% ($+2.0$ [1.3, 2.9]), but barely moves with the image even without the report and follows the report's label on the substitute. Neither adds image competence (Appendix G). The flip rate rises with the embedding distance between the images; adjusting for distance and label change, the odds ratio for report presence is 0.12 [0.09, 0.17] (Appendix C). The classifier surrogate agrees (Appendix N). The effect replicates on a second paired-report dataset, the Indiana University collection, with different-patient substitutes: 4.67% with the report against 16.28% without on 300 reports, $+11.6$ points [10.2, 13.2], and the same per-finding pattern including the support-devices reversal (Appendix H).

The effect also survives every prompt variant tested. Manually annotated report labels give the same picture on the 800 trials whose label changes, 0.378 against 0.519 (Appendix F). It holds under a second question template, with the answer options in either order, with an explicit Uncertain option, and under two further substitute seeds (Appendices O, P, P and I). Warning the model that the report may not describe the image and may be wrong leaves the effect intact, a paired difference of $+0.28$ points [$-0.13$, 0.74] (Appendix E). On the headline unit, text-block order does not move the report-present rate, $+0.2$ points [$-0.1$, 0.5], and lowers the report-absent rate by 4.3 points; on the one-question design the report-present rate doubles (Table 1; Appendix K).

**Readout validation.** This validation uses an explicit answer instruction and covers both designs; the headline estimate of 4.70% against 17.07% is a result for the lowercase first-token readout under a prompt that carries no answer instruction. On the one-question design, with the instruction to answer Yes or No appended, the lowercase readout agrees with the generated answer on 98.8% of report-present and 91 to 93% of report-absent cases, and the three readouts agree on the report-presence difference (Appendix Table 39). Without the instruction the lowercase readout gives a smaller report-absent rate than the token-family readout, so the report-absent value depends on the readout choice; the direction of the report effect does not (Appendix L). The primary estimate is robust to the readout: under the same instruction the lowercase and token-family readouts give 15.3 [14.2, 16.4] and 15.5 [14.5, 16.5] points on the same trials (Table 2). The instruction raises the report-absent rate, from 18.14% to 20.09% under the lowercase readout, and leaves the report-present rate within a hundredth of a point of 4.78%, so the instructed and the original prompt measure the same effect at different scale. Under every readout the direction is the same and every interval on the difference excludes zero, with ratios of 4.2 to 4.5.

**Text content.** Text content also affects image sensitivity. In the one-question design, length-matched neutral prose and clinical boilerplate produce flip rates of 8.2% and 7.1%, compared with 10.3% without a report and 1.4% with the full report. Removing the sentences that mention the queried finding raises the rate to 11.6%. A sentence stating the reference answer reduces it to 3.1%, and replacing the target sentences with the opposite label reduces source-label accuracy to 4.3% (Appendix Table 35, Figure 4). These controls suggest that answer-bearing text contributes to the reduced sensitivity, although they do not isolate its contribution from changes in length and context.

### 4.2 Modality ablation: report versus image

On the all-14 design, scored against report-derived labels, the report alone reaches a pooled balanced accuracy of 0.770; adding the image lowers it to 0.717, the image with the question gives 0.665, and the question alone is at chance, answering yes throughout (Appendix Table 40). The two inputs interact: pooled over trials, the image raises balanced accuracy by 0.165 [0.158, 0.172] without the report and lowers it by 0.053 with it, a pooled interaction of $-0.218$. Macro-averaged over the 14 findings the image adds 0.076 without the report and the interaction is $-0.128$ [$-0.141$, $-0.115$], still negative. Pooling over trials matches the paper's estimand, the average over trials; the macro average weights each finding equally and shows that the image-only gain is not uniform, since several findings sit at chance in that cell. Report only is the best cell on 9 of the 14 findings. The one-question design (Appendix Table 43) also ranks report only above the full arm. Both designs are scored against report-derived labels, so a fall when the image is added measures movement away from the report's label, which need not be a visual error; neither the all-14 interaction nor the one-question ranking shows that the image misleads.

### 4.3 Secondary analyses

**Per-finding agreement, chance-corrected.** Raw agreement rewards a skewed responder, so we also report Cohen's $\kappa$ (Cohen, 1960) per finding (Appendix Table 10; all 14 in Table 60). Against chance the ranking changes: for support devices, agreement is 0.946 and $\kappa$ is 0.370, less than half the next lowest, because predictions are strongly skewed toward yes.

**Flip rate by question type.** On SLAKE, with no report, perceptual questions flip more often than knowledge questions: organ 81.0% and modality 80.0% against abnormality 36.0% and knowledge-graph 25.9% (Appendix Table 7). Sample sizes vary by category and no ordered-alternative test was run; comparisons are exploratory.

The layer-wise readouts, the report-token attention knockout and their calibration are secondary analyses and are reported in Appendix Q.

### 4.4 Negative results: steering and probe transfer

Two mechanistic analyses were tried and both are negative (Appendix Q). Difference-of-means steering did not move the share of answers that track the substituted image below the scale that degraded the outputs, where the targeted direction did not separate from a random one, and changed it by 0.1 points at the largest scale. A linear probe fit under a frozen protocol and run once on a disjoint

patient cohort fell from its test-split discrimination to chance. The paper makes no localisation claim beyond that bound.

## 5 LIMITATIONS

**Label validity.** The labels are report-derived throughout. The label-changed split measures consistency with the report, not with the image; the manually annotated report labels of Appendix F check the automated extraction, not the image. We did not obtain independent image annotations for this evaluation; the only image-level labels are the classifier surrogate of Appendix N.

The audit says how much the readout moves with the image, and the opposite-label control says the readout follows the text when the two conflict. Neither says whether following the text was wrong: a report is an expert reading, and deferring to an accurate one may be rational.

No chronology was supplied to the model, so Appendix E is a retrospective stratification, in which later and earlier substitutes give nearly the same report-agreement rate. An instruction asking the model to treat the report as historical context raised the flip rate by about half a point, and an explicit warning that the report may not describe the image and may be wrong changed it by $+0.28$ points $[-0.13, 0.74]$ against the plain framing; these were framing interventions with no temporal consistency enforced, so they do not test the handling of reports that are in fact stale, and neither restores image sensitivity: under the warning the readout still matches the substituted image's label on only 8.49% of the label-changed pairs (Appendix E). The certainty and direction splits of Appendix J, and the mismatch question asked there, inherit the same limitation: the substitute side of a disagreement is that study's report label, so they show where the answer stops moving with the image and where no flag is raised, not where the report was wrong. The near-perfect specificity of that flag follows from answering yes almost everywhere, not from calibration. The validation that would close this gap is independent image annotation: readers blinded to the report shown to the model, adjudication of their disagreements, and a classification of each disagreement's clinical significance, which would say when reduced image sensitivity coincides with a clinically important error. Until then the conclusions are those of a methodological audit.

**Clinical implications.** If a model is used to verify a report or to give a second reading, an answer that barely moves with the image can give a false impression of independent corroboration: an error in the report would be repeated rather than caught, and a discrepancy between report and image would pass undetected. These are potential risks under those uses. This study measures the behaviour; it does not establish that it leads to errors in patient care or to harm.

**Readout validity.** The primary estimate is the generated answer under the explicit answer instruction; the original protocol, kept as a sensitivity analysis, reads the lowercase first token of a prompt that does not request a yes or no. Neither readout can flag conflicting or missing evidence, and both record a shift that does not cross the boundary as no change. The two agree to 0.2 points on the primary question set (Table 2), so the readout choice does not carry the effect (Appendix L). The report-absent one-question rate depends on the readout (10.3% on the paper's tokens, 17.4% on the token families); the report-present rate does not.

**Sampling and prompt dependence.** One substituted image per case, drawn without matching on view, positioning or the other thirteen labels, so a substitute can carry the same finding, which lowers the flip rate even for an image-conditioned model.

The flip rate depends on text-block order, and the headline unit was rerun under both orders to bound that dependence (Table 1). The report effect survives the swap but shrinks: $+7.8$ points [6.9, 8.8] with the text block first against $+12.4$ [11.2, 13.6] with the image first, on identical trials, so the direction and its significance are robust to order while the magnitude is not. The order effect itself is a null with the report present, $+0.2$ points $[-0.1, 0.5]$, which does not reproduce the one-question design's $+1.4$ [0.6, 2.3], and it is negative with the report absent, $-4.3$ points $[-5.4, -3.3]$. That pooled figure is a net of large offsetting per-finding effects rather than a uniform shift: text first lowers the report-absent rate on nine of the 14 findings and raises it on five, over a range from $-21.7$ to $+14.3$ points, and support devices moves from flipping on no trial to 8.07%. No flip rate in this paper should be read as an order-independent property of the model. The headline rests on one dataset, the MIMIC-CXR test split, 293 patients at a single institution, frontal chest radiographs and English reports. A replication on the Indiana University collection, a second institution with different-patient

substitutes and no labels, gives the same direction at a smaller size (Appendix H); it checks the report effect only, not any label-dependent result.

**Generalisation.** The headline rests on one model family and one snapshot; the model card lists MIMIC-CXR and SLAKE in training, which does not by itself establish exposure to these test cases. Whether this split was held out of the pre-training mixture is not disclosed and we could not determine it. The audit is a within-case counterfactual, so it measures dependence on the image even where a case was memorised: memorisation would have to reproduce the substituted study's label to mask the effect. MedGemma-4B shows the same interaction with a larger effect.

The effect is not specific to one lineage or one size: under the answer instruction it replicates in direction and significance across three general-domain families, MedGemma, Qwen3.5 at 9B and 27B, and LLaVA-NeXT on Mistral-7B; in the two medical models tested it is absent (CheXagent-2-3b) or an order of magnitude smaller (LLaVA-Med v1.5), and neither reads the substituted image above chance (Appendix G). On the label-conditioned one-question set no model tested reads the substituted image above chance once the report is removed, so what generalises there is report anchoring, not image competence; on the all-14 unit the image does carry signal for MedGemma, which is why that claim is stated for the one-question set only (Section 4.2, Appendix G).

Every layerwise readout uses the uncalibrated logit lens: the tuned lens (Belrose et al., 2023) scored below it on both readout datasets (Appendix Table 62), and activation patching (Meng et al., 2022) leaves the depth under-determined, so no depth claim rests on a projection, and the knockout below identifies when direct access to report-token positions matters rather than where report information stops being used.

A calibration pass puts the shallowest depth at which the projection tracks the model's own answer at layer 46 of 62 (Appendix Q); everything shallower, the layer-25 feature of Figure 5 included, sits where the projection does not clear the best constant predictor of that answer, and that feature is a property of the report-present prompt rather than of the network. The one depth statement the paper does make is causal rather than projected, and it is a bound and not a locus: a cumulative attention block on the report restores image sensitivity when it begins at or below layer 22 and not from layer 28 upward, report-specific over layers 0 to 20, so the boundary sits near layer 24 of 62; the knockout is not equivalent to deleting the report, no single layer carries the effect, and cross-condition residual patching does not separate from a wrong-donor control at any depth (Appendix Q). These metrics say *whether* the readout changes with the image, not competence.

## 6 Conclusion

On 41,587 paired MIMIC-CXR trials, the 13 finding-specific questions, removing the report raises the rate at which MedGemma-27B's generated answer changes with the image from 4.43% to 19.93%, a paired difference of 15.5 points [14.5, 16.5] under an explicit answer instruction; the original prompt with a lowercase first-token readout gives 4.78% against 18.14%, 13.4 points [12.2, 14.6]. Continuous answer margins are also more sensitive to the image without the report. The direction of the report effect persists across the evaluated controls, under generated-answer scoring of both designs, and across three model families and two model sizes within two of them; since none of those models shows above-chance reading of the substituted image on the one-question set; on six finding-specific questions asked of every case, chosen by a prespecified prevalence rule so the selection cannot score on the label, the report effect holds at 13.1 to 20.9 points in all three (Appendix Table 21), what replicates there is report anchoring rather than image competence. These findings show that report availability changes measured image sensitivity. Sensitivity to an image swap, agreement with a report-derived label and diagnostic correctness are three different things, and this audit measures only the first: independent image annotations are needed to determine when the behaviour produces visually incorrect answers, and until they exist the result is a methodological audit, not a measure of clinical accuracy.

## ETHICS STATEMENT

This work uses MIMIC-CXR and MIMIC-CXR-JPG under the PhysioNet credentialed-access data use agreement, together with three public benchmarks under their own terms. The images and reports are de-identified at source and were read in place under that agreement; no image, report or derived identifier is redistributed with this paper, and no attempt was made to re-identify any subject. The study is a retrospective audit of model behaviour on existing records, involves no human subjects and no clinical deployment, and no reader study is reported here. The evaluation is diagnostic rather than prescriptive: it measures when a model's answer stops depending on the image and reports a failure mode, and the results should not be read as validating any system for clinical use. Every label is derived from report text, so the audit measures consistency with the report and not correctness against the image, a limitation stated in Section 5 that also bounds what any safety claim here can support. Licences: MIMIC-CXR was used under the PhysioNet Credentialed Health Data License 1.5.0 and its data use agreement; VQA-RAD is the CC0 1.0 release archived on OSF, SLAKE the authors' CC BY 4.0 release, OmniMedVQA its public open-access release under its published terms, which pass through each source dataset's licence, and ProbMed its gated release under its dataset-card terms, an MIT licence over images that keep their upstream licences; the IU X-Ray collection is the NLM Open-i release under CC BY-NC-ND 4.0, de-identified at source. MedGemma weights were used under the Health AI Developer Foundations terms of use, Qwen3.5-9B and Qwen3.5-27B (Qwen Team, 2026), LLaVA-NeXT on Mistral-7B (Liu et al., 2024; 2023) and the TorchXRayVision classifier under their Apache 2.0 licences. No dataset was redistributed, and no image or report left institutional hardware. The authors have read the ICLR Code of Ethics and adhere to it.

## REPRODUCIBILITY STATEMENT

Every number in this paper is backed by a committed run record holding the exact command, the resolved model revision, the input checksums, the environment capture and the metrics, and the appendices state the cohort, prompts and estimator behind each result. Appendix B gives the cohort construction, the question set and the pairing rules; Appendix I gives the sensitivity of the primary effect to the pairing and label choices, including the seeds used to redraw the substitute image. Prompts are quoted verbatim where they are introduced, the readout is defined in Section 3, and all intervals are patient-clustered percentile bootstraps whose draw count and seed are stated with the number. The supplementary material contains an anonymized copy of the library, the analysis scripts, the job scripts, the test suite and, for every run, its record: the exact command, the resolved model revision, the input checksums, the environment capture and the aggregate metrics. It contains no images, reports, labels or per-case outputs, and any derived study or subject identifier is redacted, because MIMIC-CXR is credentialed data that cannot be redistributed and must be obtained from PhysioNet under the same agreement.

## REFERENCES

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4971–4980, 2018. doi: 10.1109/CVPR.2018.00522. arXiv:1712.00377.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Janan Arslan, Kurt Benke, Sebastián Andrés Cajas Ordóñez, Rowell Castro, Leo Anthony Celi, Gustavo Adolfo Cruz Suarez, Roben Delos Reyes, Justin Engelmann, Ari Ercole, Almog Hilel, Leo Kinyera, Maximin Lange, Torleif Markussen Lunde, Mackenzie J. Meni, Felipe Ocampo Osorio, Anna E. Premo, Jana Sedlakova, and Pritika Vig. Engineering framework for curiosity-driven and humble AI in clinical decision support. *BMJ Health & Care Informatics*, 33(1):e101877, 2026. doi: 10.1136/bmjhci-2025-101877.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023.

Quang Bui, Shlok Jaiswal, Samuel Paik-Heintz, Kevin Zhou, Kaushik Madapati, Krittaphas Chaisutyakorn, Noah Dane Hebdon, Dimitrios Proios, Sebastián Andrés Cajas Ordóñez, Kacper Dobek, Boya Zhang, Aly Dhedhi, Ahram Han, Kushul Reddy Palakala, Rahul Gorijavolu, Jacques Kpodonu, and Leo Anthony Celi. Loud or silent? a reusable framework for per-modality failure analysis in multimodal clinical AI. *arXiv preprint arXiv:2608.01462*, 2026.

Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, Emily B. Tsai, Andrew Johnston, Cameron Olsen, Tanishq Mathew Abraham, Sergios Gatidis, Akshay S. Chaudhari, and Curtis Langlotz. Chexagent: Towards a foundation model for chest x-ray interpretation. *arXiv preprint arXiv:2401.12208*, 2024.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. doi: 10.1177/001316446002000104.

Joseph Paul Cohen, Joseph D. Viviano, Paul Bertin, Paul Morrison, Parsa Torabian, Matteo Guarrera, Matthew P. Lungren, Akshay Chaudhari, Rupert Brooks, Mohammad Hashir, and Hadrien Bertrand. TorchXRayVision: A library of chest X-ray datasets and models. In *Proceedings of the 5th International Conference on Medical Imaging with Deep Learning*, volume 172 of *Proceedings of Machine Learning Research*, pp. 231–249, 2022. arXiv:2111.00595.

Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer Antani, George R. Thoma, and Clement J. McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. arXiv:1705.08500.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6325–6334, 2017. doi: 10.1109/CVPR.2017.670. arXiv:1612.00837.

Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. OmniMedVQA: A new large-scale comprehensive evaluation benchmark for medical LVLM. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22170–22183, 2024. doi: 10.1109/CVPR52733.2024.02093. arXiv:2402.09181.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 590–597, 2019. doi: 10.1609/aaai.v33i01.3301590. arXiv:1901.07031.

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, 2019a. doi: 10.1038/s41597-019-0322-0.

Alistair E. W. Johnson, Tom J. Pollard, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G. Mark, Seth J. Berkowitz, and Steven Horng. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019b.

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(1):180251, 2018. doi: 10.1038/sdata.2018.251.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13872–13882, 2024. doi: 10.1109/CVPR52733.2024.01316. arXiv:2311.16922.

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. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2023a.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023b. arXiv:2306.03341.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023c. arXiv:2305.10355.

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)*, pp. 1650–1654, 2021. doi: 10.1109/ISBI48211.2021.9434010.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2304.08485.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge. Blog post, `https://llava-vl.github.io/blog/2024-01-30-llava-next/`, January 2024.

Mahshad Lotfinia, Sebastian Ziegelmayer, Lisa Adams, Daniel Truhn, Andreas Maier, and Soroosh Tayebi Arasteh. Vision-language models for chest radiography do not always need the image. *arXiv preprint arXiv:2606.17710*, 2026.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *Conference on Language Modeling (COLM)*, 2024. arXiv:2310.06824.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. arXiv:2202.05262.

nostalgebraist. interpreting GPT: the logit lens. Blog post, LessWrong, August 2020. `https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens`.

Qwen Team. Qwen3.5: Towards native multimodal agents. `https://qwen.ai/blog?id=qwen3.5`, February 2026.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt 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)*, pp. 15504–15522, 2024. doi: 10.18653/v1/2024.acl-long.828. arXiv:2312.06681.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018. arXiv:1809.02156.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. MedGemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal LLMs. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9568–9578, 2024. doi: 10.1109/CVPR52733.2024.00914. arXiv:2401.06209.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

Xingyu Wan, Qiaoying Teng, Jun Chen, Yonghan Lu, Deqi Yuan, and Zhe Liu. Eliminating language bias for medical visual question answering with counterfactual contrastive training. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, Lecture Notes in Computer Science, pp. 194–204. Springer, 2025. doi: 10.1007/978-3-032-04978-0_19.

Alexa Q. Xiang, Takeshi Tohyama, Alexander Cole Bank, Quang Bui, Rahul Gorijavolu, John Anderson Garcia Henao, Nikhil Jaiswal, Akshay Kelshiker, Kaushik Madapati, Sebastian A. Cajas Ordoñez, Milit Patel, Nina Prakash, and Leo Anthony Celi. Distributed open justice oversight (DOJO): A community-driven, modality-agnostic platform for adversarial evaluation of health AI, 2026. SSRN preprint 6676818.

Qianqi Yan, Xuehai He, Xiang Yue, and Xin Eric Wang. Worse than random? an embarrassingly simple probing evaluation of large multimodal models in medical VQA. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19188–19205, 2025. doi: 10.18653/v1/2025.findings-acl.981. arXiv:2405.20421.

## A USE OF LARGE LANGUAGE MODELS

Large language models were used only for coding help, prose editing, and pre-submission feedback via ICLR's Paper Assistant Tool; they did not shape design or interpretation and are not authors.

## B DATASETS, COHORT AND QUESTIONS

| Dataset | Role in this study | Additional clinical context |
|---|---|---|
| MIMIC-CXR (Johnson et al., 2019a) | primary dataset, 14 questions per case | free-text report |
| VQA-RAD (Lau et al., 2018) | image-decisive positive control | none ($n$=52) |
| SLAKE (Liu et al., 2021) | question-type breakdown | none |
| OmniMedVQA (Hu et al., 2024) | modality breadth, 4-way MCQ | none |
| ProbMed (Yan et al., 2025) | false-premise question axis | none (false-premise question) |

Table 4: Datasets used. MIMIC-CXR, the only one pairing each image with a free-text report, carries the report-availability comparison and every headline number; VQA-RAD and SLAKE also appear in the body, as an image-decisive control and a question-type breakdown (Sections 4.1 and 4.3); OmniMedVQA and ProbMed appear only in the supplementary tables.

| Group | Item | | |
|---|---|---|---|
| Swap tier | Same patient, different study | 3,180 | 99.4% |
| | Different patient, different label | 19 | 0.6% |
| Time gap, source to substitute study (days) | Minimum / Q1 | 0.007 | 15.0 |
| | Median / mean | 141.3 | 342.0 |
| | Q3 / maximum | 542.1 | 1,829 |
| | Substitute earlier / later | 1,612 | 1,568 |
| | Gap under 1 day / under 7 days | 3.2% | 17.9% |
| | Gap under 30 days / over 365 days | 32.8% | 33.7% |
| View position, source → substitute | AP → AP / AP → PA | 1,546 | 406 |
| | PA → PA / PA → AP | 363 | 431 |
| | Either view unlabelled | 453 | |
| | Same view, both known | 1,909 / 2,746 | 69.5% |
| CheXpert labels differing (of 14) | None / exactly one | 22 | 402 |
| | Exactly two / three or more | 937 | 1,838 |
| | Median / mean | 3 | 2.89 |
| | Asked answer changes (own finding) | 2,321 | 72.6% |
| Reuse and reciprocity | Distinct substitute images / studies | 2,093 | 1,953 |
| | Maximum reuse of one image / study | 6 | 8 |
| | Pairs whose reverse pair also exists | 376 | 11.8% |
| | Source studies / with more than one pair | 2,816 | 344 |

Table 5: Composition of the 3,199 MIMIC-CXR discordant pairs (seed 0). Time gaps use the 3,180 same-patient pairs. Label differences count CheXpert findings whose positive status differs between source and substitute.

| Question | Question text | Target |
|---|---|---|
| 13 finding-specific: Atelectasis, Cardiomegaly, Consolidation, Edema, Enlarged Cardiomediastinum, Fracture, Lung Lesion, Lung Opacity, Pleural Effusion, Pleural Other, Pneumonia, Pneumothorax, Support Devices | *Does this chest X-ray show <finding>?*, finding in lower case, e.g. *Does this chest X-ray show support devices?* | `yes` iff the finding is in the image's CheXpert positive set (report-derived) |
| Composite (any finding) | *Does this chest X-ray show any acute cardiopulmonary finding?* | `yes` iff any of the 13 finding-specific targets is `yes`, `no` otherwise; the dataset's No Finding label is not used |

Table 6: The 14 question templates and target rules, 13 finding-specific and one composite. The one-question design asks each case only its primary positive finding, the first positive in fixed CheXpert order, or the composite question with target `no` on the 492 cases with none. No answer-format instruction is appended.

| Content type | $n$ | Flip rate | Concordant acc. |
|---|---|---|---|
| Organ | 42 | 80.95% [69.05, 90.48] | 0.857 [0.738, 0.952] |
| Modality | 20 | 80.00% [60.00, 95.00] | 0.900 [0.750, 1.000] |
| Position | 10 | 40.00% [10.00, 70.00] | 0.700 [0.400, 1.000] |
| Abnormality | 50 | 36.00% [24.00, 50.00] | 0.640 [0.500, 0.760] |
| Knowledge-graph | 54 | 25.93% [14.81, 38.89] | 0.593 [0.463, 0.722] |

Table 7: SLAKE by question type, no report, 176 of the 186 cases; Plane ($n$=8) and Color ($n$=2) are omitted for size. Sample sizes vary by category; comparisons are exploratory. Brackets: 95% percentile bootstrap over the cases in each category, 5,000 draws.

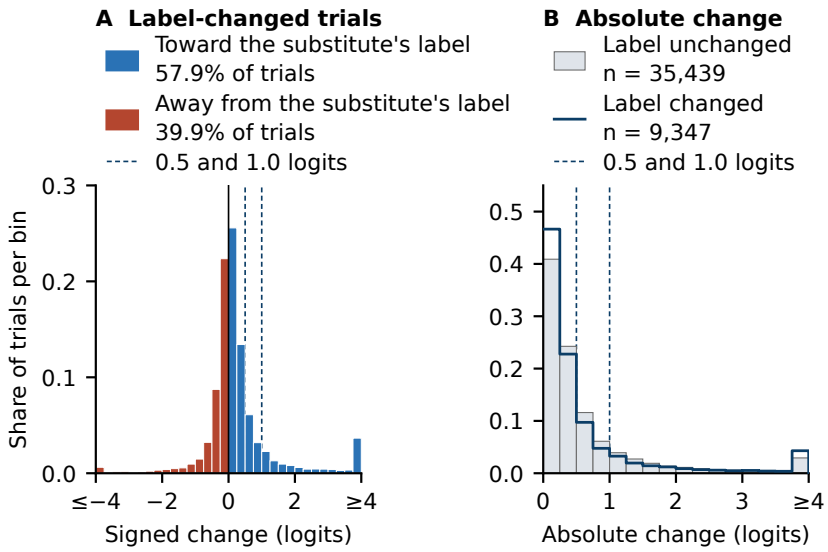



Figure 3: Paired margin change, all 14 questions, report present, image first. **(A)** Signed, target-aligned change $y_d(m_d - m_c)$ on the 9,347 label-changed trials, shaded by direction; mean 0.35 [0.29, 0.41], median 0.08. **(B)** Absolute change $|m_d - m_c|$ on the 35,439 label-unchanged and the same 9,347 label-changed trials. End bins collect the tails.

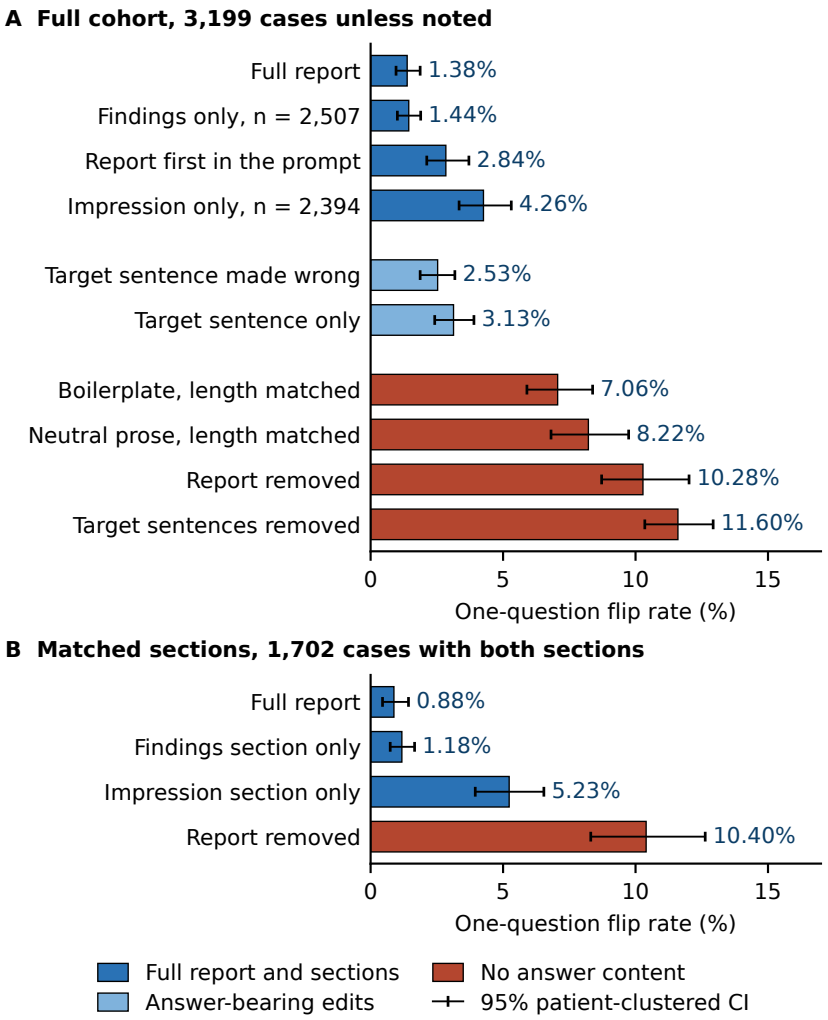



Figure 4: One-question flip rate by text condition, MedGemma-27B, patient-clustered 95% intervals. **(A)** Full cohort, 3,199 cases unless noted (Appendix Table 35). **(B)** The 1,702 cases with both sections, all scored on the same cases: findings alone within 0.3 points of the full report, impression alone 4.4 points above it [3.0, 5.7].

| Level | Count | Note |
|---|---|---|
| Patients | 293 | resampling unit of the intervals |
| Source studies | 2,816 | 344 contribute more than one case |
| Cases (one frontal image, one substitute) | 3,199 | 3,180 within patient, 19 across patients |
| Substitute studies / images | 1,953 / 2,093 | reused up to 8 / 6 times; 376 reciprocal pairs |
| Questions per case | 14 | 13 finding-specific plus one composite |
| Paired trials | 44,786 | 9,347 with the asked finding's label changed |

Table 8: Cohort accounting; pair composition in Appendix Table 5.

## C How much the image changes

Two distances between the source and substitute image, all 3,199 pairs. Flip rates are the one-question design; the regression pools both conditions (6,398 pair-condition rows, cluster-robust standard errors).

| Quantity | Value |
|---|---|
| SSIM, quartiles | 0.462 / 0.528 / 0.586 (max 0.78) |
| Pairs above the 0.9 SSIM threshold | 0 |
| Vision-tower cosine distance, quartiles | 0.395 / 0.563 / 0.745 |
| Spearman, distance against time gap | 0.16 |
| Flip, no report, nearest quartile | 8.5% [6.3, 11.0] |
| Flip, no report, farthest quartile | 14.5% [11.7, 17.4] |
| Difference | +6.0 [2.5, 9.5] |
| Flip, report, nearest to farthest | 0.6% to 1.9% |
| Odds ratio per SD of distance | 1.30 [1.16, 1.46] |
| Odds ratio, label change | 1.43 [1.05, 1.96] |
| Odds ratio, report present | 0.12 [0.09, 0.17] |

Table 9: Image distance and its relation to the flip rate. SSIM is computed at the model's 896 by 896 greyscale input; the cosine distance is between mean-pooled vision-tower embeddings, cohort-mean-centred. Brackets: 95% intervals.

## D PER-FINDING AGREEMENT

| Finding | Observed agreement | Chance-expected | Raw excess | Cohen's $\kappa$ |
|---|---|---|---|---|
| Pneumothorax | 0.992 [0.988, 0.995] | 0.919 | 0.073 | 0.899 [0.837, 0.943] |
| Any finding (composite) | 0.964 [0.957, 0.971] | 0.501 | 0.463 | 0.927 [0.913, 0.941] |
| Atelectasis | 0.951 [0.942, 0.960] | 0.698 | 0.254 | 0.839 [0.807, 0.867] |
| Cardiomegaly | 0.936 [0.923, 0.948] | 0.500 | 0.435 | 0.871 [0.845, 0.896] |
| Support devices | 0.946 [0.931, 0.959] | 0.914 | 0.032 | 0.370 [0.271, 0.451] |

Table 10: Five of 14 findings, each $n$=3,199; pooled observed agreement is 0.953. Chance-expected agreement is $p_c p_d + (1 - p_c)(1 - p_d)$ from the condition-specific yes-rates, raw excess is observed minus expected, $\kappa$ is excess over one minus expected. $\kappa$ corrects for imbalance, not for grounding. Brackets: 95% patient-clustered percentile bootstrap over the 293 patients, 2,000 draws at seed 0. $\kappa$ depends on both branches' marginals, so each draw rebuilds them from the resampled cases rather than averaging a per-case value; the three indicators are resampled jointly so the pairing is preserved. Support devices is the one finding whose interval is wide and low, [0.271, 0.451], which is what a 0.914 chance level leaves room for.

## E RETROSPECTIVE CHRONOLOGY ANALYSIS OF STALE REPORTS

No chronology was supplied to the model, so this is a retrospective stratification by chronology, not a test of whether the model reasons about staleness. Among the same-patient pairs whose asked-finding label differs between the two studies, the substitute study is later than the shown report on 1,038 pairs and earlier on 1,130 (Table 11); Table 12 gives the agreement rates. With the report present the readout follows the report's label at nearly the same rate for later and earlier substitutes, and the rate does not move with the time gap. The cell nearest the stale-report scenario, a finding labelled in a later study's report that the shown report does not mention, has the lowest report-agreement rate; whether this corresponds to clinical harm is not measured, and in this pair set it is the composite question only, because the one-question design asks about the source study's own positive finding.

| Quantity | Pairs |
|---|---|
| Same-patient pairs | 3,180 |
| Substitute without a CheXpert value | 171 |
| Substitute study with the same value | 841 |
| Asked-finding label differs (254 patients) | 2,168 |
| substitute later than the shown report | 1,038 |
| substitute earlier | 1,130 |
| Label-changing over all 3,199 pairs | 2,321 |
| Pairs with a missing or zero study-time gap | 0 |

Table 11: Cohort of the chronology analysis. The 2,168 are the same-patient subset of the 2,321 label-changing pairs of Table 5 on which the substitute study also carries a CheXpert value for the asked finding.

| Pairs | Agrees with report label | Agrees with substitute label |
|---|---|---|
| All 2,168, report present | 0.912 [0.897, 0.926] | 0.088 [0.074, 0.103] |
| All 2,168, report absent | | 0.321 [0.289, 0.351] |
| Present minus absent | | $-0.232$ [$-0.262$, $-0.202$] |
| Substitute later (1,038), report present | 0.899 [0.878, 0.919] | |
| Substitute earlier (1,130), report present | 0.924 [0.904, 0.942] | |
| Later minus earlier | $-0.025$ [$-0.054$, $+0.002$] | |
| Odds ratio per log day of time gap | 0.94 [0.87, 1.01] | |
| Later study's finding not in the shown report (188) | 0.824 [0.762, 0.878] | |

Table 12: Readout agreement by chronology on the 2,168 same-patient pairs whose asked-finding label differs between the two studies, one-question design, report-derived labels. Brackets: 95% patient-clustered percentile bootstrap over the 254 patients, 2,000 draws.

**Sensitivity to an instruction treating the report as historical or as unreliable.** A framing intervention only: one sentence before the report stating that it describes an earlier study of the same patient and that the question is to be answered from the current image, under the same answer instruction as Appendix L. Temporal consistency was not enforced and no study dates were checked, so this measures sensitivity to the instruction, not performance on genuinely stale reports.

| Quantity | Value |
|---|---|
| Flip, historical framing | 1.72% [1.25, 2.25] |
| Flip, plain Report label, same instruction | 1.22% |
| Paired difference | $+0.50$ [0.07, 0.94] |
| Same under the generated answer | $+0.60$ points |
| Accuracy against the concordant label, original image | 92.0% |
| Accuracy against the concordant label, substitute | 92.2% |
| Gap to the no-report condition (18.26%) | 16.54 [14.46, 18.86] |
| Against report present, no instruction (1.41%) | $+0.31$ [$-0.27$, 0.87] |

Table 13: Historical-context framing. The framing does not restore the sensitivity seen without the report. Brackets: 95% patient-clustered percentile bootstrap.

**An explicit reliability warning.** The historical framing only relocates the report in time, so the report-present condition was also rerun under a stronger sentence: the report may not describe this image and may be wrong, answer from the image, the report is unverified context. Neither framing restores the sensitivity seen without the report (Table 14).

| Quantity | Value |
|---|---|
| Flip, reliability warning | 1.50% [1.06, 1.98] |
| Flip, plain Report label, same instruction | 1.22% |
| Paired difference against the plain label | $+0.28$ [$-0.13$, 0.74] |
| Paired difference against the historical framing | $-0.22$ [$-0.67$, 0.22] |
| Flip, token-family / generated readout | 1.47% / 1.38% |
| Generated-answer shift against the plain framing | $+0.53$ [0.12, 0.97] |
| Gap to the no-report condition (18.26%) | $-16.76$ [$-18.93$, $-14.74$] |
| Substitute image: matches the report's label | 93.12% |
| Substitute image: matches the image's label | 32.89% |
| on the 2,321 label-changing pairs | 8.49% [7.22, 9.93] |
| Format compliance, own / substitute | 99.94% / 99.87% |

Table 14: Reliability warning, report present, same answer instruction. The warning does not restore image sensitivity: on the substituted image the readout still follows the report's label on 93% of pairs. Brackets: 95% patient-clustered bootstrap.

**On the pairs where the framing is true.** The historical instruction was applied to every pair but is true only where the substitute image is a later study than the one the shown report describes. Chronology reconstructed from the signed study gap splits the 3,180 same-patient pairs into 1,568 later and 1,612 earlier, with no equal-date or undated pairs. No difference between the strata was detected (Table 15): the instruction does not measurably restore image sensitivity on the pairs where what it says is the case. This is the nearest the data comes to a stale-report condition without independently dated images; no report was rewritten, and the absence of a detected difference is not evidence that stale reports are handled correctly.

| Quantity | Later (1,568) | Earlier (1,612) |
|---|---|---|
| Flip under the historical framing, readout | 1.72% | 1.74% |
| Difference | $-0.02$ [$-1.03$, $+1.06$] | |
| Flip under the historical framing, generated answer | 1.66% | 0.99% |
| Difference | $+0.67$ [$-0.27$, $+1.66$] | |

Table 15: The historical framing where it is true against where it is not. Intervals resample the 276 patients once per draw and score both strata on that draw, since a patient can contribute to both.

## F Sensitivity analysis using manually annotated report labels

We repeated the analysis using the manually annotated report labels supplied with MIMIC-CXR-JPG: the 687 test-set reports annotated by a single radiologist to evaluate the automated labelers (Johnson et al., 2019b). This checks sensitivity to automated label extraction, not the findings in the displayed images: the annotator read the report, not the image. Applied at study level, 291 of the 3,199 pairs (90 patients) have manual labels on both studies, 4,074 of the 44,786 trials. On those studies the automated and manual report labels agree on 0.943 [0.938, 0.947] (Cohen's $\kappa$ 0.795; automated sensitivity 0.894, precision 0.773 against the manual label), which estimates agreement between automated and manual labels in this subset. Table 16 gives the readout on the 800 trials whose manual label of the asked finding differs between the two studies: the report effect keeps its direction. Scored against the manual rather than the automated label, concordant accuracy is 0.627 (balanced 0.712) with the report and 0.485 (balanced 0.644) without; the report-absent readouts answer `yes` on 73% of trials, so the balanced accuracies are the fair comparison.

| Statistic | Report present | Report absent | Absent minus present |
|---|---|---|---|
| Flip rate, 800 manual-label-changed trials | 3.13% [1.87, 4.48] | 12.50% [9.60, 16.13] | +9.4 points [6.1, 13.3] |
| Flip rate, 3,274 same-label trials | 4.40% | | |
| Agrees with the substitute's manual label, 800 trials | 0.378 [0.342, 0.410] | 0.519 [0.487, 0.552] | +0.141 [0.100, 0.187] |
| Agrees with the source report's label, 800 trials | 0.616 [0.580, 0.651] | | |
| Agrees with the substitute's manual label, 150 trials, one-question design | 0.053 [0.022, 0.088] | 0.147 [0.087, 0.208] | |
| Agrees with the report, 717 trials whose substitute's manual label contradicts it | 0.633 [0.596, 0.671] | 0.481 [0.444, 0.518] | |
| The same 717 trials, one-question design | 0.979 [0.956, 1.000] | | |

Table 16: Readout under the manually annotated report labels of MIMIC-CXR-JPG, with and without the report, all 14 questions unless stated. Brackets: 95% patient-clustered percentile bootstrap over the 90 covered patients, 2,000 draws; unbracketed cells have none.

## G Other models: MedGemma-4B, a second family and a third

MedGemma-4B run through the same four one-question cells on the same 3,199 cases shows the same interaction with a larger effect.

| Quantity | MedGemma-4B | MedGemma-27B |
|---|---|---|
| One-question flip, report present | 3.16% [2.46, 3.92] | |
| One-question flip, report absent | 19.01% [16.64, 21.40] | |
| Paired difference | 15.85 [13.45, 18.25] | |
| Ratio | 6.0 [4.8, 7.8] | |
| Balanced accuracy, original image, report present | 0.812 | |
| Balanced accuracy, original image, report absent | 0.810 | 0.479 |
| Specificity, report absent | 0.77 | 0.06 |
| Answers yes, report absent, of 3,199 | 2,422 | 2,911 |
| All-14 flip, report present | 7.42% | |
| All-14 flip, report absent | 22.52% | |
| All-14 paired difference | 15.1 [13.8, 16.3] | |
| All-14 ratio | 3.0 [2.8, 3.3] | |

Table 17: MedGemma-4B against MedGemma-27B on the same cases, each image scored against its own study label. The two report-absent balanced accuracies describe the same cell in different models: without a report the 27B answers yes almost everywhere, so its specificity collapses. Brackets: 95% patient-clustered bootstrap.

Qwen3.5-9B (Qwen Team, 2026), a non-Gemma image-text model, was run through the one-question design on the same cases, report present and absent, both images, with the same prompt, its own chat template, thinking disabled and images capped at 802,816 pixels. Its token-family readout takes the argmax between the summed probabilities of the yes and no families at the first generated position. Without an answer instruction the model answered `yes` on 3,197 of 3,199 report-absent cases, a saturated prior under which the effect could not be established, so the instruction of Appendix L was appended; every generation was then a bare Yes or No.

| Quantity | Value |
|---|---|
| Flip, report present | 1.3% [0.9, 1.7] |
| Flip, report absent | 14.5% [12.5, 16.9] |
| Paired difference | 13.3 [11.2, 15.6] |
| Ratio | 11.6 |
| Agreement with the report's label, report present | 0.916 |
| Balanced accuracy against the substituted image's label | 0.48 |
| Generated agrees with token family | 0.984 to 0.997 |
| Bare yes or no first token, no instruction | 13 to 34% |

Table 18: Qwen3.5-9B under the answer instruction, one H100. Family probabilities are summed by log-sum-exp over logits. Reading the substituted image at chance replicates report anchoring, not image competence. Brackets: 95% patient-clustered bootstrap.

**A third lineage and a scale check.** Two further models on the same 3,199 cases and 293 patients under the instruction "Answer with exactly Yes or No.": LLaVA-NeXT on Mistral-7B (Liu et al., 2024; 2023), a lineage that is neither Gemma nor Qwen, and Qwen3.5-27B, three times the size of the Qwen3.5-9B arm. The report-presence effect holds in both and scale does not shrink it (Table 22). Both readouts are validated here: format compliance is 1.000 in all eight cells, none of the 25,592 generations was unscoreable, and the generated answer agrees with the token family on at least 0.988 of cases. What replicates is report anchoring, not image competence.

| Model | Substituted image | Own label, present / absent |
|---|---|---|
| LLaVA-NeXT on Mistral-7B | 0.470 [0.452, 0.488] | 0.925 / 0.670 |
| Qwen3.5-27B | 0.516 [0.492, 0.539] | 0.943 / 0.848 |
| Qwen3.5-9B | 0.487 [0.462, 0.511] | |
| MedGemma-27B, token family | 0.480 | |
| MedGemma-4B, paper readout | 0.518 [0.492, 0.542] | |
| CheXagent-2-3b | 0.508 | 0.757 / 0.841 |
| LLaVA-Med v1.5 | 0.368 | 0.954 / 0.872 |

Table 19: Balanced accuracy reading each image against its own label, report removed. No model tested reads the substituted image above chance here. The Qwen intervals include 0.5, which is insufficient evidence of above-chance reading, not evidence of chance; the LLaVA-NeXT interval sits below 0.5. The set is label-conditioned (target `yes` on 2,707 of 3,199), so this does not carry to the all-14 unit. Brackets: 95% patient-clustered bootstrap.

**Questions selected independently of the source label.** The one-question design asks a finding-specific question on the 2,707 positive source cases and the composite question on the 492 negative ones, so a rule answering yes to every finding-specific question and no to the composite scores perfectly without reading anything. To remove that shortcut the same four cells were rerun on a prespecified question set: the six findings whose source-label yes-rate lies between 20% and 80%, fixed before any result was seen. That gives 19,194 trials per model over 293 patients, 6,610 of them label-changing on the asked finding's own label (Table 21).

| Model | Image-only balanced accuracy |
|---|---|
| Qwen3.5-9B | 0.635 [0.620, 0.651] |
| Qwen3.5-27B | 0.664 [0.650, 0.678] |
| LLaVA-NeXT on Mistral-7B | 0.538 [0.524, 0.553] |

Table 20: The prespecified set: atelectasis 22.3%, cardiomegaly 27.1%, edema 20.8%, lung opacity 32.0%, pleural effusion 32.4%, support devices 32.7%. The report effect holds in every model, 13.1 to 20.9 points. The Qwen scores carry sensitivity near 0.9 and specificity 0.3 to 0.4, a yes bias, so above chance is not competent. No question-only cell was run. Brackets: 95% patient-clustered bootstrap.

**A chest-X-ray-tuned model.** CheXagent-2-3b (Chen et al., 2024), a phi-2 backbone with an X-ray SigLIP tower at 512 by 512 pixels, trained for chest radiograph interpretation, was run through the same four cells under the answer instruction with its own chat template and image tokens. It is the one model whose sign reverses (Table 22): the answer changes with the image on 23.76% [21.38, 26.37] of substitutions with the report and 20.69% [18.36, 23.03] without, $-3.06$ points [$-5.03$, $-1.05$] under the token-family readout and $-3.01$ [$-4.99$, $-1.00$] under the generated answer. The report does not anchor it; it raises the yes rate on the report's own image from 0.745 to 0.842 and lowers balanced accuracy against that image's own label from 0.841 to 0.757. Nor does the model read the substitute: balanced accuracy against the substitute study's label is 0.508 without the report and 0.480 with it (Table 19). Domain tuning therefore removed the anchoring without adding image competence on this set. Format compliance was 1.000 in three cells and 0.996 in the fourth (12 unscoreable generations); the run used the checkpoint's own environment (transformers 4.40) because its remote code requires it, and its balanced accuracies are point estimates. No question-only cell was run.

**A biomedical model.** LLaVA-Med v1.5 on Mistral-7B (Li et al., 2023a), instruction-tuned on PubMed figure-caption data, was run through the same four cells. Its checkpoint is in the original LLaVA layout, so the weights were renamed into the transformers LLaVA layout with the vision tower taken from inside the checkpoint and a tokenizer whose model file is byte-identical to the published one; images are padded to square as the checkpoint specifies, 576 image tokens, the Mistral instruction template. The direction of the report effect replicates and is significant, 0.09% [0.00, 0.21] with the report against 2.13% [1.37, 3.00] without, $+2.03$ points [1.25, 2.93] under the token-family readout and $+1.03$ [0.55, 1.60] under the generated answer, but the size is an order of magnitude below every other family: the model barely moves with the image even without the report. No case has identical logits on the two images, and the shift grows without the report (median absolute change in $\log p(\texttt{yes})$ 0.005 with it, 0.048 without), so the pipeline is not the cause. Balanced accuracy against the substitute study's label is 0.368 without the report and 0.376 with it, below chance, which on a label-conditioned set means the movement that exists follows the text prior; against the report's own label it is 0.954 with the report and 0.872 without (Table 19). Format compliance was 0.993 and 0.981. The model's intended use is research only, and its accuracies are not comparable across families at different input resolutions.

| | Generated-answer flip rate | | Report effect | Image-only balanced accuracy | | Follows sub. label |
|---|---|---|---|---|---|---|
| Model | Report present | Report absent | points | Original image | Substitute image | no report |
| Qwen3.5-9B | 6.04% [5.5, 6.6] | 19.09% [17.1, 21.4] | +13.1 [11.2, 15.1] | 0.635 [0.620, 0.651] | 0.635 [0.620, 0.651] | 56.23% |
| Qwen3.5-27B | 6.39% [5.9, 6.9] | 23.90% [22.0, 25.9] | +17.5 [15.8, 19.3] | 0.664 [0.650, 0.678] | 0.664 [0.649, 0.679] | 57.70% |
| LLaVA-NeXT-Mistral-7B | 4.33% [3.9, 4.8] | 25.24% [24.1, 26.3] | +20.9 [19.9, 21.9] | 0.538 [0.524, 0.553] | 0.538 [0.524, 0.551] | 50.17% |

Table 21: Cross-model results on six finding-specific questions selected by a prespecified prevalence rule and asked of every case, 19,194 trials per model, generated answers under the explicit instruction. Balanced accuracy is against each image's own report-derived label with no report shown. Both substitute-image columns use the asked finding's own label on the substitute study, on which 6,610 of the 19,194 trials change; the last column is the share of those trials whose no-report answer follows the substitute's label. The pairs record's coarse flag, which tracks the substitute's whole positive set, would count 9,531 and was used in an earlier version of this table. Brackets: 95% patient-clustered percentile bootstrap over 293 patients, 2,000 draws; the report effect resamples patients once per draw and scores both conditions on that draw.

| Model | Report present | Report absent | Paired difference, points |
|---|---|---|---|
| MedGemma-27B | 0.84% [0.50, 1.27] | 17.69% [15.88, 19.61] | +16.9 [15.0, 18.8] |
| Qwen3.5-9B | 1.3% [0.9, 1.7] | 14.5% [12.5, 16.9] | +13.3 [11.2, 15.6] |
| Qwen3.5-27B | 0.88% [0.56, 1.22] | 20.48% [17.95, 23.26] | +19.60 [17.00, 22.40] |
| LLaVA-NeXT Mistral-7B | 2.59% [1.97, 3.31] | 15.35% [13.65, 17.13] | +12.75 [11.10, 14.44] |
| CheXagent-2-3b (chest X-ray tuned) | 23.76% [21.38, 26.37] | 20.69% [18.36, 23.03] | −3.06 [−5.03, −1.05] |
| LLaVA-Med v1.5 Mistral-7B (biomedical) | 0.09% [0.00, 0.21] | 2.13% [1.37, 3.00] | +2.03 [1.25, 2.93] |

Table 22: Six models on one axis: one-question design, 3,199 cases and 293 patients, token-family readout under the instruction to answer Yes or No, paired difference absent minus present in points. Brackets: 95% patient-clustered percentile bootstrap, 2,000 draws.

## H A SECOND PAIRED DATASET: IU X-RAY

The report effect was rerun on the Indiana University chest X-ray collection (Demner-Fushman et al., 2016), distributed by NLM Open-i under CC BY-NC-ND 4.0 and de-identified at source: a different institution, different reporting style, and no overlap with MIMIC-CXR. The primary protocol was kept exactly: MedGemma-27B, the 13 finding-specific questions, the report as findings plus impression, the answer instruction, 8-token greedy generation, both readouts. Of 3,955 reports, 3,809 had a usable frontal image and a non-empty report; the first 300 by identifier were run, 3,900 trials per cell. The collection has no longitudinal studies and no patient identifier, so every substitute is the frontal image of a different report, the paper's random tier, and the absolute rates are not comparable to the same-patient MIMIC headline; the paired report effect is. It has no CheXpert labels, so there is no accuracy, no balanced accuracy and no label-changed split. Intervals resample the 300 reports.

The effect replicates (Table 23): with the report the generated answer changes with the image on 4.67% of trials, without it on 16.28%, a paired difference of 11.6 points [10.2, 13.2], against 15.5 on MIMIC-CXR; the report-present rate is nearly identical across the two datasets and the report-absent rate is lower here. The per-finding pattern is the MIMIC one, including the one reversal: support devices flips more with the report than without (Table 24).

| Readout | Report present | Report absent | Difference | Ratio |
|---|---|---|---|---|
| Generated answer | 4.67% [4.00, 5.31] | 16.28% [14.77, 17.85] | +11.6 [+10.2, +13.2] | 3.5 |
| Token family | 4.67% [4.00, 5.31] | 16.18% [14.67, 17.74] | +11.5 [+10.1, +13.0] | 3.5 |
| Lowercase first token | 5.87% [5.10, 6.72] | 15.38% [13.72, 17.10] | +9.5 [+7.9, +11.2] | 2.6 |

Table 23: IU X-Ray, MedGemma-27B, 300 reports, 13 finding-specific questions, 3,900 trials per cell, different-report substitutes. Brackets: 95% report-clustered percentile bootstrap, 2,000 draws.

| Finding | Report present | Report absent | Difference |
|---|---|---|---|
| Atelectasis | 2.0% | 43.0% | +41.0 [+34.7, +46.3] |
| Cardiomegaly | 0.7% | 19.3% | +18.7 [+14.7, +23.0] |
| Consolidation | 0.7% | 6.3% | +5.7 [+3.0, +8.7] |
| Edema | 0.0% | 10.7% | +10.7 [+7.3, +14.3] |
| Enlarged cardiomediastinum | 0.7% | 8.0% | +7.3 [+4.3, +10.7] |
| Fracture | 0.0% | 1.3% | +1.3 [+0.3, +2.7] |
| Lung lesion | 8.0% | 28.3% | +20.3 [+15.3, +25.3] |
| Lung opacity | 8.7% | 47.0% | +38.3 [+32.0, +44.3] |
| Pleural effusion | 1.0% | 14.7% | +13.7 [+10.0, +17.7] |
| Pleural other | 0.3% | 6.3% | +6.0 [+3.3, +9.0] |
| Pneumonia | 0.0% | 10.3% | +10.3 [+7.0, +14.0] |
| Pneumothorax | 0.0% | 1.7% | +1.7 [+0.3, +3.3] |
| Support devices | 38.7% | 14.7% | −24.0 [-31.0, -17.0] |

Table 24: IU X-Ray per finding, generated answer, 300 trials each. Brackets: 95% report-clustered percentile bootstrap.

# I SENSITIVITY OF THE PRIMARY EFFECT TO PAIRING AND LABEL CHOICES

| Subset or weighting | Trials | Report present | Report absent | Difference, points |
|---|---|---|---|---|
| All 14 questions | 44,786 | 4.70% [4.35, 5.06] | 17.07% [15.78, 18.39] | +12.4 [11.2, 13.6] |
| 13 finding-specific questions (primary) | 41,587 | 4.78% [4.43, 5.16] | 18.14% [16.83, 19.46] | +13.4 [12.2, 14.6] |
| Cross-patient pairs excluded | 44,520 | 4.70% [4.34, 5.07] | 16.99% [15.70, 18.26] | +12.3 [11.1, 13.4] |
| Same known view | 26,726 | 4.61% [4.15, 5.10] | 15.21% [13.85, 16.60] | +10.6 [9.3, 11.9] |
| Both views known | 38,444 | 4.80% [4.43, 5.21] | 16.80% [15.44, 18.20] | +12.0 [10.8, 13.3] |
| Known view mismatch | 11,718 | 5.23% [4.68, 5.76] | 20.43% [18.43, 22.49] | +15.2 [13.3, 17.2] |
| Equal-study weighting | 44,786 | 4.70% [4.36, 5.06] | 17.01% [15.71, 18.30] | +12.3 [11.2, 13.5] |
| Equal-patient weighting | 44,786 | 4.46% [4.05, 4.89] | 18.75% [17.62, 20.01] | +14.3 [13.2, 15.5] |

Table 25: Primary effect under alternative pairings, weightings and question sets, all 14 questions unless stated, each subset resampled over its own patients. Same known view: AP to AP or PA to PA; both views known: 453 blank-view pairs excluded. Brackets: 95% patient-clustered percentile bootstrap, 10,000 draws for the primary row and 2,000 for the subsets.

**Label policies** (asked finding, both images, raw CheXpert values).

| Quantity | Uncertain negative | Uncertain positive | Uncertain excluded |
|---|---|---|---|
| Label-changed trials | 9,347 | 10,626 | 8,660 |
| Trials retained | 44,786 | 44,786 | 42,032 |
| Flip, label changed | 4.30% | 4.65% | 4.53% |
| Flip, label unchanged | 4.81% | 4.72% | 4.79% |
| Changed minus unchanged | −0.50 [−1.03, 0.02] | −0.07 [−0.56, 0.42] | −0.26 [−0.82, 0.27] |
| Concordant balanced accuracy | 0.717 | 0.724 | 0.725 |
| Mean aligned margin change | 0.35 | 0.34 | 0.37 |

Table 26: Label policy for the asked finding, report present; the first column is the paper's policy. Unmentioned findings excluded rather than negative keeps 6,997 trials and gives +0.63 [−0.17, 1.48]. Brackets: 95% patient-clustered bootstrap.

**All 14 questions.** The composite question included, 44,786 trials, for comparison with the 13-question primary.

| Quantity | MedGemma-27B | MedGemma-4B |
|---|---|---|
| Trials | 44,786 | 44,786 |
| Label-changed trials | 9,347 | 9,347 |
| Concordant accuracy | 0.648 [0.635, 0.662] | 0.673 [0.661, 0.686] |
| Concordant balanced accuracy | 0.717 [0.710, 0.725] | 0.708 [0.701, 0.714] |
| Discordant accuracy | 0.589 [0.576, 0.602] | 0.610 [0.598, 0.624] |
| Flip, label changed / unchanged | 4.30% / 4.81% | 7.28% / 7.46% |
| Flip, report present, lowercase | 4.70% [4.35, 5.06] | 7.42% |
| Flip, report absent, lowercase | 17.07% [15.78, 18.39] | 22.52% |
| Paired difference, lowercase | 12.4 [11.2, 13.6] | 15.1 [13.8, 16.3] |
| Flip, report present / absent, generated answer | 4.26% / 20.94% | |
| Paired difference, generated answer | 16.7 [15.6, 17.7] | |

Table 27: All 14 questions, the composite included. It contributes 931 of the 9,347 label changes and alone gives 2.13% against 34.10%, 32.0 points [29.0, 34.8], so including it raises the pooled effect from 15.5 to 16.7. Brackets: 95% patient-clustered bootstrap.

**Substitute resampling.** The substitute image was redrawn under two further seeds on the same 3,199 cases, reports, questions and concordant images, one-question design, report present and absent, both images. With the report the flip rate was 1.38%, 1.34% and 1.31% under seeds 0, 1 and 2, a spread of

0.06 points; without it 10.28%, 10.00% and 10.32%, a spread of 0.31 points. The paired effect of removing the report was 8.91, 8.66 and 9.00 points. A 95% interval on that effect which resamples both the patients and the substitute draw is [7.32, 10.65]. Which cases flip changes with the substitute: with the report 66 cases flip under at least one seed and 19 under all three, without it 554 and 139. The size of the report effect does not.

| Model | Label | Report | Correct, correct | Correct, wrong | Wrong, correct | Wrong, wrong | Total |
|---|---|---|---|---|---|---|---|
| MedGemma-27B | Changed | present | 297 | 5,730 | 3,215 | 105 | 9,347 |
| MedGemma-27B | Changed | absent | 1,058 | 4,014 | 3,945 | 330 | 9,347 |
| MedGemma-27B | Unchanged | present | 22,073 | 929 | 774 | 11,663 | 35,439 |
| MedGemma-27B | Unchanged | absent | 14,428 | 3,085 | 3,171 | 14,755 | 35,439 |
| MedGemma-4B | Changed | present | 520 | 5,577 | 3,090 | 160 | 9,347 |
| MedGemma-4B | Changed | absent | 1,824 | 3,566 | 3,503 | 454 | 9,347 |
| MedGemma-4B | Unchanged | present | 22,560 | 1,484 | 1,159 | 10,236 | 35,439 |
| MedGemma-4B | Unchanged | absent | 16,641 | 3,832 | 3,975 | 10,991 | 35,439 |

Table 28: Correctness on the original then the substitute image, each against its own study's label for the asked finding, all 14 questions. Concordant correct is columns one and two, discordant correct one and three, flips one and four on changed rows and two and three on unchanged rows.

Of the 9,347 label-changed trials, 61.3% [60.2, 62.4] are correct on the original image and wrong on the substitute. Conditional on being correct on the original, that share is 0.951 with the report [0.943, 0.958]. Without the report it is 0.791 [0.766, 0.816]. On the 35,439 label-unchanged trials a change of correctness is exactly a flip: 1,703 trials change correctness with the report and 6,256 without. Both accuracies of Table 3 follow from the counts: with the report, the 27B readout is correct on the original image on 29,029 of 44,786 trials (0.648) and on the substitute on 26,359 (0.589), and the four flip cells sum to 2,105. Balanced accuracy on the substitute is 0.633 [0.626, 0.640] for 27B and 0.625 for 4B. Every accuracy in this paper is scored on the asked finding's own label for the study whose image is shown, except the accuracy columns of the text-control table, which are agreement with the source study's label and are named as such there.

**Margins.** Table 29 gives the margin statistics on the same 44,786 trials. Per finding, the aligned change with and without the report: pleural effusion 0.52 and 3.94, support devices 1.12 and 1.59, edema 0.32 and 2.45, pneumothorax 0.28 and 2.38, consolidation 0.18 and 2.31, cardiomegaly 0.28 and 0.99, the composite question 0.09 and 1.39; fracture, pleural other and enlarged cardiomediastinum are near zero in both.

| Statistic | Report present | Report absent | Within-trial difference |
|---|---|---|---|
| Starting margin $m_c$ on the original image, quartiles | $-3.47$ / $-0.31$ / 8.23 | $-1.13$ / 3.03 / 7.92 | |
| Flip rate, lowest / highest quartile of $\lvert m_c \rvert$ | 15.1% / 0.15% | 37.9% / 4.3% | |
| Mean absolute paired change | 0.690 [0.657, 0.726] | 2.307 [2.203, 2.415] | 1.62 [1.53, 1.70] |
| Aligned change, label-changed trials | 0.35 [0.29, 0.41] | 1.44 [1.26, 1.62] | |
| Aligned change, no-to-yes (4,637 trials) | 0.63 [0.53, 0.74] | | |
| Aligned change, yes-to-no (4,710 trials) | 0.07 [0.05, 0.09] | | |
| Aligned change, averaged over the 14 findings | 0.246 [0.199, 0.297] | 1.227 [1.050, 1.407] | |
| Absolute change above 0.5, share of trials | 32.2% [30.8, 33.5] | 78.0% [77.0, 79.0] | |

Table 29: Margin changes on the 44,786 all-14 trials. The absolute change is larger without the report on 82.9% of trials. Section 4.1 pools the aligned change over trials; the finding-averaged row weights findings equally. Brackets: 95% patient-clustered percentile bootstrap, 2,000 draws; unbracketed cells have none.

## J REPORT CERTAINTY AND DIRECTION OF DISAGREEMENT

Two questions from a clinical reading, answered on the all-14 trials with no new model run. Report status is the raw CheXpert value of the asked finding in the source report: 1.0 certain positive, 0.0 certain negative, $-1.0$ hedged, blank not mentioned; for the composite question, positive if any real column is 1.0, hedged if any is $-1.0$ and none 1.0, negative if No Finding is 1.0. CheXpert uncertain is a proxy for hedged language, not a reading of the text. Direction is the change of the asked finding's label between the two studies; "negative" pools certain negative, hedged and not mentioned (Table 34 separates them), and "follows the substitute" means the answer moved with the image, not that it was right. Intervals use the draws of Appendix I.

A hedge does not release the image: the hedged flip rate sits within two points of the certain one and far below the same trials' report-absent rate (Table 31), and the high yes rate with and without the report is a response tendency rather than a reading of the hedge. Direction matters more. Negative-to-positive trials lose most of their agreement with the substitute to the report (Table 33), and what stops the answer moving with the image is a report that denies the finding or does not mention it, not one that hedges (Table 34).

**The mismatch question.** A separate run asked one question that names no finding, "Does the report correctly describe this image?", report present, on the report's own image and on the substitute, both models, 3,199 pairs each, greedy generation beside the logit readout. Its target is the change in the 13 finding labels between the two studies: identical (55 pairs), positive to negative only (880), negative to positive only (818) or mixed (1,446). Both models almost never flag (Table 30). Specificity on the identical pairs is 100% for both, which follows from answering yes almost everywhere rather than from calibration; sensitivity to any profile change is 1.27% [0.82, 1.78] for 27B and 0.19% [0.03, 0.41] for 4B, giving balanced accuracies of 0.506 and 0.501. What signal exists runs in the clinically important direction, 2.84 points [1.45, 4.48] for 27B, and 4B never flags a positive-to-negative or mixed pair. No generation was unscoreable.

| Model | Label change | Pairs | Flag, own image | Flag, substitute | Difference |
|---|---|---|---|---|---|
| MedGemma-27B | Any | 3,199 | 0.41% [0.20, 0.64] | 1.25% [0.80, 1.75] | +0.84 [+0.43, +1.34] |
| MedGemma-27B | None (same 13 labels) | 55 | 0.00% | 0.00% | 0.00 |
| MedGemma-27B | Positive to negative only | 880 | 0.11% [0.00, 0.39] | 0.34% [0.00, 0.82] | +0.23 [−0.22, +0.73] |
| MedGemma-27B | Negative to positive only | 818 | 0.86% [0.26, 1.55] | 3.18% [1.78, 4.81] | +2.32 [+0.97, +3.89] |
| MedGemma-27B | Mixed | 1,446 | 0.35% [0.07, 0.67] | 0.76% [0.38, 1.25] | +0.41 [+0.14, +0.78] |
| MedGemma-4B | Any | 3,199 | 0.03% [0.00, 0.10] | 0.19% [0.03, 0.40] | +0.16 [+0.03, +0.36] |
| MedGemma-4B | None (same 13 labels) | 55 | 0.00% | 0.00% | 0.00 |
| MedGemma-4B | Positive to negative only | 880 | 0.00% | 0.00% | 0.00 |
| MedGemma-4B | Negative to positive only | 818 | 0.12% [0.00, 0.41] | 0.73% [0.13, 1.54] | +0.61 [+0.11, +1.38] |
| MedGemma-4B | Mixed | 1,446 | 0.00% | 0.00% | 0.00 |

Table 30: "Does the report correctly describe this image?", report present, logit readout. Flag: the model answers no. Own image: the report's own study, where a flag is a false alarm. Brackets: 95% patient-clustered bootstrap; a row with no interval had zero flags on both sides.

| Model | Report status | Trials | Flip, report | Flip, no report | Same answer | Yes, report | Yes, no report |
|---|---|---|---|---|---|---|---|
| MedGemma-27B | Certain, positive | 9,121 | 1.75% [1.50, 2.03] | 9.38% [8.20, 10.74] | 81.6% | 83.4% | 93.7% |
| MedGemma-27B | Certain, negative | 2,438 | 1.39% [0.99, 1.85] | 23.91% [21.31, 26.50] | 54.8% | 10.7% | 54.0% |
| MedGemma-27B | Hedged | 1,279 | 2.50% [1.64, 3.46] | 18.53% [15.65, 21.66] | 82.2% | 82.1% | 86.3% |
| MedGemma-27B | Not mentioned | 31,948 | 5.88% [5.42, 6.36] | 18.68% [17.32, 20.04] | 73.3% | 40.5% | 60.1% |
| MedGemma-4B | Certain, positive | 9,121 | 2.88% [2.49, 3.30] | 21.38% [19.14, 23.67] | 78.1% | 76.7% | 82.5% |
| MedGemma-4B | Certain, negative | 2,438 | 4.76% [3.89, 5.65] | 23.46% [20.79, 26.23] | 74.2% | 16.4% | 23.3% |
| MedGemma-4B | Hedged | 1,279 | 5.00% [3.81, 6.21] | 30.02% [27.15, 32.96] | 74.4% | 74.2% | 69.0% |
| MedGemma-4B | Not mentioned | 31,948 | 9.01% [8.33, 9.66] | 22.47% [21.27, 23.70] | 77.8% | 35.0% | 49.7% |

Table 31: Image sensitivity by the report's certainty on the asked finding, all 14 questions. Status is the raw CheXpert value in the source report. Same answer: the answer on the original image with and without the report. Yes: share answered yes on the original image. Brackets: 95% patient-clustered bootstrap.

**Hedging read from the text.** CheXpert uncertain is a proxy, so the split was repeated on the report text with a lexicon of 49 hedging terms supplied by a clinical reviewer, graded weak negation, possibility, probable and differential; comparative and technical terms are not hedges. A trial is hedged when a sentence mentioning the asked finding matches a graded term, found with the per-label keyword patterns of the text controls; the composite question is excluded. The text-derived group is much larger than CheXpert's uncertain class, 7,414 trials against 1,279, and overlaps it little: 10.7% is CheXpert-uncertain and 39.9% CheXpert-positive, since "likely atelectasis" is graded probable and labelled positive. The conclusion does not move (Table 32), and excluding trials whose only hedge is a threshold descriptor changes nothing.

| Model | Text group | Trials | Flip, report | Flip, no report | Difference | Same answer | Yes, report | Yes, no report |
|---|---|---|---|---|---|---|---|---|
| MedGemma-27B | Mentioned, no hedge | 12,069 | 1.36% [1.14, 1.58] | 19.55% [17.98, 21.21] | +18.19 [+16.64, +19.86] | 80.74% [79.70, 81.81] | 53.58% [50.36, 56.44] | 62.33% [58.97, 65.32] |
| MedGemma-27B | Hedged | 7,414 | 1.61% [1.30, 1.95] | 17.98% [16.47, 19.57] | +16.37 [+14.85, +17.97] | 82.94% [81.65, 84.13] | 70.70% [67.24, 73.93] | 73.98% [71.10, 76.68] |
| MedGemma-27B | Hedged, thresholds excluded | 6,012 | 1.68% [1.35, 2.07] | 18.23% [16.63, 19.92] | +16.55 [+14.88, +18.21] | 82.02% [80.61, 83.30] | 65.88% [62.08, 69.60] | 71.29% [68.12, 74.37] |
| MedGemma-27B | Not mentioned | 22,104 | 7.72% [7.11, 8.36] | 17.43% [16.02, 18.78] | +9.71 [+8.46, +11.01] | 70.91% [69.61, 72.22] | 38.40% [36.21, 40.37] | 63.34% [61.54, 65.04] |
| MedGemma-4B | Mentioned, no hedge | 12,069 | 2.96% [2.57, 3.34] | 18.76% [17.44, 20.12] | +15.80 [+14.47, +17.22] | 83.41% [82.46, 84.29] | 54.49% [51.34, 57.40] | 57.13% [53.80, 60.21] |
| MedGemma-4B | Hedged | 7,414 | 3.16% [2.70, 3.67] | 22.54% [20.91, 24.25] | +19.38 [+17.85, +21.08] | 82.74% [81.49, 83.94] | 69.05% [65.55, 72.25] | 67.26% [63.72, 70.53] |
| MedGemma-4B | Hedged, thresholds excluded | 6,012 | 3.18% [2.69, 3.68] | 22.44% [20.81, 24.26] | +19.26 [+17.64, +21.06] | 82.72% [81.36, 83.98] | 64.47% [60.40, 68.33] | 64.82% [60.88, 68.63] |
| MedGemma-4B | Not mentioned | 22,104 | 11.36% [10.54, 12.13] | 23.52% [22.14, 24.90] | +12.17 [+10.92, +13.41] | 75.80% [74.86, 76.76] | 30.09% [28.18, 31.94] | 49.11% [46.87, 51.30] |

Table 32: Image sensitivity by the report's wording, read with a clinician-supplied lexicon, 13 finding-specific questions. Hedged: a sentence mentioning the finding matches a graded term. Columns as in Table 31. Brackets: 95% patient-clustered bootstrap.

| Model | Direction | Report | Trials | Flip | Follows substitute | Follows report |
|---|---|---|---|---|---|---|
| MedGemma-27B | Positive to negative | present | 4,710 | 1.53% [1.17, 1.91] | 12.72% [11.26, 14.54] | 87.28% |
| MedGemma-27B | Positive to negative | absent | 4,710 | 14.46% [12.84, 16.25] | 17.18% [15.34, 19.17] | 82.82% |
| MedGemma-27B | Negative to positive | present | 4,637 | 7.12% [6.18, 8.12] | 62.82% [59.81, 65.62] | 37.18% |
| MedGemma-27B | Negative to positive | absent | 4,637 | 15.25% [13.49, 17.12] | 90.45% [88.78, 91.92] | 9.55% |
| MedGemma-4B | Positive to negative | present | 4,710 | 1.97% [1.48, 2.54] | 15.58% [13.83, 17.70] | 84.42% |
| MedGemma-4B | Positive to negative | absent | 4,710 | 24.48% [22.21, 26.91] | 33.48% [30.53, 36.72] | 66.52% |
| MedGemma-4B | Negative to positive | present | 4,637 | 12.66% [11.44, 13.84] | 62.02% [59.23, 64.59] | 37.98% |
| MedGemma-4B | Negative to positive | absent | 4,637 | 24.26% [21.90, 26.85] | 80.87% [77.77, 83.64] | 19.13% |

Table 33: Direction of the label change on the 9,347 label-changed trials, all 14 questions. Follows substitute: the answer on the substitute image matches that study's label; follows report: it matches the report's. Brackets: 95% patient-clustered bootstrap.

| Model | Report says | Report | Trials | Flip | Follows substitute |
|---|---|---|---|---|---|
| MedGemma-27B | Certain negative | present | 671 | 2.83% [1.58, 4.33] | 18.93% [15.00, 22.88] |
| MedGemma-27B | Certain negative | absent | 671 | 15.80% [12.31, 19.66] | 93.74% [91.14, 95.95] |
| MedGemma-27B | Hedged | present | 258 | 1.55% [0.36, 3.23] | 91.09% [87.20, 94.42] |
| MedGemma-27B | Hedged | absent | 258 | 8.91% [4.96, 13.55] | 94.57% [91.83, 97.13] |
| MedGemma-27B | Not mentioned | present | 3,708 | 8.28% [7.13, 9.46] | 68.80% [66.12, 71.20] |
| MedGemma-27B | Not mentioned | absent | 3,708 | 15.59% [13.77, 17.63] | 89.56% [87.61, 91.22] |
| MedGemma-4B | Certain negative | present | 671 | 9.69% [7.56, 12.12] | 28.32% [24.58, 32.40] |
| MedGemma-4B | Certain negative | absent | 671 | 37.11% [31.32, 43.11] | 57.68% [49.91, 65.38] |
| MedGemma-4B | Hedged | present | 258 | 4.65% [2.41, 7.17] | 85.66% [81.30, 89.78] |
| MedGemma-4B | Hedged | absent | 258 | 22.48% [17.17, 27.97] | 86.43% [81.57, 90.57] |
| MedGemma-4B | Not mentioned | present | 3,708 | 13.75% [12.34, 15.16] | 66.48% [63.65, 68.99] |
| MedGemma-4B | Not mentioned | absent | 3,708 | 22.06% [19.78, 24.59] | 84.68% [82.31, 86.81] |

Table 34: The negative-to-positive direction (report negative, substitute study positive) split by what the report says about the asked finding, all 14 questions. Follows substitute: share answered yes on the substitute image. Brackets: 95% patient-clustered percentile bootstrap.

## K TEXT-CONTROL HIERARCHY

One-question design, full arm, both images, MedGemma-27B. Differences are paired against the plain full report on each condition's cases. The full-report row is a re-run of the one-question full arm

on the same 3,199 cases: it gives 1.38% (44 flips) against the 1.41% (45 flips) of Table 1, a one-case difference between two runs of the same arm. Removing only the sentences that name the asked finding is not separated from removing the whole report (+1.3 points [−0.3, 2.7]). On the 1,702 cases carrying both sections, the findings section alone is within 0.3 points of the full report [−0.3, 0.9]. The impression alone is 4.4 points above it [3.0, 5.7]. Text-block order on the same design: with the text first the flip rate is 2.84% against 1.41% image first (Table 1); the per-arm counts are in Table 55.

| Text shown with the image | Flip rate | vs. full report, points | Source-label agreement, original | Source-label agreement, substitute |
|---|---|---|---|---|
| Full report | 1.38% [0.96, 1.87] | reference | 0.931 [0.919, 0.941] | 0.930 [0.919, 0.940] |
| Findings section only (2,507 cases) | 1.44% [1.01, 1.89] | +0.2 [−0.2, 0.6] | 0.927 [0.915, 0.939] | 0.925 [0.913, 0.937] |
| Target sentences replaced by the opposite label | 2.53% [1.87, 3.18] | +1.2 [0.3, 1.9] | 0.043 [0.034, 0.053] | 0.046 [0.037, 0.056] |
| One sentence stating the label | 3.13% [2.42, 3.90] | +1.8 [0.9, 2.6] | 0.845 [0.819, 0.868] | 0.848 [0.822, 0.871] |
| Impression section only (2,394 cases) | 4.26% [3.34, 5.31] | +3.1 [2.2, 4.1] | 0.873 [0.853, 0.891] | 0.866 [0.847, 0.883] |
| Clinical boilerplate, length matched | 7.06% [5.90, 8.38] | +5.7 [4.5, 7.0] | 0.787 [0.759, 0.813] | 0.764 [0.736, 0.790] |
| Neutral prose, length matched | 8.22% [6.81, 9.74] | +6.9 [5.4, 8.4] | 0.799 [0.771, 0.823] | 0.767 [0.737, 0.794] |
| Report removed | 10.28% [8.72, 12.02] | +8.9 [7.5, 10.6] | 0.773 [0.744, 0.799] | 0.746 |
| Target sentences removed | 11.60% [10.35, 12.93] | +10.2 [9.0, 11.6] | 0.743 [0.717, 0.767] | 0.726 [0.699, 0.751] |

Table 35: Text-control hierarchy, one-question design, 3,199 cases unless stated; accuracy is scored against the source study's label. Sentences naming the asked finding were found by keyword lists (55 cases had none); section variants keep cases whose report has that section. Brackets: 95% patient-clustered percentile bootstrap, 2,000 draws; unbracketed cells have none.

## L READOUT VALIDATION

The one-question design rerun with "Answer with exactly Yes or No." appended, eight tokens generated per case, both report conditions and both images; the all-14 design was then rerun the same way (Table 2). Three readouts are compared: the paper's lowercase first-token argmax, the token-family argmax over the yes and no variants, and the parsed generated answer. Answer order is tested in Appendix P.

| Quantity | Value |
|---|---|
| Format compliance, one-question design | 100% |
| Format compliance, all-14 design | at least 0.9999 in every condition |
| Without the instruction | no answers unscoreable within eight tokens |
| Readout tokens | lowercase yes (4443), no (1904) |
| Variant token ids, family readout only | Yes 10784, 11262, 8438; No 3771, 951, 2301 |
| Mean first-position probability of the read tokens | below $10^{-4}$ in every condition |
| Lowercase agrees with generated, report present | 98.8% |
| Lowercase agrees with generated, report absent | 91 to 93% |
| Instruction effect on the report-absent rate, lowercase | +8.0 points |
| Same effect, token-family readout | +0.3 [−1.4, 2.0] |

Table 36: Readout validation. The two read tokens are the unique one-token continuations of the generation prompt. Without the instruction the model answers "Yes" only when the label is yes, so the report-absent rate is a property of the readout choice.

| Prompt | Arm | Lowercase first token | Token families | Generated answer |
|---|---|---|---|---|
| With instruction | report present | 1.22% [0.81, 1.67] | 0.84% [0.50, 1.27] | 0.84% [0.50, 1.27] |
| With instruction | report absent | 18.26% [16.20, 20.57] | 17.69% [15.88, 19.61] | 17.66% [15.90, 19.55] |
| With instruction | paired difference | −17.0 [−19.4, −14.9] | −16.9 [−18.8, −15.0] | −16.8 [−18.7, −15.0] |
| No instruction | report present | 1.38% [0.96, 1.87] | 1.09% [0.69, 1.54] | 1.84% [1.36, 2.37] |
| No instruction | report absent | 10.28% [8.72, 12.02] | 17.35% [15.31, 19.56] | 11.57% [9.99, 13.32] |
| No instruction | paired difference | −8.9 [−10.6, −7.5] | −16.3 [−18.5, −14.2] | −9.7 [−11.5, −8.1] |

Table 37: Flip rates of the one-question design under three readouts and two prompts, 3,199 cases; paired difference is report present minus absent, in points. No-instruction generated answers are three-state (yes, no, unscoreable). Brackets: 95% patient-clustered percentile bootstrap, 2,000 draws.

**The 14-question design under the instruction.** All 14 questions, 44,786 trials on the same 3,199 cases and 293 patients, rerun with the instruction and greedy generation, so the same trials can be scored three ways (Table 2). The effect survives on its own unit under every readout and every interval on the difference excludes zero.

| Readout | Paired difference | Ratio |
|---|---|---|
| Lowercase first token | 16.1 points | 4.5 to 4.9 |
| Token families | 16.7 points | 4.5 to 4.9 |
| Generated answer | 16.7 points | 4.5 to 4.9 |
| Headline prompt, no instruction | 12.37 points | |

Table 38: All 14 questions under the answer instruction. The headline prompt carries no instruction and gives 4.70% against 17.07% on the same trials, so the instructed prompt checks the readout rather than replacing the headline.

Table 2, the all-14 design under the instruction, is in the main text (Section 4.1).

| Prompt | Readout | Report present | Report absent | Paired difference, points |
|---|---|---|---|---|
| With instruction | lowercase first token | 1.22% | 18.26% | −17.0 [−19.4, −14.9] |
| With instruction | token families | 0.84% | 17.69% | −16.9 [−18.8, −15.0] |
| With instruction | generated answer | 0.84% | 17.66% | −16.8 [−18.7, −15.0] |
| No instruction | lowercase first token | 1.38% | 10.28% | −8.9 [−10.6, −7.5] |
| No instruction | token families | 1.09% | 17.35% | −16.3 [−18.5, −14.2] |
| No instruction | generated answer | 1.84% | 11.57% | −9.7 [−11.5, −8.1] |

Table 39: Readout validation, secondary one-question design (3,199 cases, 293 patients): flip rate under three readouts and two prompts, present minus absent in points. Per-cell intervals in Appendix Table 37; the all-14 design under the instruction is in Appendix Table 2. Brackets: 95% patient-clustered percentile bootstrap, 2,000 draws.

## M MODALITY ABLATION ON BOTH DESIGNS

| Finding | Full | Image + question | Report + question | Question only |
|---|---|---|---|---|
| Atelectasis | 0.623 [0.603, 0.646] | 0.500 [0.500, 0.500] | 0.823 [0.807, 0.839] | 0.500 [0.500, 0.500] |
| Cardiomegaly | 0.766 [0.743, 0.788] | 0.656 [0.630, 0.682] | 0.808 [0.789, 0.827] | 0.500 [0.500, 0.500] |
| Consolidation | 0.834 [0.809, 0.858] | 0.645 [0.609, 0.685] | 0.887 [0.861, 0.909] | 0.500 [0.500, 0.500] |
| Edema | 0.849 [0.829, 0.868] | 0.658 [0.637, 0.681] | 0.882 [0.867, 0.897] | 0.500 [0.500, 0.500] |
| Enlarged cardio-mediastinum | 0.480 [0.429, 0.530] | 0.482 [0.432, 0.533] | 0.511 [0.469, 0.559] | 0.500 [0.500, 0.500] |
| Fracture | 0.938 [0.903, 0.965] | 0.586 [0.532, 0.642] | 0.906 [0.857, 0.947] | 0.500 [0.500, 0.500] |
| Lung lesion | 0.696 [0.666, 0.725] | 0.499 [0.483, 0.512] | 0.772 [0.737, 0.804] | 0.500 [0.500, 0.500] |
| Lung opacity | 0.601 [0.583, 0.621] | 0.503 [0.501, 0.505] | 0.688 [0.668, 0.707] | 0.500 [0.500, 0.500] |
| Pleural effusion | 0.833 [0.811, 0.854] | 0.740 [0.712, 0.768] | 0.911 [0.897, 0.924] | 0.500 [0.500, 0.500] |
| Pleural other | 0.608 [0.492, 0.717] | 0.529 [0.428, 0.637] | 0.580 [0.475, 0.694] | 0.500 [0.500, 0.500] |
| Pneumonia | 0.714 [0.684, 0.742] | 0.537 [0.505, 0.569] | 0.710 [0.677, 0.742] | 0.500 [0.500, 0.500] |
| Pneumothorax | 0.853 [0.781, 0.903] | 0.711 [0.663, 0.758] | 0.830 [0.749, 0.889] | 0.500 [0.500, 0.500] |
| Support devices | 0.535 [0.525, 0.547] | 0.500 [0.500, 0.500] | 0.784 [0.757, 0.808] | 0.500 [0.500, 0.500] |
| Any finding (composite) | 0.732 [0.712, 0.750] | 0.524 [0.510, 0.543] | 0.696 [0.680, 0.712] | 0.500 [0.500, 0.500] |
| Pooled balanced accuracy | 0.717 [0.710, 0.725] | 0.665 [0.658, 0.672] | 0.770 [0.764, 0.776] | 0.500 [0.500, 0.500] |
| Macro-averaged balanced accuracy | 0.719 [0.706, 0.732] | 0.576 [0.564, 0.588] | 0.770 [0.759, 0.782] | 0.500 [0.500, 0.500] |
| Pooled micro accuracy | 0.648 [0.635, 0.662] | 0.504 [0.493, 0.516] | 0.772 [0.764, 0.781] | 0.204 [0.194, 0.213] |
| Sensitivity | 0.834 [0.817, 0.847] | 0.937 [0.926, 0.945] | 0.767 [0.751, 0.781] | 1.000 [1.000, 1.000] |
| Specificity | 0.601 [0.580, 0.623] | 0.394 [0.374, 0.415] | 0.773 [0.760, 0.787] | 0.000 [0.000, 0.000] |

Table 40: Balanced accuracy per finding, pooled and macro-averaged, four modality cells, all-14 design (44,786 trials, report-derived labels, majority 0.796). Interaction, image effect with the report minus without: $-0.218$ [$-0.224$, $-0.212$] pooled over trials. Macro-averaged over findings: $-0.128$ [$-0.141$, $-0.115$]. Brackets: 95% patient-clustered percentile bootstrap, 2,000 draws.

| | Full | | Image + question | | Report + question | |
|---|---|---|---|---|---|---|
| Finding | Sens. | Spec. | Sens. | Spec. | Sens. | Spec. |
| Atelectasis | 0.999 | 0.247 | 1.000 | 0.000 | 0.993 | 0.652 |
| Cardiomegaly | 0.893 | 0.639 | 0.879 | 0.432 | 0.833 | 0.783 |
| Consolidation | 0.957 | 0.711 | 0.791 | 0.499 | 0.909 | 0.865 |
| Edema | 0.968 | 0.730 | 0.956 | 0.361 | 0.958 | 0.806 |
| Enlarged cardiomediastinum | 0.406 | 0.554 | 0.594 | 0.370 | 0.219 | 0.804 |
| Fracture | 0.925 | 0.950 | 0.280 | 0.893 | 0.839 | 0.972 |
| Lung lesion | 0.952 | 0.441 | 0.976 | 0.022 | 0.911 | 0.633 |
| Lung opacity | 0.991 | 0.212 | 1.000 | 0.005 | 0.969 | 0.407 |
| Pleural effusion | 0.990 | 0.675 | 0.908 | 0.572 | 0.988 | 0.833 |
| Pleural other | 0.694 | 0.521 | 0.444 | 0.613 | 0.486 | 0.673 |
| Pneumonia | 0.679 | 0.749 | 0.746 | 0.327 | 0.560 | 0.861 |
| Pneumothorax | 0.723 | 0.983 | 0.607 | 0.814 | 0.670 | 0.991 |
| Support devices | 0.994 | 0.077 | 1.000 | 0.000 | 0.973 | 0.594 |
| Any finding (composite) | 0.589 | 0.876 | 0.994 | 0.055 | 0.444 | 0.947 |
| Pooled | 0.834 | 0.601 | 0.937 | 0.394 | 0.767 | 0.773 |

Table 41: Sensitivity and specificity per finding, three modality cells, all-14 design (44,786 trials, report-derived labels). Question only is omitted: it answers yes on every trial, so its sensitivity is 1.000 and its specificity 0.000 for every finding. These are the decomposition of the balanced accuracies in Table 40, which is their mean, and they are point estimates from the committed per-finding summary with no intervals computed. The decomposition is what the balanced accuracy hides: on the full arm, atelectasis, lung opacity and support devices sit at 0.999, 0.991 and 0.994 sensitivity against 0.247, 0.212 and 0.077 specificity, so a balanced accuracy near 0.6 for those findings is a near-constant yes rather than a discrimination. Removing the image raises specificity on all 14 findings, which is the same ordering the pooled row shows.

| Finding | Report present | Report absent | Absent minus present |
|---|---|---|---|
| Atelectasis | 6.81% | 4.63% | −2.2 |
| Cardiomegaly | 6.06% | 26.23% | +20.2 |
| Consolidation | 3.84% | 30.73% | +26.9 |
| Edema | 2.50% | 28.63% | +26.1 |
| Enlarged cardiomediastinum | 4.78% | 24.29% | +19.5 |
| Fracture | 0.78% | 7.00% | +6.2 |
| Lung lesion | 6.19% | 26.95% | +20.8 |
| Lung opacity | 5.10% | 12.38% | +7.3 |
| Pleural effusion | 5.22% | 29.60% | +24.4 |
| Pleural other | 2.31% | 21.76% | +19.5 |
| Pneumonia | 2.50% | 33.17% | +30.7 |
| Pneumothorax | 0.31% | 12.32% | +12.0 |
| Support devices | 11.03% | 1.44% | −9.6 |
| Any finding (composite) | 2.13% | 34.10% | +32.0 |
| Pooled, 14 questions | 4.26% | 20.94% | +16.7 [15.6, 17.7] |
| Pooled, 13 finding-specific (primary) | 4.43% | 19.93% | +15.5 [14.5, 16.5] |

Table 42: Image-swap flip rate of the generated answer per finding under the explicit answer instruction, all-14 design, 3,199 cases per finding. Two findings reverse the pooled direction: atelectasis and support devices flip less often without the report than with it. The pooled increase is therefore a positive pooled effect, not a consistent effect across findings. Intervals on the pooled rows are patient-clustered, 10,000 draws; per-finding rows are point estimates from the committed per-trial record.

Table 43 gives the same four cells on the secondary one-question design.

| Arm | Report | Image | Accuracy | Balanced acc. | Δ vs. full, paired |
|---|---|---|---|---|---|
| Full (reference) | ✓ | ✓ | 0.931 [0.919, 0.941] | 0.908 | reference |
| Report + question | ✓ | × | 0.913 [0.900, 0.925] | 0.927 | −0.018 [−0.026, −0.009] |
| Question only | × | × | 0.846 [0.822, 0.869] | 0.500 | −0.084 [−0.109, −0.062] |
| Image + question | × | ✓ | 0.773 [0.744, 0.799] | 0.479 | −0.158 [−0.184, −0.134] |

Table 43: Modality ablation, one question per case (3,199 cases, 293 patients), question always present. Target `yes` on 2,707 cases and `no` on the 492 asked the composite question, so the majority rate 0.846 equals question-only accuracy. Both accuracy columns are on the original image, scored against its own study's label; the MedGemma-4B counterparts are in Appendix G. Brackets: 95% patient-clustered percentile bootstrap, 10,000 draws; unbracketed cells have none.

## N CLASSIFIER-SURROGATE IMAGE LABELS

A surrogate for image-level labels, not a reading. A DenseNet-121 from TorchXRayVision (Cohen et al., 2022), trained on CheXpert only (Irvin et al., 2019) and so on neither these images nor these reports, scored all 3,300 shown images. Eleven heads map one to one onto the finding questions; the composite question takes their maximum. Pleural other and support devices have no head, leaving twelve questions per case. Thresholds are applied to the library's operating-point-normalised score, on which each head's CheXpert-fitted operating point sits at 0.5, and a trial is excluded when either image falls within 0.1 of 0.5. The rule was fixed before any flip was examined.

| Quantity | Value |
|---|---|
| Images scored | 3,300 |
| Questions per case | 12 (of 14) |
| Trials | 38,388 |
| Predicted label change | 2,249 |
| Predicted label preserved | 22,469 |
| Excluded band | 13,670 |
| AUROC against report labels, eleven heads | 0.53 to 0.74 |
| AUROC, derived any-finding score | 0.76 |
| Called positive at the 0.5 rule | 18% to 92% |
| Report-label prevalence | 3% to 32% |
| Raw-probability rule: trials retained | 29,147 |
| Flip, report present, change / preserved | 7.44% / 3.64% |
| Flip, report absent, change / preserved | 31.24% / 13.82% |

Table 44: Classifier surrogate: setup, agreement with the report labels, and the robustness rule. The raw-probability rule thresholds the sigmoid directly and leaves every AUROC unchanged. Ten of eleven heads call more than half of images positive, so the surrogate over-calls.

| Report and classifier-predicted subset | Flip rate | Agrees with substitute label | Paired report difference, points |
|---|---|---|---|
| Present, label change (2,249) | 7.20% [6.03, 8.57] | 0.422 [0.395, 0.451] | reference |
| Present, label preservation (22,469) | 4.44% [3.99, 4.87] | 0.697 [0.678, 0.715] | reference |
| Absent, label change | 37.35% [32.63, 41.87] | 0.659 [0.624, 0.689] | +30.1 [25.3, 34.6] |
| Absent, label preservation | 12.06% [10.98, 13.17] | 0.846 [0.831, 0.861] | +7.6 [6.6, 8.7] |
| Absent, change supported by classifier and report labels (396) | 51.01% [41.35, 58.41] | 0.745 [0.686, 0.790] | +44.9 [36.1, 52.2] |

Table 45: Flip rates where the classifier surrogate predicts a change in the queried finding or the same label for both images; last row: change supported by both classifier and report labels. Brackets: 95% patient-clustered percentile bootstrap, 2,000 draws; 293 patients on the classifier-only rows, 286 on the last.

Where the surrogate predicts a change in the queried finding, the readout without the report flips three times as often as where it predicts the same label for both images; the report-derived split shows no such separation on the same findings (4.55% against 4.69% with the report, 16.91% against 17.68% without).

## O A SECOND QUESTION TEMPLATE

The one-question design re-run with "Is <finding> present on this chest X-ray?" in place of "Does this chest X-ray show <finding>?", report present and absent, both images, no answer instruction, same cases.

| Quantity | Default template | Second template |
|---|---|---|
| Flip, report present | 1.38% | 1.06% [0.70, 1.42] |
| Flip, report absent | 10.28% | 8.28% [6.55, 10.15] |
| Report-presence difference | $-8.9$ [$-10.6$, $-7.5$] | $-7.2$ [$-9.2$, $-5.5$] |
| Balanced accuracy, report absent | 0.479 | 0.621 |
| Specificity, report absent | 0.06 | 0.29 |
| Balanced accuracy, report present | 0.908 | 0.908 |

Table 46: Second question template against the default on the same cases. The composite question becomes "Is any acute cardiopulmonary finding present on this chest X-ray?". The template-by-report interaction is $+1.7$ points [$-0.0$, $+3.4$]. Brackets: 95% intervals.

## P ANSWER ORDER IN THE INSTRUCTION

The four passes carrying the answer instruction, repeated with the options reversed from "Answer with exactly Yes or No." to "Answer with exactly No or Yes.", paired case by case on the same 3,199 cases.

| Quantity | Default | Reversed | Difference |
|---|---|---|---|
| Flip, report present | 1.22% | 0.97% | $-0.25$ [$-0.60$, 0.09] |
| Flip, report absent | 18.26% | 18.54% | $+0.28$ [$-0.42$, 0.99] |
| Report-presence difference | $-17.04$ | $-17.57$ | $-0.53$ [$-1.36$, 0.30] |
| Flip, report absent, generated answer | | | $+1.22$ [0.51, 1.90] |
| Flip, report absent, token families | | | $+1.09$ |

Table 47: Answer order under the paper's readout unless stated. Reversing it changed the answer on 0.59% to 1.78% of cases, a small shift toward the option listed first, an order of magnitude below the report effect. Brackets: 95% intervals.

**An explicit abstention option.** The same cases rerun with Yes, No or Uncertain allowed.

| Quantity | Value |
|---|---|
| Uncertain rate, report present | 1.5% [1.0, 2.0] |
| Uncertain rate, report absent | 0% |
| Change in that rate under substitution | $+0.06$ [$-0.10$, 0.22] |
| Flip, report present | 1.1% |
| Flip, report absent | 18.3% |
| Paired difference | 17.2 [15.2, 19.3] |
| Same difference, forced yes or no | 16.8 [15.0, 18.7] |

Table 48: Abstention option. A change to or from Uncertain counts as a change. The model abstains only when a report is present and at the same rate whether or not the image agrees, so the option neither weakens nor explains the report effect.

# Q LAYER-WISE READOUTS AND ATTENTION INTERVENTIONS

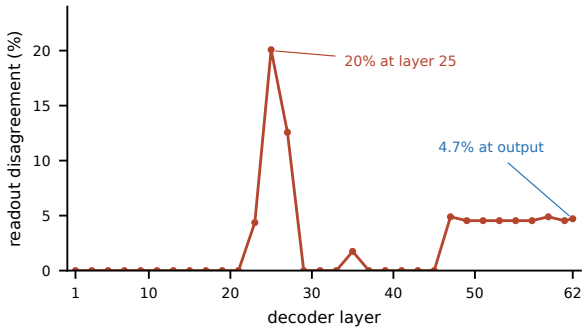



Figure 5: Uncalibrated logit-lens projection: share of 44,786 trials, 14 findings pooled, whose projected yes/no readout differs between the two images, by layer. Below layer 46 (shaded) the projection does not clear the best constant predictor of the model's output. The layer-25 peak, 20.0% with the report against 0.7% without, is a property of the report-present prompt rather than of the network and supports no claim.

## Q.1 IMAGE SENSITIVITY ACROSS DECODER LAYERS

We next examined how image substitutions affect projected answer scores across decoder layers, with and without the report. We evaluated these trajectories alongside agreement with the final answer to assess where the intermediate readout was informative. We then used attention interventions to test the contribution of direct access to report-token positions (Figure 6; full analysis in Appendix Q).

At each layer we project the residual stream through the output head and take the yes-minus-no margin, then average the absolute change in that margin between the two images over the 44,786 trials (Figure 6A). The change is small through the middle of the network and rises from about layer 46: with the report it is 0.04 at layer 25 and 1.31 at layer 46, ending at the published 0.690 at the output; without the report it is 0.05 and 4.54, ending at 2.307. The report-absent trajectory sits above the report-present one at every layer from 46 onward, which is the layerwise form of the headline effect. Figure 6B shows how far each layer's projected answer agrees with the final answer: agreement is 0.55 with the report and 0.67 without at layer 25, against a majority-class floor, and it reaches 0.95 and 0.93 at layer 46. The intermediate readout carries a limited, non-persistent signal before layer 46 and tracks the final answer after it, so the layer-25 feature reported in Appendix Q is retained as an observation about the projection and not read as the point at which the answer is decided. Figure 6C shows the attention intervention against its two controls; the trajectory and the intervention answer different questions, one describing projected responses across depth and the other the effect of interrupting a specified pathway, and Section Q.2 states what the intervention does and does not establish.

## Q.2 AT WHAT DEPTH DOES THE REPORT TAKE HOLD?

Cumulative blocking of attention to report-token positions increased image-swap sensitivity when initiated at early decoder layers, and the effect diminished for interventions initiated later. These results identify when direct access to report-token positions affects the measured response; they do not localise all downstream use of report information, since report-derived content may already have been copied into other positions before a later block starts. The estimates: masking the text stream's attention to the report raises the image-swap flip rate from 1.50% to 7.75% when the block begins at layer 0, against 12.00% with no report, and is at baseline from layer 32. Single-layer blocking never exceeds 2.50%, which is consistent with redundant or distributed pathways and does not show that no layer carries the effect. The 14.00% at layer 22 overshoots the no-report rate and falls in the window where the question-and-instruction control also moves, so the report-specific comparison is confined to layers 0 to 20, where the paired difference clears zero against both controls and the two arms flip disjoint case sets. A higher flip rate under the block is not by itself a restored reading of the substituted image. Cross-condition residual patching adds no localisation, being a null

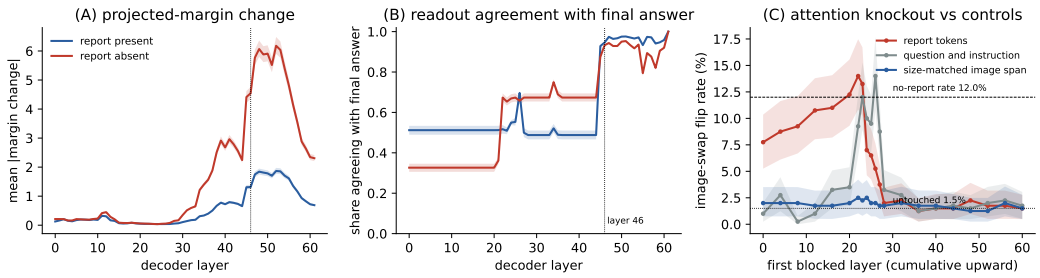



Figure 6: **(A)** Mean absolute change in the projected yes/no margin between the two images at each decoder layer, report present and absent, 44,786 trials, patient-clustered 95% intervals; the layer-61 values are the published 0.690 and 2.307. **(B)** Share of trials whose projected answer at each layer equals the final answer. **(C)** Report-token attention knockout, cumulative from the plotted layer upward, against the question-and-instruction span and a size-matched image span, 400 one-question cases.

against its wrong-donor control, 14.07% against a correct-donor maximum of 12.25% (Appendix Q, Figure 7). This is a MIMIC-CXR result: the image-token knockout run on three further datasets does not separate from its control on VQA-RAD or OmniMedVQA, and the control comparison was unavailable for SLAKE, so it is reported as a negative and not generalised.

This appendix asks how the measured response to an image substitution evolves across decoder depth and what an attention intervention adds. Four approaches were evaluated. Reading the answer off each layer fails for a measurable reason: the projection does not track the model's own answer until layer 46 of 62. The tuned lens scored below the uncalibrated lens on both readout datasets. The settling depth gives one-question medians of 26 with the report and 23 without, but a shuffled-pairing null reproduces the same per-case difference. A per-layer probe (Alain & Bengio, 2016), trained on the residual stream to predict whether the swap changes the answer, transfers to chance on a disjoint cohort. Attention knockout is the one intervention that separates from its controls: blocking the report's influence restores image sensitivity when the block starts shallow and does nothing when it starts deep, which bounds when access to report-token positions matters without localising report information that has already left them. Cross-condition patching adds nothing. Nothing in the paper rests on the layerwise projection.

### Q.3 The projection and its calibration

Both analyses are exploratory and negative; neither supports a mechanistic claim. The layerwise logit-lens projection (nostalgebraist, 2020), the fraction of report-present trials whose projected readout differs between the two runs at each of the 62 layers, describes the projection only (Figure 5); no accuracy, flip rate or margin change depends on it, and its output-layer value is the flip rate of the headline. A calibration pass asked at each layer whether the projected yes/no agrees with the model's own forced-choice output, balanced over the two classes and scored against the best constant predictor. Table 49 gives the result: the projection clears both baselines only from layer 46, the layer-25 feature is a property of the report-present prompt, and the settling-depth statistic carries nothing about the pairing. No projected claim in this paper is anchored to a depth below layer 46; the only depth statement below it is the causal bound from the attention knockout, which does not use the lens.

| Quantity | Value |
|---|---|
| Shallowest layer clearing both baselines | 46 of 62 [46, 46] |
| Balanced agreement, layer 46, present / absent | 0.909 / 0.760 |
| Balanced agreement at layer 25 | 0.648 [0.628, 0.668] |
| Balanced agreement, layers 28 to 45 | within 0.002 of 0.500 |
| Layer-25 disagreement, present / absent | 0.2000 / 0.0067 |
| Difference | 19.3 points [17.2, 21.3] |
| Settling depth, one-question medians, report present / absent | 26 / 23 |
| Paired mean difference | 1.56 layers [0.88, 2.21] |
| Same per-case median under a shuffled-pairing null | 3.0 |
| Still apart vs flips, image first / text first | 48 vs 6 / 122 vs 33 |
| Curve endpoint, report present / absent | 0.0470 [0.044, 0.051] / 0.1707 [0.158, 0.183] |

Table 49: Calibration of the layerwise projection. Still apart: projected distributions at the final layer differ by more than 0.05 in total variation whichever answer wins, so it exceeds the flip count (Appendix Tables 64 and 55). Brackets: 95% patient-clustered bootstrap.

### Q.4 Report-token attention knockout

**Causal depth: report-token attention knockout and cross-condition patching.** Two interventions measure the effect of access to report-token representations, MedGemma-27B, one-question design, 400 cases and 282 patients, 300 of them decisive, at 21 of 62 layers, with a 2,000-draw patient-clustered bootstrap. The first masks attention at decoder layer $L$ so that chosen query positions cannot read the report's key columns, cumulatively from $L$ upward or at $L$ alone, with the query scope either the answer position or every non-image position outside the blocked span. The second patches the report-absent residual into the report-present run against a wrong-donor control.

Table 50 gives the numbers; Figure 7 the curve. Cumulative blocking restores image-swap sensitivity when it begins at layers 0 to 20 and not from layer 28 upward, so the boundary sits near layer 24; no single layer carries the effect; the report-specific claim is confined to layers 0 to 20 by three readings, the paired difference against both controls, disjoint flipped cases against the controls, and a destination near the measured no-report operating point rather than a destroyed one.

Two limits bound the reading. The instrument has a ceiling, since the answer position still reads the report directly and the report's tokens still shift every later position, so its extent should not be read as an effect size, and the peak over-restores with an unchanged destination, added disruption rather than more removal. Patching is a null and adds no localisation. Content copied out of the report positions at an early layer is untouched by a later block, and a higher flip rate is not by itself a restored reading of the substituted image.

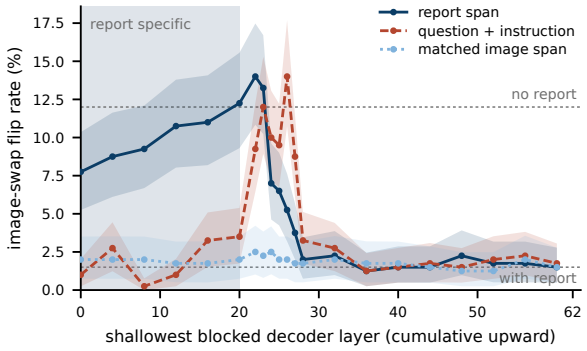



Figure 7: Report-token attention knockout, cumulative from the plotted layer upward, wider query scope, 400 one-question cases. Blocking the report restores image sensitivity toward the no-report rate only when the block starts at or below layer 22; the question-and-instruction control moves only over layers 22 to 27. Shading: layers 0 to 20.

| Quantity | Value |
|---|---|
| Flip, cumulative block, layer 0 / 20 / 22 / 28 | 7.75% / 12.25% / 14.00% / 2.00% |
| Baseline from layer | 32 |
| In-run endpoints, report present / absent | 1.50% / 12.00% |
| Single-layer block, maximum | 2.50% |
| Answer-position-only block, maximum | 2.00% |
| Blocking question and instruction instead | flips over layers 22 to 27 only |
| Clears zero against both controls at layers | 0, 4, 8, 12, 16, 20 |
| Flipped-case overlap, layers 0 to 27 | Jaccard 0.000 to 0.069 |
| Flipped-case overlap, layers 32 to 60 | Jaccard 0.250 to 0.500 |
| Report-label agreement, report arm, layers 0 to 20 | 0.6875 to 0.7400 |
| Same, question-and-instruction arm | 0.1725 to 0.2125 |
| Magnitude-matched image-span control | 0.0125 to 0.0250 at every depth |
| Size-matched report-span control, peak | 4.75% |
| Over-restoration at the peak, destination | 14.00% against 12.00%, agreement 0.7075 |
| Patching, wrong-donor control | 14.07% [10.78, 17.65] |
| Patching, correct donor, maximum | 12.25% |

Table 50: Report-token attention knockout and cross-condition patching, 400 cases. Flip rates use the cumulative, wide-scope variant unless stated. The wrong-donor control exceeds the correct donor at every depth, so patching localises nothing. Brackets: 95% patient-clustered bootstrap.

### Q.5 Interventions that did not work: steering and probe transfer

**Steering.** Scored on 1,721 test trials (12 findings, 150 cases each, the trials whose answer was the same on both images) by whether the steered answer on the substituted image matches that image's answer key, an accuracy against the substituted key and not the flip rate of Table 3. No arm separates from a random direction below the scale that degrades the answer (Table 51, Table 52). The prompt is the main experiment's and the readout is the model output; only the largest scale is shown, the direction, layer and magnitude are as run, and the record carries no confidence intervals.

**Probe transfer.** A linear probe (Alain & Bengio, 2016) fit under a frozen protocol was run once on 650 patients disjoint from every patient used anywhere else in this study. Every row of Table 53 rejects equality; the probe AUROC falls by 0.334, to a value not separable from chance. What the repository history verifies is narrower than a timestamp before the draw: the protocol hash appears in every output of the run, which started on 2026-08-04, while the protocol file entered version control on 2026-09-03. The protocol was fixed before evaluation; that it was fixed before the cohort was drawn is the authors' statement, not something the history can establish.

| Quantity | Value |
|---|---|
| Steering: baseline matches with the substituted key | 835 of 1,721 (0.485) |
| Difference-of-means direction, layer 20, scale 6,000 | 837 (0.486) |
| Every layer / random direction of same norm | $-18$ / $-14$ trials |
| Degradation of 1,209 correct: layer 20 / all / random | 0.167 / 0.614 / 0.328 |
| Probe: layer / findings / disjoint patients | 20 / 4 / 650 |
| Split rule | seed 0 from the training pool, disjoint from every cached store |
| Test | two-sided, $\alpha$=0.05, 10,000 resamples, one run |
| Frozen baselines | case-weighted over 13 findings, 0.952 |
| Probe AUROC change on transfer | $-0.334$, to chance |

Table 51: Steering (Marks & Tegmark, 2024; Turner et al., 2023; Li et al., 2023b; Rimsky et al., 2024) and probe transfer. Both are negative and exploratory.

| Condition | Tracks-substitute rate | Degradation rate | $\Delta$ vs. scale 0 |
|---|---|---|---|
| Targeted, layer 20 | 0.486 (837/1,721) | 0.167 | $+0.001$ |
| Global, all layers | 0.475 (817/1,721) | 0.614 | $-0.010$ |
| Random direction | 0.477 (821/1,721) | 0.328 | $-0.008$ |

Table 52: Difference-of-means steering on MIMIC-CXR, 12 findings, 150 cases each, largest scale (6,000). Tracks-substitute rate: share of 1,721 same-answer trials whose steered answer on the substitute matches its answer key, 0.485 before steering; not the flip rate of Table 3. Degradation: share of 1,209 originally correct answers turned wrong. No intervals.

| Statistic | Fresh cohort ($n$=650) | Test split | Gap (fresh − test) |
|---|---|---|---|
| Unchanged-answer rate | 93.95% [93.53, 94.37] | 95.22% | $-1.3$ [$-1.7$, $-0.9$] |
| Never-diverged rate | 57.10% [55.63, 58.47] | 62.30% | $-5.2$ [$-6.7$, $-3.8$] |
| Settling layer, median | 61.0 [61.0, 61.0] | 60.8 | $+0.2$ [0.2, 0.2] |
| Probe AUROC, layer 20 | 0.458 [0.414, 0.512] | 0.792 | $-0.334$ [$-0.379$, $-0.280$] |
| AURC (Geifman & El-Yaniv, 2017), probe | 0.414 [0.390, 0.437] | 0.490 | $-0.076$ [$-0.099$, $-0.053$] |
| AURC, max-softmax | 0.346 [0.323, 0.369] | 0.374 | $-0.028$ [$-0.051$, $-0.005$] |

Table 53: Frozen-protocol replication on a disjoint MIMIC-CXR cohort; the pre-specified two-sided test at $\alpha$=0.05 rejects equality in every row. Baselines are case-weighted over 13 findings (rates) or the four probe findings (AUROC, AURC) and are fixed constants. Brackets: 95% patient-clustered percentile bootstrap, 10,000 draws; unbracketed cells have none. The settling-layer row is a property of the projection, whose median sits well above the layer-46 informative depth in this arm and some 20 layers below it on the one-question unit, where the paired report-present minus report-absent difference is 1.56 layers [0.88, 2.21] and a shuffled-pairing null reproduces it; the row supports no depth claim.

### Q.6 The knockout does not transfer as a depth instrument outside MIMIC-CXR

The report-token knockout is possible only on MIMIC-CXR, the one dataset here that pairs an image with a report. The nearest manipulation elsewhere blocks attention to the image tokens instead, and on the three datasets where it was run it does not separate from its own control (Table 54), so no cross-dataset depth claim is made. Two of the three records also fail the repository's run-record schema and are listed among its grandfathered exceptions. The negative is reported because it bounds the paper: the depth result rests on the report-token knockout and its two controls on MIMIC-CXR, and on nothing else.

| Dataset | $n$ | Image block | Control | Control type |
|---|---|---|---|---|
| VQA-RAD | 52 | 48.1% [34.6, 61.5] | 51.9% [38.5, 65.4] | count-matched random tokens |
| OmniMedVQA | | 8.0% [5.6, 10.8] | 83.6% [80.1, 86.9] | random-key, not matched |
| SLAKE | 186 | 36.6% [29.6, 43.6] | none run | sweep marked exploratory |

Table 54: Answer-change rate when every image token is blocked, against each run's own control. No control sits below its treatment, so the manipulation is not a depth instrument outside MIMIC-CXR. Brackets: 95% intervals.

## R COHORT AND LABELS

Arm: the manipulation a row applies. Flip rate: share of pairs whose final answer differs between the original and substituted image. Unchanged-answer rate: its complement, an agreement statistic. Settling layer: first layer from which the lens-projected answer no longer changes, a property of the projection. Never diverged: share of pairs whose projected answers never differ at any layer. n/r: not reported, unlike a measured zero. Answer space and readout: MIMIC-CXR, yes/no from the first answer-token logits with the report in context; OmniMedVQA, one of four option letters; VQA-RAD, SLAKE and ProbMed, yes/no from the final-token answer logits.

How the MIMIC-CXR cohort and the other case pools were paired and labelled. Case pools by release, which bound every later denominator (the MIMIC-CXR cohort is fixed in the main text): OmniMedVQA, swap-changed case mining, pool 127,995, 427 admissible pairs, 33 swap-changed cases; SLAKE, swap-changed case inventory, pool 186, 90 swap-changed cases, swap-changed rate 0.484; VQA-RAD, pool ceiling mining, 52 admissible pairs, leakage-rule ceiling 66; VQA-RAD, pool relaxation, image reuse, 52 admissible pairs, leakage-rule ceiling 66. The OmniMedVQA answer-key audit flags 104 of 427 cases (0.244) whose correct option's text appears in the question; original-image accuracy is 0.576 overall and 0.607 on the 323 unflagged cases, and adversarial relabelling of every flagged case would bound it between 0.459 and 0.703; flip and unchanged-answer rates never read the key. On VQA-RAD, the automatic unchanged-answer label and the answer key's own image-decisive marking agree on 0.726 of 266 question-image rows (92 questions), Cohen's kappa 0.430.

## S ROBUSTNESS OF THE MAIN RESULT

The headline design re-run with changed prompt, input order, resolution, report section, decoding and image content; none of it enters the main-text estimates. Swap tier, flip rate pooled and over pairs whose substitute shares the source or organ (hard, as in the main experiment) or not (easy): OmniMedVQA, same-source against cross-source, 427 pairs: 0.068 pooled, 0.068 hard; VQA-RAD, same-organ against different-organ, 52 pairs: 0.615 pooled, 0.526 hard, 0.857 easy. Modality ablation on the MIMIC-CXR cohort, one question per case (its labelled finding), accuracy against the report-derived label as in Table 43: one input dropped (3,199 cases): 0.931 full prompt, 0.913 report plus question, 0.773 image plus question, 0.846 question alone; the image adds 0.018; one input dropped, re-run (3,199 cases): 0.931 full prompt, 0.913 report plus question, 0.773 image plus question, 0.846 question alone; the image adds 0.017. Response-bias floor of the unchanged-answer rate on MIMIC-CXR, the agreement a responder answering from its marginal alone would reach: over 7 per-finding evaluations, each over that finding's 3,199 pairs, the floor runs from 0.500 to 0.919.

| Manipulation | Arm | $n$ | Flip rate | Flips | Settling layer | Never div. | Still apart |
|---|---|---|---|---|---|---|---|
| prompt wording | clinical wording, yes/no order | 185 | 1.62% [0.55, 4.66] | 3 | 59.0 | 64.32% | n/r |
| prompt wording | direct wording, no/yes order | 185 | 4.32% [2.21, 8.30] | 8 | 62.0 | 64.86% | n/r |
| prompt wording | direct wording, yes/no order | 185 | 1.62% [0.55, 4.66] | 3 | 59.0 | 63.24% | n/r |
| prompt wording | impression wording, yes/no order | 185 | 1.08% [0.30, 3.86] | 2 | 59.0 | 61.08% | n/r |
| modality order | image before text | 200 | 3.00% | 6 | 59.0 | 25.50% | 48 |
| modality order | text before image | 200 | 16.50% | 33 | 62.0 | 2.50% | 122 |
| resolution | downsampled 2x | 200 | 2.00% | 4 | n/r | n/r | 47 |
| resolution | downsampled 4x | 200 | 2.00% | 4 | n/r | n/r | 51 |
| resolution | native resolution | 200 | 3.00% | 6 | n/r | n/r | 48 |
| report section | findings and impression | 200 | 3.00% | 6 | 57.0 | n/r | 48 |
| report section | findings only | 200 | 2.50% | 5 | 59.0 | n/r | 51 |
| report section | impression only | 200 | 25.50% | 51 | 58.0 | n/r | 97 |

Table 55: MIMIC-CXR under another prompt wording, modality order, resolution or report section; Settling layer is the median over pairs, Still apart the pairs whose projections differ at the final layer. The main experiment: image first, native, both sections. Brackets: Wilson 95% intervals; unbracketed cells have none. These are 200-case pilots and are superseded on the full cohort: modality order is rerun on all 3,199 cases in Table 1 (2.84% text first against 1.41% image first), and impression only is rerun on the 2,394 cases that carry an impression, 4.26% [3.34, 5.31], and on the 1,702 cases carrying both sections in Appendix Table 35. Read the 25.50% here as a 200-case pilot, not as the paper's estimate.

| Dataset | Arm | $n$ | Samples | Swap flip, greedy | Floor at $T$=0 | Floor at $T$=1 |
|---|---|---|---|---|---|---|
| MIMIC-CXR | repeated decoding, 40-case pilot | 40 | 5 | 0.00% | 0.00% | 42.50% |
| OmniMedVQA | repeated decoding, 40-case pilot | 40 | 10 | 5.00% | 0.00% | 20.00% |
| VQA-RAD | repeated decoding, full pool | 52 | 10 | 61.54% [48.08, 75.00] | 0.00% [0.00, 0.00] | 21.15% [11.54, 32.69] |

Table 56: Decoding-noise floor: answer-change rate between repeats at temperature 0 and 1.0, beside the greedy swap flip rate. Runs decode greedily, so the temperature-0 floor is zero by construction; the temperature-1 floor is a same-input control. Brackets: 95% bootstrap intervals; unbracketed cells have none.

| Dataset | Arm | $n$ | Agreement | 95% CI | Agree | Disagree |
|---|---|---|---|---|---|---|
| MIMIC-CXR | blank image | 200 | 97.50% | n/r | 195 | 5 |
| SLAKE | blank image | 186 | 55.38% | [48.39, 62.37] | 103 | 83 |
| VQA-RAD | blank image | 52 | 61.54% | n/r | n/r | n/r |

Table 57: Blank-image condition (image tokens present, content blank): share of cases whose answer with the blank image equals the answer with the real image. Agreement means only that the answer did not change. 95% CI: patient-level bootstrap, 10,000 draws. n/r: not recorded for that run.

| Dataset | Arm | $n$ | Clean acc. | Blur, max | Noise, max | Occlusion, max |
|---|---|---|---|---|---|---|
| OmniMedVQA | graded blur, noise and occlusion | 427 | 0.576 [0.529, 0.623] | 7.49% [5.15, 10.07] | 6.79% [4.45, 9.37] | 9.84% [7.03, 12.65] |
| SLAKE | graded blur, noise and occlusion | 186 | 0.710 [0.645, 0.774] | 24.19% [18.28, 30.65] | 17.20% [11.83, 22.58] | 44.62% [37.63, 51.61] |
| VQA-RAD | graded corruption | 52 | 0.731 [0.615, 0.846] | 28.85% | 34.62% | 38.46% [25.00, 51.92] |

Table 58: Graded image corruption: share of cases whose answer changes at the strongest blur, noise and occlusion, beside clean-image accuracy. VQA-RAD: corrected readout 0.731 supersedes the earlier double-normed 0.654. Brackets: 95% intervals (OmniMedVQA, bootstrap, 2,000 draws; SLAKE and VQA-RAD, bootstrap); unbracketed cells have none.

| Dataset | Arm | $n$ | Lowest dose | Highest dose | Trend $\rho$ | Trend $p$ |
|---|---|---|---|---|---|---|
| MIMIC-CXR | report length, text-side confound | 185 | 2.13% | 0.00% | −0.057 | 0.443 |
| MIMIC-CXR | swap severity, label Hamming distance | 185 | 0.00% [0.00, 29.91] | 2.27% [0.63, 7.91] | 0.044 | 0.549 |
| SLAKE | swap severity, pixel distance | 186 | 47.83% [32.61, 63.04] | 62.50% [47.92, 75.00] | 0.115 | 0.115 |
| VQA-RAD | swap severity, divergence tertiles | 52 | 12.50% | 100.00% | n/r | n/r |

Table 59: Swap-severity dose-response: flip rate in the lowest and highest bucket (each over its pairs), Spearman's rank correlation with severity across buckets and its two-sided $p$; Arm names the severity measure. Brackets: 95% intervals (MIMIC-CXR, Wilson; SLAKE, bootstrap); unbracketed cells have none.

| Finding | Observed | Chance-expected | Raw excess | Kappa |
|---|---|---|---|---|
| Atelectasis | 0.951 | 0.698 | 0.254 | 0.839 [0.807, 0.867] |
| Cardiomegaly | 0.936 | 0.500 | 0.435 | 0.871 [0.845, 0.896] |
| Consolidation | 0.931 | 0.559 | 0.373 | 0.845 [0.818, 0.868] |
| Edema | 0.963 | 0.514 | 0.449 | 0.923 [0.906, 0.940] |
| Enlarged Cardiomediastinum | 0.951 | 0.505 | 0.446 | 0.901 [0.883, 0.919] |
| Fracture | 0.988 | 0.861 | 0.126 | 0.910 [0.866, 0.948] |
| Lung Lesion | 0.914 | 0.514 | 0.401 | 0.824 [0.800, 0.846] |
| Lung Opacity | 0.960 | 0.763 | 0.197 | 0.832 [0.793, 0.865] |
| Pleural Effusion | 0.940 | 0.504 | 0.436 | 0.880 [0.860, 0.898] |
| Pleural Other | 0.956 | 0.500 | 0.455 | 0.911 [0.894, 0.927] |
| Pneumonia | 0.950 | 0.583 | 0.367 | 0.880 [0.856, 0.903] |
| Pneumothorax | 0.992 | 0.919 | 0.073 | 0.899 [0.837, 0.943] |
| Support Devices | 0.946 | 0.914 | 0.032 | 0.370 [0.271, 0.451] |
| No Finding | 0.964 | 0.501 | 0.463 | 0.927 [0.913, 0.941] |

Table 60: Cohen's kappa with condition-specific marginals, MIMIC-CXR, 14 findings: Observed is the finding's unchanged-answer rate, Chance-expected is $p_c p_d + (1 - p_c)(1 - p_d)$ from the two conditions' yes-rates, Kappa the excess over $1 -$ expected. Per row: 3,199 pairs per finding. Brackets on kappa: 95% patient-clustered percentile bootstrap over the 293 patients, 2,000 draws at seed 0, resampling the agreement and both yes-rate indicators jointly so each draw rebuilds the marginals kappa depends on.

## T OTHER DATASETS AND MODELS

These experiments evaluate the swap protocol on datasets without reports and through a second lens; they are not part of it. Accuracy under the swap on OmniMedVQA: all 427 of 427 swaps change the correct answer, so accuracy is scored against each condition's ground truth (chance 0.25): 0.579 on the original image, 0.457 on the substituted one, a drop of 0.122 pooled, $-0.003$ on anatomy identification and 0.654 on disease diagnosis. Text-side symmetry, share of cases whose answer changes: on OmniMedVQA (427 cases), 0.232 for a question swap with the image unchanged, 0.021 for an option swap, 0.068 for an image swap and 0.750 for a uniform option reroll; on SLAKE (186 cases), 0.419 for a question swap with the image unchanged, 0.484 for an image swap and 0.500 for a uniform option reroll; on VQA-RAD (52 cases), 0.269 for a question swap with the image unchanged, 0.615 for an image swap. Open-ended generation, a frozen rule mapping the free-text answer to an option: on OmniMedVQA, 427 of 427 answers mapped and 426 also cleared the 0.500 first-token probability floor (chance 0.250); on SLAKE, 111 of 186 answers mapped and 0 also cleared the 0.500 first-token probability floor; on VQA-RAD, 4 of 52 answers mapped and 0 also cleared the 0.500 first-token probability floor.

| Dataset | Arm | $n$ | Both inputs | Question only | Image, no question | Neither | Image adds |
|---|---|---|---|---|---|---|---|
| OmniMedVQA | one input dropped, or both | 427 | 0.576 [0.529, 0.623] | 0.543 [0.496, 0.590] | 0.508 [0.461, 0.555] | 0.515 [0.468, 0.564] | +0.033 [0.002, 0.061] |
| SLAKE | one input dropped, or both | 186 | 0.710 [0.645, 0.774] | 0.500 [0.430, 0.570] | 0.500 [0.430, 0.570] | 0.500 [0.430, 0.570] | +0.210 [0.118, 0.301] |
| VQA-RAD | inputs dropped | 52 | 0.731 [0.615, 0.846] | 0.538 [0.404, 0.673] | 0.500 | 0.500 | +0.192 [0.038, 0.346] |

Table 61: Modality ablation without a report: accuracy against the answer key with both inputs, question only, image under a neutral prompt, or neither; chance 0.25 on OmniMedVQA, 0.5 elsewhere. VQA-RAD: corrected readout 0.731 supersedes the earlier 0.654. Brackets: 95% bootstrap intervals; unbracketed cells have none.

| Dataset | Arm | $n$ | Concordant | Discordant | Flip rate |
|---|---|---|---|---|---|
| SLAKE | logit lens, full pool | 186 | 0.710 [0.645, 0.774] | 0.710 [0.645, 0.774] | 48.39% [41.40, 55.91] |
| SLAKE | tuned lens, full pool | 186 | 0.661 [0.591, 0.731] | 0.661 [0.591, 0.726] | 34.41% [27.96, 41.40] |
| VQA-RAD | logit lens | 52 | 0.731 | 0.731 | 61.54% |
| VQA-RAD | tuned lens, 2,000 steps | 52 | 0.673 | 0.673 | 34.62% |

Table 62: Readout comparison: concordant and discordant accuracy and flip rate of the answer projected at the final decoder layer by the output head (logit lens) or trained translator (tuned lens), not the model's output. Brackets: 95% bootstrap intervals, 10,000 draws; unbracketed cells have none.

| Dataset | Half | $n$ | Unchanged | Flip | Never div. |
|---|---|---|---|---|---|
| OmniMedVQA | development half | 175 | 97.71% [95.43, 99.43] | 2.29% [0.57, 4.57] | 0.57% |
| OmniMedVQA | replication half | 252 | 90.08% [86.51, 93.65] | 9.92% [6.35, 13.49] | 7.54% |
| SLAKE | development half | 92 | 50.00% [39.13, 59.78] | 50.00% [40.22, 60.87] | 13.04% |
| SLAKE | replication half | 94 | 53.19% [43.62, 62.77] | 46.81% [37.23, 56.38] | 10.64% [5.32, 17.02] |

Table 63: Frozen-protocol replication without a report: split by swap group fixed before any outcome was seen, analysis fixed on the development half, replication half run once. Unchanged and flip rates are complements. Brackets: 95% bootstrap intervals; unbracketed cells have none.

## U Exploratory mechanistic analyses

Layer-wise projections and interventions, descriptive of the lens projection; the main text draws no mechanistic claim from them. Activation patching over flip cases only: OmniMedVQA, patching with no-op control, 29 flip cases of 427 over 62 layers: patching restores the original answer most often at layer 60 and on more than half the flip cases first at layer 44; never-diverged share 0.047; VQA-RAD, activation patching, 32 flip cases of 52 over 62 layers: patching restores the original answer most often at layer 58 and on more than half the flip cases first at layer 31; never-diverged share 0.038. Tuned-lens training for MedGemma-27B, one run each, not used by the main experiment: WikiText, 100 steps, best validation KL 829.1 at step 100, 100 steps, 208 s wall time; WikiText-103, 1,000 steps, best validation KL 233.4 at step 900, 1,000 steps, 1,270 s wall time; WikiText-103, 2,000 steps, best validation KL 195.4 at step 1,700, 2,000 steps, 2,492 s wall time, final validation KL 197.6; seed reproducibility check, 60 steps. Probes, a linear classifier of swap-induced answer change on the residual at the stated layer: OmniMedVQA, cross-model probe, LLaVA-1.5-7B, 22 swap-changed cases of 427, out-of-fold balanced accuracy 0.572 against a matched shuffled null of 0.474; OmniMedVQA, leave-one-modality-out transfer, layer 45, 29 swap-changed cases of 427, 0 transfer folds above the null; VQA-RAD, leave-one-category-out transfer, layer 20, 32 swap-changed cases of 52; VQA-RAD, probe curve and stability, 32 swap-changed cases of 52, out-of-fold balanced accuracy 0.516 against a matched shuffled null of 0.444. Saliency baselines against the probe, area under the risk-coverage curve (lower is better) for abstention: OmniMedVQA (layer 45, 427 cases), 0.354 with the probe's score, 0.450 with image-token attention mass and 0.437 with the input gradient; VQA-RAD (layer 20, 52 cases), 0.199 with the probe's score, 0.210 with image-token attention mass and 0.320 with the input gradient.

| Dataset | Arm | $n$ | Flip rate | Unchanged | Peak div. | Layers |
|---|---|---|---|---|---|---|
| MIMIC-CXR | test split, 300 cases | 300 | 2.00% | 98.00% | 0.078 | 62 |
| MIMIC-CXR | test split, 976 cases | 976 | 1.43% | 98.57% | 0.076 | 62 |
| MIMIC-CXR | train split, 300 cases | 300 | 0.67% | 99.33% | 0.077 | 62 |
| MIMIC-CXR | train split, 976 cases | 976 | 0.92% | 99.08% | 0.074 | 62 |
| MIMIC-CXR | validation split, 300 cases | 300 | 1.33% | 98.67% | 0.072 | 62 |
| MIMIC-CXR | validation split, 880 cases | 880 | 1.02% | 98.98% | 0.076 | 62 |
| ProbMed | cross-image, same question | 2,634 | 39.86% | 60.14% | 0.633 | 62 |
| ProbMed | full test set, 27B | 6,158 | 68.87% | 31.13% | 1.000 | 62 |
| ProbMed | full test set, 4B | 6,158 | 71.65% | 28.35% | 1.000 | 34 |

Table 64: Settling depth by split: Peak div. is the median over pairs of the peak per-layer divergence between the two images' projected answer distributions, Layers the decoder depth. The MIMIC-CXR rows are earlier, smaller samples than Table 3. No intervals were computed.

| Dataset | Arm | $n$ | Cells | Computable | Significant | Median, min | Median, max |
|---|---|---|---|---|---|---|---|
| MIMIC-CXR | settling-definition grid | 3,199 | 24 | n/r | n/r | 58.0 | 62.0 |
| OmniMedVQA | definition grid, unchanged-answer cases | 427 | 24 | 15 | 7 | 42.0 | 59.5 |
| VQA-RAD | definition grid, unchanged-answer cases | 52 | 24 | 5 | 0 | 35.0 | 61.0 |

Table 65: Settling-definition sensitivity over a divergence-metric by threshold grid: grid cells, computable cells, cells where the two conditions' ordering stayed significant, and the range of the median settling layer over computable cells. No intervals were computed.

| Dataset | Arm | $n$ | Layers | Peak layer | Peak-layer change | All-layer change | Random-key change |
|---|---|---|---|---|---|---|---|
| OmniMedVQA | image attention masked, layer sweep | 427 | 62 | 8 | 8.20% [5.62, 11.01] | 7.96% [5.62, 10.77] | 83.61% [80.09, 86.89] |
| SLAKE | image attention masked, layer sweep | 186 | 62 | 0 | 4.30% [1.61, 7.53] | 36.56% [29.57, 43.55] | n/r |
| VQA-RAD | image attention masked, layer sweep | 52 | 62 | 23 | 34.62% [21.15, 48.08] | 48.08% [34.62, 61.54] | 51.92% [38.46, 65.38] |

Table 66: Attention knockout, layer-swept: share of cases whose answer changes with text-to-image attention masked at every layer, at the most effective layer only, or at random key positions of matched size. Nothing is localised. Brackets: 95% bootstrap intervals; unbracketed cells have none.

| Dataset | Arm | $n$ | Layer | Cases steered | Flip | Selective flip | Random flip |
|---|---|---|---|---|---|---|---|
| OmniMedVQA | positive control, answer axis | 427 | 45 | 427 | 18.27% | n/r | 21.78% |
| SLAKE | positive control, answer axis | 186 | 61 | 186 | 56.99% [50.00, 63.98] | 6.23% | n/r |
| VQA-RAD | positive control, answer axis | 52 | 40 | 52 | 50.00% | n/r | 11.54% |
| MIMIC-CXR | non-linear direction, layer 20 | n/r | 20 | n/r | n/r | 4.92% | n/r |
| MIMIC-CXR | per-layer efficacy sweep | 200 | 48 | 187 | 41.18% [34.22, 48.13] | 38.02% | 19.79% [14.44, 25.67] |
| OmniMedVQA | layer sweep, sub-saturation scales | 427 | 45 | 398 | 2.01% | 1.60% | 1.51% |
| SLAKE | steering l61 relative hi | 186 | 61 | 98 | 38.78% | 0.00% | 38.78% |
| VQA-RAD | difference-of-means, random null | 52 | 40 | 20 | 10.00% | 10.00% | 10.00% |
| VQA-RAD | per-layer efficacy sweep | 52 | 8 | 20 | 60.00% [40.00, 80.00] | 7.37% | 70.00% [50.00, 90.00] |

Table 67: Steering, positive controls (all cases steered) and sweeps (unchanged-answer cases), best cell: flip, share whose steered readout differs from the unsteered; selective flip, minus new errors among correct cases; random flip, largest under a magnitude-matched random direction. Brackets: 95% bootstrap intervals; unbracketed cells have none.

| Dataset | Arm | $n$ | Probe layer | AURC probe | AURC max-softmax | AURC random |
|---|---|---|---|---|---|---|
| OmniMedVQA | abstention regime sweep | 427 | 45 | 0.354 | 0.304 | 0.382 |
| SLAKE | AURC, held-out layer selection | 186 | 35 | 0.179 | 0.228 | 0.169 |
| VQA-RAD | AURC, swap-grouped folds | 52 | 20 | 0.199 | 0.134 | n/r |

Table 68: Abstention: area under the risk-coverage curve (lower is better) when the probe's score, the answer's maximum softmax or a random score decides which cases to abstain on; the probe is a linear classifier of swap-induced answer change. No intervals were computed.

# V RUNS THAT DID NOT YIELD AN INTERPRETABLE RESULT

Experiments whose records judge the result not interpretable: obsolete runs on an earlier cohort, and a readout that disagrees with the model's output. Reported for completeness; none is compared with the main text. Scale and finetuning on the MIMIC-CXR atelectasis cohort, one question per case, logit lens against the report-derived label, obsolete runs not comparable with Table 3: 4B against 27B, settling depth (3,199 cases), relative settling depth 0.735 for MedGemma-4B against 1.000 for MedGemma-27B under the report prompt; Gemma-3 against MedGemma, accuracy (3,199 cases), accuracy 0.433 for Gemma-3-27B against 0.415 for MedGemma-27B at layer 41 of 62, report condition unstated, majority baseline 0.777. The tuned-lens readout on the ProbMed test set, 6,158 pairs: flip rate 0.509, unchanged-answer rate 0.491, never-diverged share 0.008, for the answer projected by the text-trained translator at the final layer, which matches the model's output on 0.607 of cases; not interpretable.