CXR-CONTRABENCH: AUDITING POLARITY REVERSALS BEYOND AGGREGATE ACCURACY

Anonymous authors
Paper under double-blind review

ABSTRACT

A medical vision-language model can identify a finding on a chest X-ray and then deny it after a single answer option changes. We call this directional failure *negated-option attraction*. CXR-ContraBench measures it with a paired protocol that holds the image, question, reference finding, and retained options fixed while one distractor is replaced by the negated target, and records every answer transition, including correct answers that become false denials and wrong answers that recover. On 260 paired CheXpert items, MedGemma-4B-it turns 89 of 158 initially correct answers into false denials and Qwen2.5-VL-7B turns 41 of 116, conditional reversal rates of 56.33% and 35.34%. The aggregate accuracy of Qwen2.5-VL nevertheless rises by 5.77 percentage points because recoveries offset these reversals. Because the target-paired answer set admits an image-free structural ceiling, image-conditioned validity is evaluated separately: matched, opposite-label, shuffled, and blank images change which statement is true while the text stays fixed. Under this diagnostic, expected polarity updating is limited, a polarity-explicit prompt changes answer behavior with a benefit that depends on the measured endpoint, and in-format adaptation reaches perfect format accuracy without paired image-conditioned correctness. The transition persists across ten wording families, option-free single-sentence generation, OpenI, frontier models, and 135,754 CheXpert records from 59,173 studies. Average accuracy should therefore be interpreted together with whether previously correct judgments are preserved. Code is available at https://anonymous.4open.science/r/cxr-contrabench/.

1 INTRODUCTION

Medical vision-language models are increasingly evaluated by whether they select the right finding from a chest radiograph. A polarity error is more specific than an ordinary wrong answer: it asserts the opposite clinical statement. Consider an image labeled positive for consolidation and a question asking which listed finding is present. A model may select *Consolidation* from three affirmative options, then select *No consolidation* when one distractor is replaced by that phrase. The image, the question, and the correct answer are unchanged, yet a correct identification has become a denial of the labeled finding, as Fig. 1a illustrates. We call this transition *negated-option attraction*.

Aggregate accuracy cannot expose this failure because it is the net of opposing row-level flows. Replacing a distractor can recover an initially wrong answer on one item while turning an initially correct answer into a false denial on another, and the two flows cancel in the average, so a configuration whose score rises may still have introduced new false denials. For clinically directional errors, the relevant question is not only how many answers are correct, but which previously correct answers were lost and where they went. This emphasis on error direction mirrors clinically informed evaluation, in which error categories are not interchangeable under one aggregate score (Yu et al., 2023).

Negation has mostly been studied as a property of an input or query: NegBench evaluates negated captions and retrieval queries, NegVQA negates the question, and clinical work studies polarity-aware report adaptation (Alhamoud et al., 2025; Zhang et al., 2025; Ko & Park, 2025; Abbasi et al., 2026). These evaluations ask whether a model understands a negated input. We ask instead whether a fixed image and a fixed question preserve a correct answer when a negated statement enters the answer set, and we account for every paired answer transition rather than reporting a single score.

CXR-ContraBench implements this measurement as a paired protocol on chest radiographs. Each item fixes the image, the presence question, the reference-present finding, the option count, and the retained

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.

option positions; the only change between control and intervention is that one reference-absent distractor is replaced by the negated target. Because both answer sets are scored on the same rows, every response falls into exactly one transition state.

On 260 paired CheXpert items, MedGemma-4B-it converts 89 of 158 initially correct answers into false denials and Qwen2.5-VL-7B converts 41 of 116. The accuracy of Qwen2.5-VL nevertheless rises by 5.77 percentage points because its recoveries outnumber its reversals, whereas LLaVA-Med-1.5-7B never selects the negated option on the same items.

A high score on this construction is not by itself evidence of image understanding: because the inserted option always negates the target finding, an image-free rule that selects the affirmative member of the displayed pair is correct on every target-paired item. We therefore evaluate image-conditioned validity separately, with a diagnostic that keeps the question and options fixed while matched, opposite-label, shuffled, and blank images change which statement is true. Blank-image probes reveal strong answer preferences for some models in the absence of usable visual evidence, which motivates this separate analysis and frames how prompting and adaptation gains are interpreted.

Practical use. The paired audit is intended as a regression test when comparing checkpoints, prompts, or adaptation strategies. Two configurations can have similar or even oppositely ranked aggregate accuracy while differing in whether previously correct answers are preserved. A configuration that improves the mean score but introduces new correct-to-false-denial transitions is exposed by the paired flow, and an intervention that reduces those reversals can be checked separately for image-conditioned correctness. CXR-ContraBench therefore supports model-selection and update auditing rather than estimating the prevalence of errors in clinical deployment.

Our contributions are threefold. (1) *Paired measurement.* We define negated-option attraction as a paired correct-to-false-denial transition and report it with complete row-level flows, so that reversals, recoveries, other regressions, and invalid outputs remain visible alongside accuracy. (2) *Validity interpretation.* We identify the structural ceiling of target-paired answer sets and pair the primary result with an image-validity diagnostic that changes condition-specific truth while keeping the text fixed, separating format gains from image-conditioned correctness. (3) *Breadth.* We characterize how the transition depends on prompt, wording, answer format, data source, scale, and model family, from option-free generation to a 135,754-record CheXpert construction and three frontier API systems. Together, these analyses show why model improvements should be evaluated not only by average accuracy, but also by whether previously correct judgments are preserved.

2 Related Work

Negation benchmarks. NegBench evaluates negated captions, retrieval queries, and multiple-choice templates, including medical data and generative models (Alhamoud et al., 2025); NegVQA constructs negated visual questions (Zhang et al., 2025); CXR-Align and NAST study clinical polarity and negation-aware adaptation for chest radiographs (Ko & Park, 2025; Abbasi et al., 2026), with related evidence for contrastive medical encoders (Vu & Sheshappanavar, 2025). These works establish negation evaluation in medical and generative settings. Our distinction is operational: the same image, the same question, one answer-option substitution, and a paired flow that records where each previously correct answer goes.

Negative flips, selection bias, and visual dependence. Negative-flip analyses measure sample-wise regressions across model updates so that an improved average is not read as uniform improvement (Yan et al., 2020); we apply the same accounting to one model under an answer-set intervention rather than across model versions. Multiple-choice evaluation can itself introduce option-selection and position biases in language and multimodal models (Zheng et al., 2024; Pezeshkpour & Hruschka, 2024; Zeno et al., 2025), and MMBench's CircularEval perturbs answer presentation (Liu et al., 2024); such findings motivate controlled comparisons that hold answer presentation fixed except for the intended intervention. Complementary-image VQA pairs one question with images that require different answers (Goyal et al., 2017), text can dominate visual evidence (Deng et al., 2025), and related audits caution that high multimodal accuracy does not by itself establish dependence on, or robustness to, the visual input, motivating explicit visual interventions when grounding is the intended capability (Lotfinia et al., 2026; Cheng et al., 2025). We combine these perspectives into a clinical directional transition with complete flow accounting and a separate image-validity diagnostic under fixed text.

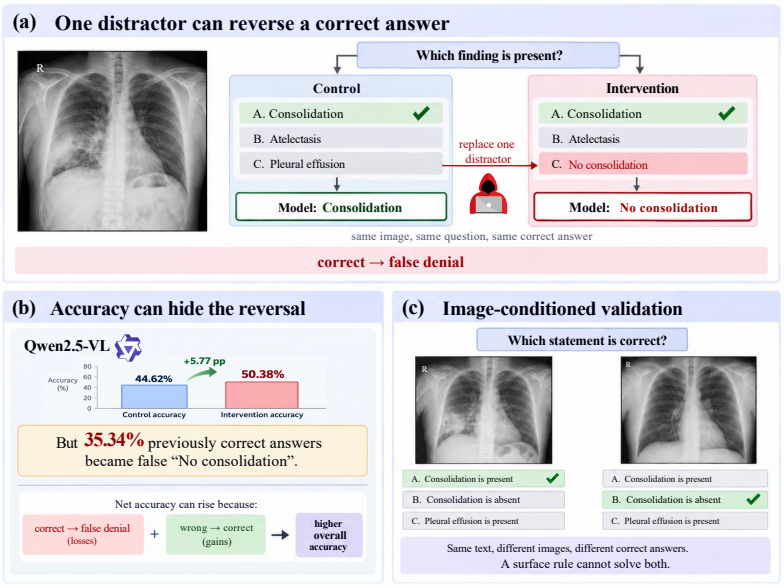


Figure 1: **Controlled polarity audit.** (a) With the image, question, and reference answer fixed, replacing one distractor with the negated target turns a correct selection into a false denial; the radiograph is a schematic illustration, not a case from the evaluation logs. (b) Aggregate accuracy can hide the reversal: Qwen2.5-VL rises from 44.62% to 50.38% while 35.34%, or 41 of 116, of its initially correct answers become *No* X. (c) Image-validity diagnostic: the text stays fixed while the image changes, so different images make different statements true and a surface rule cannot satisfy both.

Medical vision-language models and benchmarks. Medical VQA benchmarks broaden coverage, scale, and trustworthiness analysis (Lau et al., 2018; Liu et al., 2021; Zhang et al., 2023b; Xia et al., 2024; Chen et al., 2024a; Zhou et al., 2024; Pal et al., 2025; Royer et al., 2024; Xie et al., 2024; Lee et al., 2025), and recent medical VLMs span instruction-tuned assistants, contrastive pretraining, and report-generation systems (Li et al., 2023; Sellergren et al., 2025; Chen et al., 2024b; Moor et al., 2023; Zhang et al., 2023a; Wu et al., 2023; 2025; Thawakar et al., 2023; Pellegrini et al., 2023; Wang et al., 2023). We evaluate representative backbones from these families and interpret cross-model differences behaviorally rather than attributing them to a single architecture or pretraining source.

3 CXR-ContraBench: Paired Protocol and Measurements

3.1 Task and Experimental Units

Each primary item consists of an image input I_i, an affirmative presence question q_i, a reference-present finding x_i, and two reference-absent distractors y_i, z_i. Missing and uncertain labels remain distinct from absence. A model response is a selected option under the complete answer set, $\hat{a}_i^{(k)} = f_\theta(I_i, q_i, O_i^{(k)})$. The primary paired set contains 260 paired records from 127 CheXpert studies (Irvin et al., 2019); in this slice each study corresponds to one patient. Records, studies, and patients are separate experimental units. Supporting protocols on ReXVQA, OpenI, and the CheXpert training split retain their native questions and scoring rules and are summarized in Appendix A; only the 260-item paired set carries the row-level directional transition accounting used in the primary analysis. The locally evaluated backbones are MedGemma-4B-it (Sellergren et al., 2025), Qwen2.5-VL-7B (Bai et al., 2025), and LLaVA-Med-1.5-7B (Li et al., 2023), referred to below as MedGemma, Qwen2.5-VL, and LLaVA-Med.

3.2 Controlled Answer-Set Substitution

The control and intervention options are $O_i^{(0)} = [x_i, y_i, z_i]$ and $O_i^{(1)} = [x_i, y_i, \text{No } x_i]$; the image, question, reference, number of choices, and retained option positions are unchanged. Selecting No-x_i contradicts

the positive reference and is counted as a false denial. A No-y_i option in the paired-distractor arm is scored by its displayed role: it states a true absent finding while remaining nonresponsive to the requested present-finding question.

3.3 DIRECTIONAL TRANSITIONS

Let $c_i^{(k)} = \mathbf{1}\{\hat{a}_i^{(k)} = x_i\}$. For paired control and intervention outputs define

$$R = \{i : c_i^{(0)} = 1,\ \hat{a}_i^{(1)} = \text{No}\, x_i\}, \qquad I = \{i : c_i^{(0)} = 1,\ \hat{a}_i^{(1)} = \text{invalid}\}, \tag{1}$$

$$H = \{i : c_i^{(0)} = 1,\ c_i^{(1)} = 0,\ \hat{a}_i^{(1)} \notin \{\text{No}\, x_i, \text{invalid}\}\}, \qquad G = \{i : c_i^{(0)} = 0,\ c_i^{(1)} = 1\}. \tag{2}$$

R is correct-to-false-denial, H is any other valid new error, I is a correct-to-invalid loss, and G is recovery; invalid output is tracked as a separate flow. We report

$$\text{RR} = \frac{|R|}{\sum_i c_i^{(0)}}, \qquad \text{RF} = \frac{|R|}{N}, \qquad \Delta\text{Acc} = \frac{|G| - |R| - |H| - |I|}{N}. \tag{3}$$

Each RR carries its model-specific control-correct denominator; RF, the invalid rate, and the complete transition matrix keep denominator changes and always-wrong behavior visible.

3.4 IMAGE-VALIDITY DIAGNOSTIC

The image-validity diagnostic uses the same 260 units and four image conditions. Every real-image row contains the literal question "Which of the following findings is present on this chest X-ray study?" and the same three lettered options. The matched image shows the target finding; the deterministic shuffle keeps the target role; the opposite-label donor removes the target; blank rows have no gold and are unscored. Importantly, the opposite-label row is not interpreted as ordinary accuracy on the original presence question; it is a statement-truth probe over the fixed set of displayed statements. Matched and shuffle rows are scored against the positive-target statement and opposite-label rows against the negated statement, so all real-image rows use condition-specific statement truth. We report condition statement correctness, the expected positive-to-negative transition, PairCorrect, which requires both real-image roles to be correct, and shuffle stability separately; these rows measure image sensitivity under a declared truth rule and do not identify the internal cause of a response.

3.5 SCORING, EXECUTION, AND UNCERTAINTY

Every contrast is fixed before scoring: the paired protocol checks that image, question, option count, retained positions, and reference finding are identical across arms, and the image protocol checks that question and options are identical across conditions, as Appendix A details. All completed rows remain in the denominator; parse failures and outputs outside the displayed options are recorded as invalid attempts. Blank-image rows have no clinical gold and are excluded from accuracy, but their selections are retained as answer-preference diagnostics. Uncertainty is a percentile bootstrap over study clusters with 2,000 replicates, because a study can contribute several views. The three base checkpoints share one answer-set renderer; MedGemma and Qwen2.5-VL consume all available study images, whereas LLaVA-Med uses the first image only. Raw responses and parsed semantic choices are stored separately.

3.6 STRUCTURAL BASELINES

Target-paired answer sets admit image-free checks that bound what a score can establish. First, an option-only counterpart rule that selects the affirmative member of the displayed X/No-X pair is correct on 235/235 items of the direct-presence protocol, because the inserted option always negates the target. Second, conditional remapping applies the same counterpart substitution only when a model selected No-X; on that construction it reaches 96.60% for MedGemma and 95.32% for Qwen2.5-VL, and it performs no additional visual assessment. Third, on the 260-item four-arm set, remapping a negated selection to its counterpart is correct on 179/179 MedGemma and 123/123 Qwen2.5-VL rows when the negated option is the target, and on 0/226 and 0/165 rows when it negates a non-target. These values quantify answer-set structure, so deterministic remapping is a construction-dependent baseline, not independent visual inference.

Table 1: **One negated option reverses correct answers while accuracy can rise, and visual changes produce limited polarity updating.** The image, question, and reference finding are fixed; only the displayed option set (left) or the image (right) changes. $\bar{X}$ denotes "No X." Cell tint marks RR: < 10, 10–30, ≥ 30.

	Option intervention, accuracy (%)				Off-target arms (%)		Image intervention (%)	
Backbone	$[X, Y, Z]$	$[X, Y, \bar{X}]$	Δ	RR R/N_c	$[X, Y, \bar{Y}]$	$[X, Z, \bar{Y}]$	Blank $\bar{X}$ chosen	Absent img. $X \to \bar{X}$
MedGemma-4B	60.77%	29.62%	-31.15	**56.33** 89/158	5.38%	11.92%	96.54%	17.69%
Qwen2.5-VL	44.62%	50.38%	+5.77	**35.34** 41/116	1.54%	25.00%	72.69%	11.54%
LLaVA-Med	49.62%	77.31%	+27.69	0.00 0/129	41.92%	58.08%	0.00%	0.00%

The complete mutually exclusive row-level flows are $CC/R/H/G/WW/Invalid$ = $(69/89/0/8/94/0)$ for MedGemma, $(75/41/0/56/88/0)$ for Qwen, and $(127/0/2/74/57/0)$ for LLaVA-Med. All-item RF is 34.23%, 15.77%, and 0.00%, respectively; study-cluster RR 95% CIs are 48.99–63.74, 26.53–44.17, and 0.00–0.00, respectively. The LLaVA interval is zero for every empirical bootstrap replicate.

Evaluation principles. CXR-ContraBench follows three principles for interpreting controlled multimodal evaluations. First, predictions are compared on the same item before and after a precisely specified intervention, so that changes are attributable to the manipulated input rather than to a different sample set. Second, we report the direction of answer changes rather than only their net effect on accuracy, distinguishing recoveries from regressions that can cancel in the aggregate. Third, we explicitly test whether a favorable score can be obtained without using the evidence the task is intended to measure. These principles motivate the paired transition analysis, structural baselines, and image-validity diagnostic below.

4 Directional Answer Transitions

4.1 Primary paired result

Table 1 reports the option and image interventions on the paired CheXpert set, and Fig. 2 visualizes the row-level flows. Three observations follow.

The reversal is large and backbone-dependent. Among answers that MedGemma gets right under $[X, Y, Z]$, 56.33%, or 89 of 158, become "No X" when the third option is replaced by the negated form of the present finding. For Qwen2.5-VL the rate is 35.34%, or 41 of 116, and for LLaVA-Med it is zero. The study-cluster 95% intervals of 48.99 to 63.74 and 26.53 to 44.17 exclude small effects for the first two backbones.

Aggregate accuracy combines opposing flows. Qwen2.5-VL gains +5.77 pp under $[X, Y, \bar{X}]$ while reversing 35.34% of its correct answers, because its wrong-to-correct flow of 56 recoveries is larger than its correct-to-denial flow of 41 reversals. MedGemma loses 31.15 points because 89 reversals dominate 8 recoveries. LLaVA-Med gains 27.69 points while selecting the negated option zero times: replacing a competitive distractor with an option it never chooses leaves a two-way affirmative choice. Complete row-level transition counts are visualized in Fig. 2 and retained in the released row-level analysis.

Visual changes produce limited polarity updating. The right-hand columns of Table 1 keep the option set $[X, Y, \bar{X}]$ and change only the image. On a blank image MedGemma and Qwen2.5-VL select the negated option 96.54% and 72.69% of the time, a measure of answer preference in the absence of usable image evidence rather than of clinical accuracy. When a confirmed-present image is replaced by a confirmed-absent one, only 17.69% and 11.54% of their correct answers move to the negated option. MedGemma and Qwen2.5-VL thus show limited expected updating under the opposite-label intervention. LLaVA-Med exhibits a different output policy, with no native No-X transition under the released parser, and is interpreted separately.

4.2 Structural controls

Negated statements can dominate task compliance. The same 260 items are evaluated under four arms, positive-only $[X, Y, Z]$, paired-target $[X, Y, \bar{X}]$, paired-distractor $[X, Y, \bar{Y}]$, and no-counterpart $[X, Z, \bar{Y}]$;

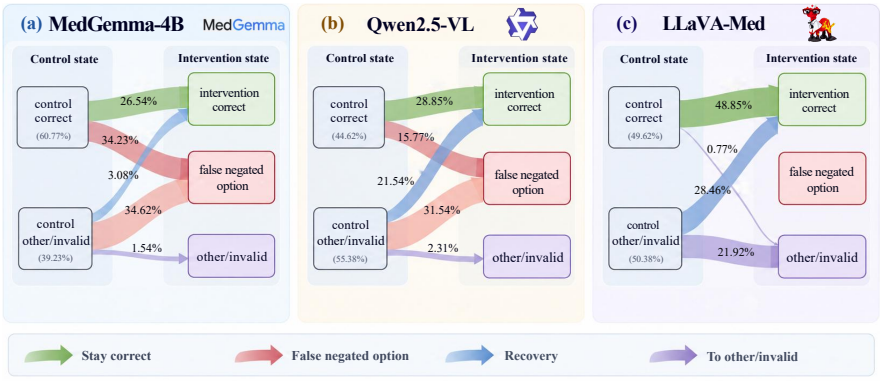


Figure 2: **Primary paired transition on the 260-item CheXpert set.** Each panel shows one backbone's row-level flow from $[X, Y, Z]$ to $[X, Y, \bar{X}]$; percentages are fractions of all 260 paired units. Recoveries and harmful reversals coexist: Qwen2.5-VL's 5.77-point gain arises because 56 recoveries outnumber 41 reversals. The figure describes observed transitions, not an internal mechanism.

the two off-target arms occupy the middle columns of Table 1. Under $[X, Y, \bar{Y}]$ the negated option is a true statement about an absent finding, while the question asks for the present finding. MedGemma's and Qwen2.5-VL's accuracies fall to 5.38% and 1.54%, and removing the affirmative counterpart Y in $[X, Z, \bar{Y}]$ raises them only to 11.92% and 25.00%, whereas LLaVA-Med reaches 41.92% and 58.08%. The attraction toward the negated statement therefore persists when that statement is off-target and when its counterpart is absent. Because the negated option in $[X, Y, \bar{X}]$ always negates the present finding, the image-free affirmative-member rule of Sec. 3.6 scores 100% on that arm and 0% on both off-target arms; paired-target accuracy must be read against this ceiling. Full structural controls and remapping diagnostics are reported in Appendix E.

5 What Do Improved Scores Establish?

5.1 Prompting and In-Format Adaptation

Table 2 scores two interventions that raise in-format accuracy; each prompt is compared with its own matched positive-only control, so reversal rates are conditional on the answers that prompt gets right.

Reasoning prompts are configuration-dependent. MedGemma's reversal stays between 51.27% and 58.68% under all four prompts. Qwen2.5-VL's conditional reversal rate is 35.34% under direct prompting, 2.00% under the polarity-explicit prompt, and 44.05% and 57.78% under the evidence-first and image-first prompts, on 116, 100, 84, and 90 control-correct rows. The polarity-explicit prompt is a real improvement on the answer-set endpoint for Qwen2.5-VL, and it is the configuration carried into the image-validity diagnostic below. Three-shot prompting is reported in Appendix C.

In-format LoRA reaches perfect format accuracy and transfers. Adapters trained in the evaluation format on CheXpert training-split questions reach 100% on the presence protocol for both backbones, up from 31.49% and 30.21%, and the lower block of Table 2 shows transfer to indirect wording, scope wording, and the no-counterpart set. Whether these gains are accompanied by image-conditioned correctness is addressed next.

5.2 Image-Validity Results

Fig. 3 reports, for each backbone, statement correctness under the condition-specific truth rule, No-X selection, PairCorrect, and the visual polarity contrasts when the text is fixed and only the image changes.

Base checkpoints. The expected matched-to-opposite target-polarity transition occurs on 17.69% of MedGemma rows and 11.54% of Qwen2.5-VL rows, and PairCorrect takes the same values; LLaVA-Med never selects No-X and therefore has 0% PairCorrect. MedGemma selects No-X frequently under every condition, including 96.54% of blank rows, whereas Qwen2.5-VL's selections are closer to balanced but

Table 2: **Prompting and in-format LoRA strongly change answer-set behavior; image validity is evaluated separately.** Each prompt is scored against its own matched positive-only control; LoRA cells are accuracy before and after adaptation, with the first row on the training format and the others on protocols the adapter never saw.

Configuration	Quantity	MedGemma-4B-it	Qwen2.5-VL-7B
Chain-of-thought prompts, correct-to-"No X" reversal (lower is safer)			
Direct	RR (%)	**56.33**	**35.34**
Polarity-explicit	RR (%)	**51.27**	2.00
Evidence-first	RR (%)	**55.84**	**44.05**
Image-first, two-pass[a]	RR (%)	**58.68**	**57.78**
In-format LoRA, accuracy base $\rightarrow$ adapted (%)			
Training format, presence protocol	acc.	31.49 $\rightarrow$ **100.00**	30.21 $\rightarrow$ **100.00**
Indirect wording, 4 families	acc.	12.74 $\rightarrow$ **87.26**	47.28 $\rightarrow$ **94.04**
Scope wording, 2 families	acc.	19.04 $\rightarrow$ **66.15**	15.96 $\rightarrow$ **54.81**
$[X, Z, \bar{Y}]$, no counterpart	acc.	11.92 $\rightarrow$ **54.23**	25.00 $\rightarrow$ **48.46**

[a] The first pass receives the image only, with no question, options, or candidate strings.

likewise change little between matched and opposite-label images. Blank rows quantify answer preference without usable image evidence; their first-letter logit gaps of +8.51 and +0.66 toward the negated option are descriptive statistics, not evidence for an internal mechanism.

Polarity-explicit prompt. The polarity-explicit prompt lowers Qwen2.5-VL's conditional reversal rate from 35.34% to 2.00%, but the same image-validity run contains 346/1,040 unparseable outputs and only 8/260 PairCorrect cases, a rate of 3.08%, with per-condition counts in Appendix C. The RR reduction therefore should not be interpreted as improved image-conditioned validity. The prompt substantially changes the model's output behavior, and its benefit depends on the endpoint being measured.

In-format LoRA. In-format LoRA reaches 100% on the original protocol and transfers to several unseen wording and answer-set conditions. Under the image-validity diagnostic, however, both adapted checkpoints are statement-correct on every matched row and on no opposite-label row, so both obtain 0.00% PairCorrect. High format accuracy therefore does not by itself imply image-conditioned correctness. Wording and structure transfer and image validity answer different questions, and both should be reported.

6 Generalization Across Wording, Format, and Models

Table 3 collects checks across wording, answer format, data source, and model family; each block keeps its native denominator and scoring rule and is not pooled with the 260-item primary flow.

6.1 Wording and option-free generation

The canonical "No X" form is one of ten wording families listed in Appendix D: four explicit families contain a literal negation token, four indirect families such as "lungs are clear of X" do not, and two scope families negate several findings or the whole option list. The transition persists without a literal "No": indirect families reach 80.27% for MedGemma, and scope families reach 70.89% and 70.69% for MedGemma and Qwen2.5-VL, while LLaVA-Med stays at or below 10.66% across families. Option-free generation removes the candidate answers entirely: for a confirmed-positive finding, the model generates one clinical sentence under each member of a matched question pair. MedGemma changes from affirmation to denial in 361/520 comparisons and Qwen2.5-VL in 238/520; the per-pair rates are in Table 3a, every attempt, including parser failures, stays in the denominator, and parser agreement with an independent text classifier is 99.23%/97.69%. This single-sentence controlled generation shows that the denial is not tied to the presence of candidate answers, but it is neither full-report generation nor the fixed-question paired intervention.

6.2 Source and scale

On the 396-record OpenI CheXStruct presence subset (Demner-Fushman et al., 2016; Lee et al., 2025), negated options account for 96.77% of MedGemma's wrong answers, 270 of 279, and for 92.58% of Qwen2.5-VL's, compared with 41.61% for LLaVA-1.5-7B and 21.19% for LLaVA-Med. These wrong-answer composition statistics on a second source complement, rather than replicate, the paired transition rate. At larger scale, a CheXpert training-split construction with 135,754 records from 59,173 studies, in which one study contributes several records and views, reports presence-side negated-option selection above 62% for both MedGemma and Qwen2.5-VL; that value is a selection rate against source labels, not a within-item transition.

6.3 Frontier models

Three API systems, recorded under their served identifiers claude-opus-5, gpt-5.5, and gemini-3.6-flash, are evaluated under the same paired principle in Table 3b and Appendix F. Directional reversals remain measurable: 14.52%, 13.73%, and 7.14% of each system's control-correct answers become No-X, with study-cluster intervals that exclude zero, while all three gain accuracy because recoveries outnumber reversals. Reversal rates under the other explicit wordings stay in a similar range for each system, and 50.77–77.69% of answers change when the image is blank. Their denominators and execution environments differ from the open-weight runs, so the comparison is a breadth check, not a ranking.

6.4 Implications for multimodal evaluation

These results illustrate a broader distinction between capability scores and response stability. An intervention can improve aggregate accuracy by repairing some mistakes while introducing new regressions, so the net score alone does not reveal which previously correct judgments were lost. Conversely, a near-perfect score can arise from answer-set structure when the construction leaks a deterministic shortcut, as the affirmative-member rule of Sec. 3.6 shows. For multimodal evaluation, a favorable score therefore raises two separate questions: whether the system is correct, and whether that correctness depends on the evidence the task is intended to test. CXR-ContraBench instantiates this distinction for polarity-sensitive chest

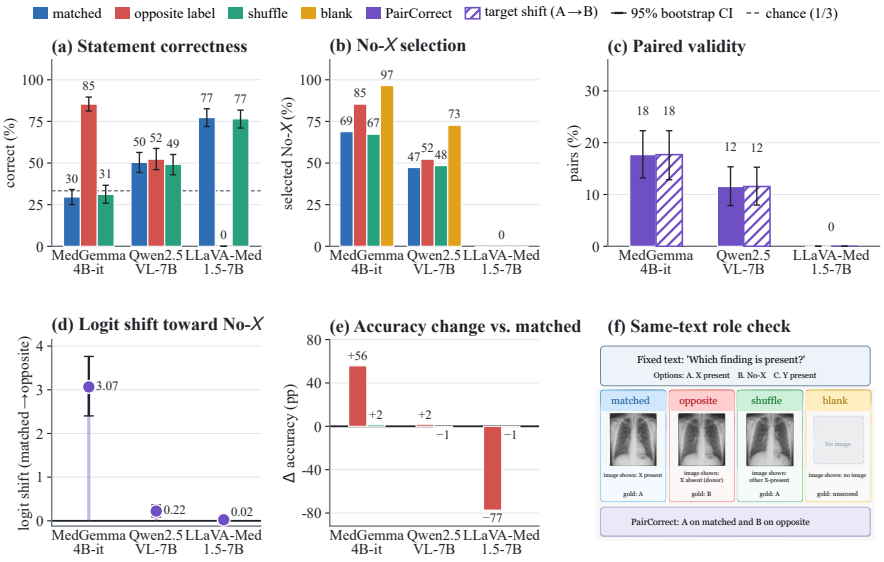


Figure 3: **Image-validity diagnostic.** Real-image conditions keep the text identical and change only the image, so the gold statement changes with the image role (f); blank rows have no clinical gold and are unscored. The opposite-label condition is scored as a fixed-option statement-truth probe rather than as ordinary presence-question accuracy. Panels: (a) statement correctness under the condition-specific gold, (b) No-X selection, (c) PairCorrect and the expected target-polarity shift, (d) logit shift and (e) accuracy difference relative to matched rows. Error bars are study-cluster bootstrap intervals; blank-image selections describe answer preference, not clinical accuracy.

Table 3: **The transition persists across wordings, without answer options, on a second source, and in frontier models.** Blocks report their native units alongside the paired result.

(a) Open-weight backbones

Track	Quantity	MedGemma-4B	Qwen2.5-VL	LLaVA-Med
Explicit wordings, 4 families	RR (%)	**62.66** 396/632	**43.97** 204/464	10.66 55/516
Indirect wordings, 4 families	RR (%)	**80.27** 480/598	**38.16** 166/435	9.50 49/516
Scope wordings, 2 families	RR (%)	**70.89** 224/316	**70.69** 164/232	1.16 3/258
Free-form, present / not present	affirm-then-deny (%)	70.77	46.92	96.92[a]
Free-form, evidence / no evidence	affirm-then-deny (%)	68.08	44.62	0.00[a]
OpenI CheXStruct presence	$\bar{X}$ share of errors (%)	96.77	92.58	21.19

(b) Frontier models

Model	RR (%) R/N_c	95% CI	Other wordings (%)	Blank change (%)
claude-opus-5	14.52 18/124	[8.41, 21.31]	9.56–19.86	50.77
gpt-5.5	13.73 21/153	[8.62, 19.70]	11.56–15.69	77.69
gemini-3.6-flash	7.14 9/126	[3.06, 12.15]	4.88–8.00	51.54

[a] LLaVA-Med differs across the two question pairs; both native rates are shown.

X-ray questions through paired transition accounting and a separate image-validity diagnostic. The same principle may be useful more broadly when small answer-set or prompt changes introduce semantically consequential alternatives, without implying that every multimodal benchmark exhibits this failure or requires this exact protocol.

7 Limitations

CXR-ContraBench is a controlled diagnostic rather than an estimate of clinical error prevalence. The primary paired analysis relies on answer sets in which a reference-positive finding and its negated statement can be compared explicitly, and the resulting rates should not be extrapolated to unconstrained radiology reports. Reference labels inherit the limitations of the source datasets, and public chest-radiograph corpora may overlap with model pretraining; the within-item paired comparison reduces image-distribution confounding but does not eliminate the influence of exposure, language behavior, answer formatting, or training distributions. The image-validity diagnostic replaces the study while keeping the displayed text fixed and evaluates condition-specific statement truth; it is not a within-patient counterfactual and does not identify the internal cause of a model response. Blank images are deliberately non-clinical probes of answer preference. Multi-view handling differs across backbones, with LLaVA-Med using the first available study image. Finally, the option-free evaluation generates single sentences rather than full radiology reports, and frontier-model results depend on API versions that we record but cannot freeze.

8 Conclusion

We presented CXR-ContraBench, a diagnostic evaluation of polarity reversals in chest X-ray vision-language models. By holding the image, question, reference finding, and retained answer positions fixed while changing one distractor, it exposes a failure that aggregate accuracy cannot describe: previously correct finding selections become false denials even when the overall score improves. The paired analysis shows substantial correct-to-false-denial transitions for MedGemma and Qwen2.5-VL, and row-level accounting explains how recoveries mask these regressions.

The experiments also show why a high score on a polarity-sensitive benchmark needs careful interpretation: target-paired answer sets admit image-free structural shortcuts, prompting effects are configuration-dependent, and in-format adaptation can suppress the original format error without guaranteeing paired image validity, so the complementary evaluations are distinct checks, not interchangeable metrics.

More broadly, these results motivate evaluating model improvements by what they preserve, not only by what they add to average accuracy. For clinically directional errors such as false denial, reporting where answers move reveals failures that a single aggregate number obscures. In this sense, the benchmark is most useful as a regression test for model, prompt, and adaptation updates, where preserving previously correct judgments is itself part of reliability.

AI USE STATEMENT

AI assistants were used for language editing, manuscript organization, and analysis scaffolding, including assistance in drafting and refactoring auxiliary analysis code. They were not treated as sources of experimental evidence. The authors specified the benchmark protocols, model configurations, scoring rules, and statistical analyses, and manually verified the data-processing logic, model executions, parsed outputs, numerical results, tables, figures, and scientific claims reported in the paper.

ETHICS STATEMENT

The study uses existing chest-radiograph resources. False denial is defined relative to the reference label; source access and redistribution terms apply.

REPRODUCIBILITY STATEMENT

Code and anonymous review artifacts are available at `https://anonymous.4open.science/r/cxr-contrabench/`. The release contains protocol-construction and evaluation code, protocol manifests, item and image identifiers, inference metadata, raw-output hashes, analysis scripts, and the code used to generate reported aggregate tables and figures. Inference metadata records the evaluated checkpoint, prompt and decoder configuration, protocol and item identifiers, and model-specific image-view policy where applicable. The paired analyses retain invalid outputs rather than dropping them, and all reported confidence intervals use 2,000 bootstrap resamples over study clusters. Additionally, the repository does not redistribute licensed chest radiographs or other third-party datasets; these must be obtained from their original sources under the corresponding access and license terms.

REFERENCES

Ali Abbasi, Mehdi Taghipour, and Rahmatollah Beheshti. NAST: Improving negation handling in medical vision-language models through negation-aware selective training. *arXiv preprint arXiv:2602.12498v2*, 2026.

Kumail Alhamoud, Shaden Alshammari, Yonglong Tian, Guohao Li, Philip Torr, Yoon Kim, and Marzyeh Ghassemi. Vision-language models do not understand negation. *arXiv preprint arXiv:2501.09425*, 2025.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Pengcheng Chen, Jin Ye, Guoan Wang, Yanjun Li, Zhongying Deng, Wei Li, Tianbin Li, Haodong Duan, Ziyan Huang, Yanzhou Su, Benyou Wang, Shaoting Zhang, Bin Fu, Jianfei Cai, Bohan Zhuang, Eric J Seibel, Junjun He, and Yu Qiao. Gmai-mmbench: A comprehensive multimodal evaluation benchmark towards general medical ai. *arXiv preprint arXiv:2408.03361*, 2024a.

Zhihong Chen, Maya Varma, Justin Xu, Magdalini Paschali, Dave Van Veen, Andrew Johnston, Alaa Youssef, Louis Blankemeier, Christian Bluethgen, Stephan Altmayer, Jeya Maria Jose Valanarasu, Mohamed Siddig Eltayeb Muneer, Eduardo Pontes Reis, Joseph Paul Cohen, Cameron Olsen, Tanishq Mathew Abraham, Emily B. Tsai, Christopher F. Beaulieu, Jenia Jitsev, Sergios Gatidis, Jean-Benoit Delbrouck, Akshay S. Chaudhari, and Curtis P. Langlotz. A vision-language foundation model to enhance efficiency of chest x-ray interpretation. *arXiv preprint arXiv:2401.12208*, 2024b.

Zijie Cheng, Ariel Yuhan Ong, Siegfried K. Wagner, David A. Merle, Lie Ju, Hanyuan Zhang, et al. Understanding the robustness of vision-language models to medical image artefacts. *npj Digital Medicine*, 8:727, 2025. doi: 10.1038/s41746-025-02108-w.

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.

Ailin Deng, Tri Cao, Zhirui Chen, and Bryan Hooi. Words or vision: Do vision-language models have blind faith in text? *arXiv preprint arXiv:2503.02199*, 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

Hanbin Ko and Chang-Min Park. Bringing CLIP to the clinic: Dynamic soft labels and negation-aware learning for medical analysis. *arXiv preprint arXiv:2505.22079*, 2025.

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*, 2018.

Hyungyung Lee, Geon Choi, Jung-Oh Lee, Hangyul Yoon, Hyuk Gi Hong, and Edward Choi. Cxreasonbench: A benchmark for evaluating structured diagnostic reasoning in chest x-rays. *arXiv preprint arXiv:2505.18087*, 2025.

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. *arXiv preprint arXiv:2306.00890*, 2023.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. *arXiv preprint arXiv:2102.09542*, 2021.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. MMBench: Is your multi-modal model an all-around player? In *Computer Vision – ECCV 2024*, volume 15064 of *Lecture Notes in Computer Science*, pp. 216–233. Springer, 2024. doi: 10.1007/978-3-031-72658-3_13.

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.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Cyril Zakka, Yash Dalmia, Eduardo Pontes Reis, Pranav Rajpurkar, and Jure Leskovec. Med-flamingo: a multimodal medical few-shot learner. *arXiv preprint arXiv:2307.15189*, 2023.

Ankit Pal, Jung-Oh Lee, Xiaoman Zhang, Malaikannan Sankarasubbu, Seunghyeon Roh, Won Jung Kim, Meesun Lee, and Pranav Rajpurkar. Rexvqa: A large-scale visual question answering benchmark for generalist chest x-ray understanding. *arXiv preprint arXiv:2506.04353*, 2025.

Chantal Pellegrini, Ege Özsoy, Benjamin Busam, Nassir Navab, and Matthias Keicher. Radialog: A large vision-language model for radiology report generation and conversational assistance. *arXiv preprint arXiv:2311.18681*, 2023.

Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2006–2017, 2024. doi: 10.18653/v1/2024.findings-naacl.130.

Corentin Royer, Bjoern Menze, and Anjany Sekuboyina. Multimedeval: A benchmark and a toolkit for evaluating medical vision-language models. *arXiv preprint arXiv:2402.09262*, 2024.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Omkar Thawakar, Abdelrahman Shaker, Sahal Shaji Mullappilly, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, Jorma Laaksonen, and Fahad Shahbaz Khan. Xraygpt: Chest radiographs summarization using medical vision-language models. *arXiv preprint arXiv:2306.07971*, 2023.

Jasmine Vu and Shivanand Sheshappanavar. The effect of negation on CLIP in medical imaging: Limitations of contrastive language-image pretraining. *arXiv preprint arXiv:2512.17121*, 2025.

Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. R2gengpt: Radiology report generation with frozen llms. *arXiv preprint arXiv:2309.09812*, 2023.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Medklip: Medical knowledge enhanced language-image pre-training in radiology. *arXiv preprint arXiv:2301.02228*, 2023.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *Nature Communications*, 2025.

Peng Xia, Ze Chen, Juanxi Tian, Yangrui Gong, Ruibo Hou, Yue Xu, Zhenbang Wu, Zhiyuan Fan, Yiyang Zhou, Kangyu Zhu, Wenhao Zheng, Zhaoyang Wang, Xiao Wang, Xuchao Zhang, Chetan Bansal, Marc Niethammer, Junzhou Huang, Hongtu Zhu, Yun Li, Jimeng Sun, Zongyuan Ge, Gang Li, James Zou, and Huaxiu Yao. Cares: A comprehensive benchmark of trustworthiness in medical vision language models. *arXiv preprint arXiv:2406.06007*, 2024.

Yunfei Xie, Ce Zhou, Lang Gao, Juncheng Wu, Xianhang Li, Hong-Yu Zhou, Sheng Liu, Lei Xing, James Zou, Cihang Xie, and Yuyin Zhou. Medtrinity-25m: A large-scale multimodal dataset with multigranular annotations for medicine. *arXiv preprint arXiv:2408.02900*, 2024.

Sijie Yan, Yuanjun Xiong, Kaustav Kundu, Shuo Yang, Siqi Deng, Meng Wang, Wei Xia, and Stefano Soatto. Positive-congruent training: Towards regression-free model updates. *arXiv preprint arXiv:2011.09161*, 2020.

Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y. Ng, Curtis P. Langlotz, Vasantha Kumar Venugopal, and Pranav Rajpurkar. Evaluating progress in automatic chest X-ray radiology report generation. *Patterns*, 4(9):100802, 2023. doi: 10.1016/j.patter.2023.100802.

Giselle Zeno, Nour Jedidi, and Steven Gomez. Choosing 'right' from wrong: A closer look at selection bias in spatial multiple-choice questions in large multimodal models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 535–544, 2025.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Angela Crabtree, Brian Piening, Carlo Bifulco, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023a.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023b.

Yuhui Zhang, Yuchang Su, Yiming Liu, and Serena Yeung-Levy. NegVQA: Can vision language models understand negation? In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 3707–3716, 2025. doi: 10.18653/v1/2025.findings-acl.191.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. arXiv:2309.03882.

Yang Zhou, Tan Li Hui Faith, Yanyu Xu, Sicong Leng, Xinxing Xu, Yong Liu, and Rick Siow Mong Goh. Benchx: A unified benchmark framework for medical vision-language pretraining on chest x-rays. *arXiv preprint arXiv:2410.21969*, 2024.

A Protocol Construction and Scoring

This appendix specifies the construction and scoring contracts underlying the main experiments. We separate the primary answer-set intervention from the image-validity diagnostic because they change different inputs and answer different questions. The former holds the image fixed and changes one distractor; the latter holds the displayed text fixed and changes the visual evidence. Both protocols are fixed before model outputs are scored, and invalid attempts remain in the analysis. Figure 4 provides an overview of the data sources and evaluation protocols used throughout the benchmark. The primary 260-item CheXpert paired set is the only protocol used for row-level directional transition accounting; the remaining protocols provide complementary structural, image-validity, linguistic, external-source, scale, and error-analysis evidence with their own units and scoring rules.

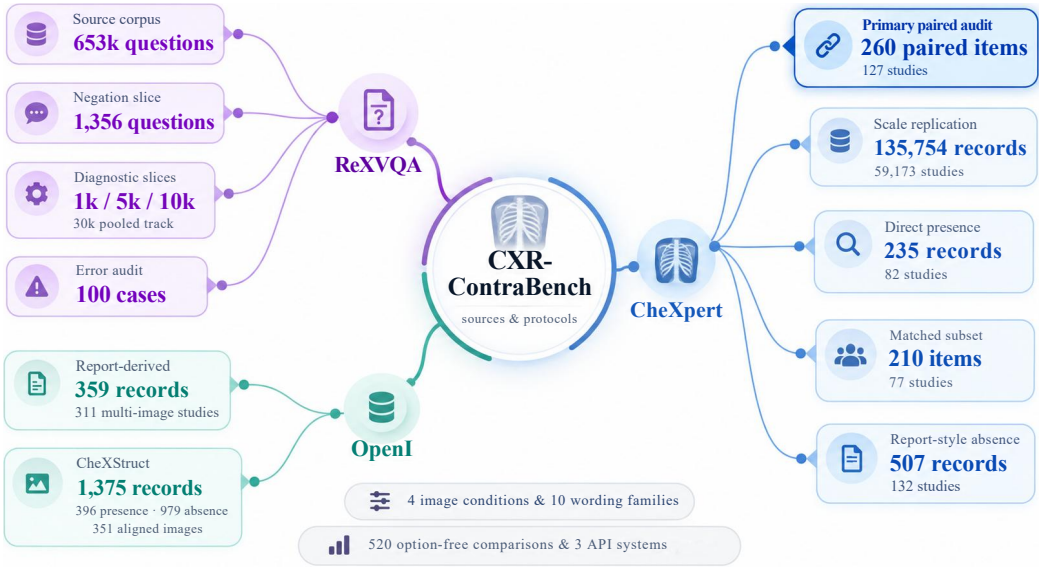


Figure 4: **Data sources and evaluation protocols in CXR-ContraBench.** The primary paired set contains 260 CheXpert units from 127 studies and supports the row-level directional transition analysis. Supporting CheXpert protocols include the 235-record direct-presence probe with its 210-record matched positive-only control, the 507-record report-style absence protocol, and a training-split scale construction with 135,754 records from 59,173 studies. OpenI provides report-derived and CheXStruct external evaluations, while ReXVQA supplies internal diagnostic slices and the clinical error audit. Counts denote the native units of each protocol and are not additive because protocols may reuse studies, images, or derived items.

A.1 Primary paired construction

The primary paired set contains 260 items from 127 CheXpert studies (Irvin et al., 2019); in this slice each study corresponds to one patient. Every item is evaluated under four arms, positive-only $[X, Y, Z]$, paired-target $[X, Y, \bar{X}]$, paired-distractor $[X, Y, \bar{Y}]$, and no-counterpart $[X, Z, \bar{Y}]$, giving 1,040 rows per model. The positive-only and paired-target arms differ by exactly one registered distractor, and the row-level directional transition accounting of the main text is computed on these two arms only. The direct-presence protocol used for the structural baselines and the prompting variants of Appendix C contains 235 records from 82 studies, with a 210-record matched positive-only control from 77 studies; the report-style absence protocol contains 507 records from 132 studies. These protocols keep their own scoring rules and are never pooled with the 260-item paired flow.

The supporting collection retains its native questions and denominators. ReXVQA (Pal et al., 2025) supplies internal 1k, 5k, and 10k slices and a pooled 30k diagnostic track, from which the clinical error audit of Appendix G samples. OpenI (Demner-Fushman et al., 2016) contributes 359 report-derived records and a CheXStruct audit (Lee et al., 2025) of 1,375 records, 396 presence and 979 absence, over 351 aligned images. A separate CheXpert training-split construction contains 135,754

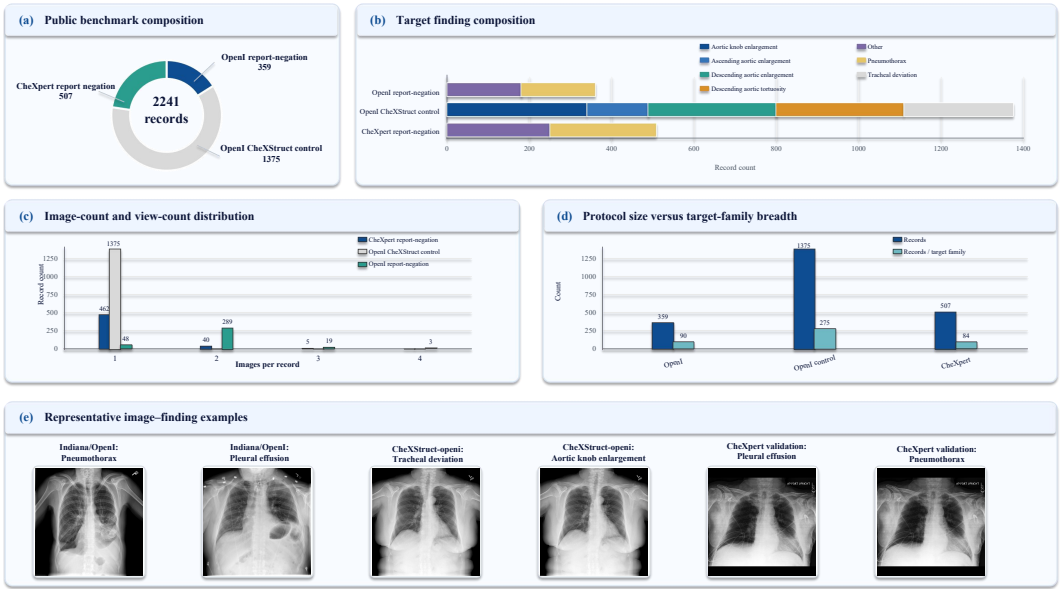


Figure 5: **Composition of the three public supporting protocols.** (a) Record counts, (b) target-finding composition, (c) images per record, (d) records per target family, and (e) representative public radiographs from OpenI and CheXpert. Panel labels use the source-protocol names, whose mapping to the protocol names of this paper is given in the text. The primary 260-unit paired set, the CheXpert presence protocols, and the training-split scale construction are not part of this figure.

records from 59,173 studies; a study can contribute several records and views, so these records are not independent patients. These collections establish breadth and scale, while the 260-item paired set alone carries the directional row-level accounting.

The three public supporting protocols are constructed as follows, and Fig. 5 summarizes their size, target-finding composition, and image-count distribution. The OpenI report-derived protocol instantiates a study-level absence question from report-extracted absent findings; its 359 records comprise 181 pneumothorax, 114 pleural effusion, 45 focal consolidation, and 19 pulmonary edema targets, and 311 of them come from multi-image studies. A positive distractor is drawn from co-mentioned present findings, so the finding concept and its negated surface form co-occur in the answer set. The OpenI CheXStruct audit uses the structured annotations released with CXReasonBench over 351 aligned images; its targets are mainly structured thoracic findings such as aortic knob enlargement, ascending and descending aortic enlargement, descending aortic tortuosity, and tracheal deviation, so its finding distribution does not overlap the CheXpert finding set. The CheXpert direct-presence protocol groups the validation split by study, requires the target finding to be cleanly present together with at least one different positive distractor, and asks which listed finding is present; 82 of 200 validation studies are constructible, and per-finding instantiation yields 235 records whose answer sets contain the target, its negated counterpart, and one other positive finding. Its matched positive-only control keeps the question but presents the present target with two absent findings expressed only in positive form; requiring a paired direct-presence item and two clean absent distractors yields 210 records from 77 studies. The report-style absence protocol requires the target to be cleanly absent and at least one other finding to be present; 132 studies are constructible, giving 507 records with 126 pneumothorax, 100 consolidation, 90 edema, 68 pleural effusion, 66 cardiomegaly, and 57 atelectasis targets. In Fig. 5, which retains the source-protocol labels, *CheXpert report-negation* denotes the report-style absence protocol, *OpenI report-negation* the report-derived protocol, and *OpenI CheXStruct control* the CheXStruct audit.

Each primary item consists of an image input, an affirmative presence question, a reference-present finding, and two reference-absent distractors. Missing and uncertain source labels remain distinct from absence. Selecting No-X when X is reference-positive is a false denial; a No-Y selection is scored by its displayed role, as a true statement about an absent finding that does not answer the present-finding question. Construction metadata is withheld from the model.

A.2 IMAGE-VALIDITY CONSTRUCTION

The image-validity diagnostic reuses the 260 items and the paired-target answer set. Every real-image row contains the literal question "Which of the following findings is present on this chest X-ray study?" and the same three lettered options. Four image conditions are evaluated: the matched study, an opposite-label donor study for which the target finding is reference-absent, a deterministic same-view shuffle that keeps the target role, and a blank image. Matched and shuffle rows require the positive target statement; opposite-label rows require the condition-specific negated statement; blank rows have no clinical gold and are unscored. Opposite-label donors must have a raw-image hash different from the matched study, and the question and option strings must be identical across conditions. The condition-specific statement rule and all real-image text are fixed before scoring. MedGemma and Qwen2.5-VL consume all available study images, whereas LLaVA-Med consumes the first image only; this view policy is declared per model. The polarity-explicit prompt audit and the LoRA audits reuse the same construction with their own prompt and checkpoint metadata, and condition-specific truth is recomputed from the image labels rather than taken from the target-role field.

A.3 PARSING AND INVALID-OUTPUT HANDLING

The three base checkpoints share one answer-set renderer and receive the complete option strings. Each inference record stores the protocol and item identifiers, study cluster, model and checkpoint, prompt and decoder identifiers, view policy, arm, question, option strings, reference role, the selected option or invalid status, and a raw-output pointer. A response enters the score only after the parser matches one displayed option; free-form text, missing letters, and out-of-set identifiers are recorded as invalid attempts and remain in the denominator. The raw response and the parsed option identifier are retained together, so surface form and semantic choice remain distinguishable. Indirect wording is scored by its full semantic scope: resolved and ruled-out statements retain their temporal or clinical meaning, and unknown or unmentioned findings retain their own semantic state. Choice outputs and likelihood-ranking outputs are stored as separate protocols. For the prompting and adaptation analyses we use the following labels: B0 is the base multiple-choice output, CoT-B0 uses a fixed handwritten reasoning prompt, three-shot uses deterministic cross-source exemplars, and LoRA is an in-format adaptation reference; these labels identify distinct execution protocols.

Backbone-specific loaders share the same protocol files and scoring code. CheXagent-2-3B is loaded with remote code, explicit attention masks, and float32 inference for stability and consumes all study images, whereas LLaVA-1.5-7B and LLaVA-Med-1.5-7B consume the first image of a study. MedGemma-27B-it runs in bfloat16 with a 64-token generation budget, because the 16-token budget used for the 4B checkpoint truncated its longer answers; its parse-failure rate is reported with its results. GPT-4o is queried through its API at temperature 0 with one image per study, and the three frontier systems of Sec. 6.3 are recorded under their served model identifiers (Appendix F).

A.4 PROTOCOL AND REPRODUCIBILITY ASSERTIONS

Before aggregation, the analysis checks item and cluster uniqueness, exact question and option text across paired arms, registered image hashes, the declared view policy, gold-slot balance, parser consistency, and the separation of scored real-image rows from blank rows. The paired answer-set protocol asserts equality of image identity, question text, option count, retained positions, and reference finding, so that the intervention changes only the registered distractor. The image protocol asserts byte equality of the question and options across conditions and distinct raw-image hashes for opposite-label donors. Complete pairs define each paired estimate; invalid records are counted as invalid attempts rather than removed.

Having fixed the construction and scoring contracts, we next quantify the uncertainty and heterogeneity of the resulting estimates.

B STATISTICAL UNCERTAINTY AND FINDING-LEVEL VARIATION

The main text emphasizes directional transitions and point estimates. Here we quantify study-cluster uncertainty and finding-level heterogeneity. Study-level resampling is used because a study can

Table 4: **Study-cluster uncertainty for the primary transition.** Percentile intervals use 2,000 bootstrap replicates over the 127 study clusters; RR conditions on control-correct rows and RF uses all 260 paired units.

Model	R/N_c	RR (%)	RF (%)	RR 95% CI	RF 95% CI
MedGemma-4B	89/158	56.33	34.23	[48.99, 63.74]	[28.15, 40.83]
Qwen2.5-VL	41/116	35.34	15.77	[26.53, 44.17]	[11.20, 20.59]
LLaVA-Med	0/129	0.00	0.00	[0.00, 0.00]	[0.00, 0.00]

Table 5: **Reversal is heterogeneous across findings and backbones.** Direct-presence protocol under direct prompting; each cell is the share of items for that finding on which the model selects the negated option.

Backbone	Atelectasis	Cardiomegaly	Consolidation	Edema	Pl. effusion	Pneumothorax
MedGemma-4B-it	11.29	**80.39**	**100.00**	**78.79**	**82.14**	**100.00**
Qwen2.5-VL-7B	**58.06**	**64.71**	**41.94**	**72.73**	**76.79**	**100.00**
LLaVA-1.5-7B	6.45	27.45	0.00	24.24	1.79	0.00
CheXagent-2-3B	1.61	0.00	3.23	0.00	8.93	0.00
LLaVA-Med-1.5-7B	0.00	0.00	0.00	0.00	0.00	0.00

contribute multiple derived items or views, while the finding-level breakdown tests whether aggregate reversal rates are concentrated in particular pathologies or model families.

For each backbone, RR conditions the correct-to-No-X count on that backbone's control-correct rows, and RF divides the same count by all 260 paired items. Table 4 reports percentile intervals from 2,000 bootstrap replicates over the 127 study clusters. The MedGemma and Qwen2.5-VL intervals exclude small effects; the LLaVA-Med interval is zero in every replicate because that backbone never selects No-X under the released parser. The complete six-state row-level counts, namely correct-to-correct, correct-to-No-X, correct-to-other-error, recovery, wrong-to-wrong, and invalid, are listed in the note to Table 1 and visualized in Fig. 2; the invalid state is empty for all three backbones on this protocol, so each net accuracy change is explained by the reversal, recovery, and other-error flows.

Table 5 breaks negated-option selection down by finding on the direct-presence protocol. Reversal is heterogeneous: MedGemma selects the negated option on every consolidation and pneumothorax item but on few atelectasis items, Qwen2.5-VL-7B selects it on a substantial share of items for every finding, and CheXagent-2-3B and LLaVA-Med-1.5-7B rarely or never select it. These finding-level differences are descriptive. They show that the aggregate rate is not driven by a single pathology, but they should not be interpreted as a universal clinical-difficulty ranking, because per-finding counts are small and the direct-presence protocol has its own 235-item denominator.

The preceding analyses hold the model configuration fixed; we next examine how prompting and adaptation change the measured behavior.

C Prompting and Adaptation

We next ask whether the measured polarity behavior changes under additional reasoning structure or adaptation to the answer format. These protocols use different prompts and, in some cases, different denominators from the primary paired set, so they are interpreted as configuration-specific diagnostics rather than pooled estimates.

Table 6 reports the direct-presence protocol (235 records) and the report-style absence protocol (507 records) across backbones, chain-of-thought prompting, three-shot prompting, and in-format LoRA. Under direct prompting, MedGemma-4B-it and Qwen2.5-VL-7B each select a negated option on 153 of 235 presence items, and the larger MedGemma-27B-it and the general-domain LLaVA-1.5-7B select it on 105 and 27 items, so the behavior is neither confined to the smallest checkpoint nor uniform across families. Chain-of-thought prompting raises accuracy for MedGemma-4B-it and Qwen2.5-VL-7B on both protocols, but it does not remove negated-option selection: 122 and 57 of

Table 6: **Negated-option selection on the direct-presence protocol and contradictions on the report-style absence protocol across backbones, prompting, and in-format adaptation.** The two protocols retain their native scoring rules and units.

		Direct presence		Report-style absence	
Backbone	**Variant**	Acc. (%)	"No X" n/235	Acc. (%)	Contrad. n/507
MedGemma-4B-it	direct	31.49	153	24.06	351
	CoT	42.98	122	32.54	282
	3-shot	29.36	160	2.17	486
	LoRA	**100.00**	0	**100.00**	0
Qwen2.5-VL-7B	direct	30.21	153	0.39	504
	CoT	54.47	57	27.22	368
	3-shot	29.79	150	28.40	354
	LoRA	**100.00**	0	**100.00**	0
MedGemma-27B-it	direct	51.91	105	31.16	335
LLaVA-1.5-7B	direct	56.60	27	27.42	348
LLaVA-Med-1.5-7B	direct	74.04	0	67.06	42
	CoT	68.09	0	11.83	446
CheXagent-2-3B	direct	37.87	7	6.51	186
GPT-4o	direct	62.55	25	2.17	470

Absence-side scaffold (`absence_step`)

Input. Question text; lettered answer options.
Fixed reasoning steps.
1. Does the question ask for an **absent** finding?
2. Which option names the **finding concept itself**?
3. Which option is the **negated surface form** ("No X")?

Decision rule. For an absence question, choose the finding name, not the negated form.
Output. Final answer: one option letter.

Presence-side scaffold (`polarity_step`)

Input. Question text; lettered answer options.
Fixed reasoning steps.
1. Does the question ask about **absence** or **presence**?
2. Which option names the **finding concept itself**?
3. Which option is the **negated surface form** ("No X")?

Decision rule. For an absence question, choose the finding name rather than the negated form; for a presence question, choose the present finding and avoid the negated form.
Output. Final answer: one option letter.

Figure 6: **Fixed chain-of-thought scaffolds of the CoT rows in Table 6.** Both scaffolds share one polarity-disambiguation skeleton; the presence-side scaffold adds an explicit branch between absence and presence questions before the final answer.

235 presence items still receive a negated option. Three-shot prompting with deterministic cross-source exemplars leaves negated-option selection on the presence protocol essentially unchanged for both backbones (160 and 150 of 235, against 153 under direct prompting) and sharply increases contradictions for MedGemma on the absence protocol, so few-shot exemplars are not a reliable remedy on these endpoints. The three-shot exemplar pools are deliberately cross-source: excluding the 82 direct-presence and 132 absence-side evaluation studies leaves no constructible held-out CheXpert validation examples, so the pools consist of 20 CheXStruct presence examples and 20 OpenI report-style absence examples, from which three polarity-correct exemplars are sampled deterministically per test call. Any gain therefore cannot be attributed to same-distribution memorization. The absence-protocol columns are exact-match diagnostics of answer format and are reported separately from the presence-side false-denial rate; a contradiction count on the absence side is not evidence about the primary transition.

The CoT rows of Table 6 use the fixed handwritten scaffolds shown in Fig. 6. The direct-presence protocol uses a scaffold that branches on question polarity, and the absence protocols use an absence-targeted scaffold; both require the model to name the question polarity, the finding concept, and the negated surface form before it commits to a single answer letter. The polarity-explicit, evidence-first, and image-first prompts of Table 2 are separate configurations with their own prompt metadata.

In-format LoRA adapters reach 100% on both protocols for both backbones and select the negated option on no item. Each adapter has rank 16 and is trained on a balanced 4,000-example CheXpert

Table 7: **In-format LoRA transfers to unseen wordings and stops selecting the off-target negated option.** Cells are base → adapted; counts are adapted-model correct over items.

	Accuracy (%)			Selection (%)
Adapter	Indirect wordings	Scope wordings	$[X, Z, \bar{Y}]$	$\bar{Y}$ under $[X, Z, \bar{Y}]$
MedGemma-4B + LoRA	12.74 → **87.26** 849/973	19.04 → **66.15** 344/520	11.92 → **54.23**	85.00 → **0.77** 2/260
Qwen2.5-VL + LoRA	47.28 → **94.04** 915/973	15.96 → **54.81** 285/520	25.00 → **48.46**	60.38 → **0.77** 2/260

training-split set, 2,000 presence and 2,000 absence questions, that uses the same question templates and answer-set layout as the evaluation protocols and shares no study with any validation evaluation study. Final training losses are near zero, 1.2×10^{-5} for MedGemma and 3×10^{-6} for Qwen2.5-VL, so the adapters fit the format almost completely; zero study overlap rules out image-level leakage but not format-specific adaptation. Table 7 tests whether this gain is confined to the adaptation format. It is not: accuracy on indirect wording rises from 12.74% to 87.26% for MedGemma and from 47.28% to 94.04% for Qwen2.5-VL, scope wording and the no-counterpart set also improve, and selection of the off-target negated option under $[X, Z, \bar{Y}]$ falls to 2 of 260 rows for both adapters. The adapters therefore transfer to several wordings and answer structures they never saw.

Transfer across wording and structure is a different question from image validity. Under the image-validity diagnostic of Sec. 5.2, both adapted checkpoints are statement-correct on every matched row and on no opposite-label row, giving 0% PairCorrect, and the polarity-explicit Qwen2.5-VL run that lowers the conditional reversal rate to 2.00% contains 346 of 1,040 unparseable outputs and 8 of 260 PairCorrect cases, with 143 of 260 matched, 23 of 260 opposite-label, and 130 of 260 shuffle rows statement-correct. High in-format accuracy and successful wording transfer therefore do not by themselves establish image-conditioned correctness, and the two endpoints should be reported together.

These interventions change the model configuration; the next section instead holds the model fixed and varies how negation is expressed.

D Negation Wording and Option-Free Generation

The canonical "No X" form makes the polarity transition easy to audit, but it raises two separate questions: whether the behavior survives alternative wording and whether it survives removal of candidate answers. We address these questions with explicit, indirect, and scope-level negation families and with single-sentence option-free generation.

Table 8 lists the ten wording families. The four explicit families contain a literal negation token ("No", "Absence of", "is not present", "No evidence of"). The four indirect families contain no "No" token and instead state clearance, exclusion, resolution, or a normal description. The two scope families negate several findings at once or the entire option list. Each family replaces the third option by the listed form of the present finding, so the item, image, question, and retained options are unchanged. Group-level reversal rates are reported in Table 3(a): indirect families reach 80.27% for MedGemma, and scope families reach 70.89% and 70.69% for MedGemma and Qwen2.5-VL, so the transition does not depend on the literal "No" token.

Table 9 applies the four explicit wordings to GPT-4o, each compared with its own repeated positive-only control on the same items. GPT-4o reverses under every wording, from 5.00% under "No X" to 19.17% under "Absence of X", while its accuracy rises under every one. The wording-by-condition interaction shows that the reversal rate of a single surface form should not be generalized to negation as a whole, and the coexistence of reversals with accuracy gains reproduces, for an API model, the pattern seen for Qwen2.5-VL in the primary analysis.

Table 10 is a secondary absence-side diagnostic. It rewrites the question of the report-style absence protocol three times and records MedGemma-4B-it's exact-match accuracy and the number of selected statements that contradict the protocol's reference. Accuracy and contradiction counts remain in the same range under the paraphrases, so the absence-side behavior is not an artifact of one question

Table 8: **Ten negation wording families.** Each family replaces the third option by the listed form of the present finding; the example uses edema. Reversal rates by family group are in Table 3(a).

Family	Example option string
Explicit, 4 families	
Bare negation	No edema
Nominal absence	Absence of edema
Predicate negation	Edema is not present
Evidential negation	No evidence of edema
Indirect, 4 families, no "No" token	
Clear-of	Lungs are clear of edema
Ruled-out	Edema has been ruled out
Resolved	Previously noted edema has resolved
Normal description	Clear lung fields; Normal heart size; Sharp costophrenic angles
Scope, 2 families	
Compound negation	No cardiomegaly or consolidation
Universal negation	None of the listed findings are present

Table 9: **GPT-4o reverses under every explicit wording while its accuracy rises under every one.** Each wording is compared with its own repeated positive-only control on the same items.

Wording	ΔAcc. (pp)	RR (%) R/N_c	95% CI
No X	+32.31	5.00 6/120	[1.64, 9.28]
Absence of X	+17.69	19.17 23/120	[12.60, 26.05]
X is not present	+20.77	11.40 13/114	[5.93, 17.65]
No evidence of X	+32.69	9.32 11/118	[4.39, 14.78]
Pooled	+25.87	11.23 53/472	[7.58, 15.40]

All four accuracy gains survive Holm correction; the wording-by-condition interaction is significant. The four control accuracies span 2.31 points, and four repeated controls agree exactly on 75.77% of items.

template. This protocol scores answer format on the absence side; its contradiction counts are not evidence for the presence-side false-denial rate.

The option-free track removes candidate answers. For each of 260 confirmed-positive image–finding records, the model generates one clinical sentence under each member of a matched question pair, present versus not present and evidence versus no evidence, giving 520 comparisons per model. A fixed rule parser records affirmation, denial, other, and invalid outputs, and a fixed text classifier provides an independent semantic check. A comparison counts as affirmation-to-denial when the model affirms the finding under the first question and denies it under the negatively cued counterpart. All attempts remain in the denominator, including uncertain and unparseable outputs. MedGemma changes from affirmation to denial in 361 of 520 comparisons and Qwen2.5-VL in 238 of 520; LLaVA-Med differs sharply between the two question pairs, so both of its native rates are shown in Table 3(a). This track establishes that the denial of a confirmed-present finding is not tied to the presence of a negated candidate answer. It changes the question wording, generates one sentence rather than a report, and is not the same intervention as the fixed-question paired protocol.

Linguistic robustness does not remove answer-set structure, so we next isolate the structural ceiling directly.

E Structural Controls

The target-paired construction intentionally places a finding and its negated surface form in the same answer set, which creates an image-free structural prior. We therefore separate two quantities: how often a model selects a negated option, and whether mapping that selection to its affirmative counterpart would be correct under the protocol.

Table 10: **The absence-side surface-copy behavior is stable across question paraphrases.** MedGemma-4B-it on the report-style absence protocol; a contradiction is a selected statement that conflicts with the protocol's reference.

Question wording	Acc. (%)	Contradictions n/507
Original	23.08	356
Paraphrase 1	22.68	283
Paraphrase 2	18.93	381
Paraphrase 3	24.65	379

Table 11: **Remapping a negated selection to its counterpart is correct when the construction makes the counterpart the gold.** Selected is how often the model chooses the negated option; correct remaps are over triggered remaps.

		MedGemma-4B-it		**Qwen2.5-VL-7B**	
Option set	**Negated concept**	Selected (%)	Correct remaps	Selected (%)	Correct remaps
$[X, Y, \bar{X}]$	present finding X	68.85 179/260	100.00 179/179	47.31 123/260	100.00 123/123
$[X, Y, \bar{Y}]$	absent finding Y	86.92 226/260	0.00 0/226	63.46 165/260	0.00 0/165
$[X, Z, \bar{Y}]$	absent Y, no counterpart	85.00 221/260	rule abstains	60.38 157/260	rule abstains

Three cases arise. When the negated option is the target (No-X under $[X, Y, \bar{X}]$), its affirmative counterpart is the gold answer by construction, so a remap is always correct. When the negated option negates a non-target (No-Y under $[X, Y, \bar{Y}]$), its counterpart Y is reference-absent, so a remap is never correct. When no affirmative counterpart is displayed ($[X, Z, \bar{Y}]$), a deterministic rule must abstain. Table 11 reports both quantities on the 260-item set. MedGemma selects the negated option on 68.85%, 86.92%, and 85.00% of rows across the three arms and Qwen2.5-VL on 47.31%, 63.46%, and 60.38%, so selection is high whether or not the negated statement is true. Remap correctness follows the construction rather than the model: 179/179 and 123/123 on the target arm, 0/226 and 0/165 on the non-target arm.

The same structure explains the remapping results on the 235-item direct-presence protocol. There, the option-only rule that always selects the affirmative member of the displayed X/No-X pair is correct on 235/235 items without reading the image or the model output. Conditional remapping, which changes only a model prediction of No-X to X, reaches 96.60% for MedGemma and 95.32% for Qwen2.5-VL on that construction; on the OpenI CheXStruct presence subset, MedGemma selects a negated option in 270 of 279 wrong responses, and 197 of those selections satisfy the unique-counterpart condition that such a rule requires. These scores measure how often the construction makes the counterpart correct, not whether the model assessed the image. These controls motivate treating deterministic remapping as a construction baseline rather than as independent visual inference.

After separating construction-dependent effects from model behavior, we next examine transfer across data sources and model families.

F External and Frontier-Model Evidence

The primary paired analysis focuses on locally evaluated backbones. We test breadth along two independent axes: data source and model family. OpenI and the large CheXpert construction evaluate source and scale, while the frontier-model study evaluates newer API-based multimodal systems under the same paired principle. Their denominators and execution environments are therefore reported separately.

On the 396-record OpenI CheXStruct presence subset, the reported statistic is the share of a model's wrong answers that are negated options: 96.77% for MedGemma, that is 270 of 279 wrong responses, 92.58% for Qwen2.5-VL, 41.61% for LLaVA-1.5-7B, and 21.19% for LLaVA-Med. The 979-record absence subset and the 359-record report-derived OpenI set retain their original exact-match rules and are not combined with the presence statistic. The CheXpert training-split construction with

Table 12: **Frontier models gain accuracy under the negated set because recoveries outnumber reversals.** Complements Table 3(b), which reports intervals, other wordings, and blank-image change.

	Accuracy (%)			Flows	
Model	Control	Negated	Δ	RR (%) R/N_c	Recovered G/N_w
claude-opus-5	47.69	68.08	+20.38	14.52 18/124	56.62 77/136
gpt-5.5	58.85	66.15	+7.31	13.73 21/153	48.60 52/107
gemini-3.6-flash	48.46	70.00	+21.54	7.14 9/126	52.24 70/134

Table 13: **Polarity confusion is the largest single category in a radiologist-facing audit of 100 MedGemma errors.** Errors are sampled from a ReXVQA slice; categories are mutually exclusive.

Category	Cases	Category	Cases
Presence versus absence	**31**	Location or laterality	11
Finding identity	20	Severity or interval change	4
Negation or question scope	18	Granularity or specificity	3
Differential diagnosis	12	Measurement or geometry	1

135,754 records from 59,173 studies reports presence-side negated-option selection above 62% for both MedGemma and Qwen2.5-VL. Because this construction scores selections against source labels rather than against a paired control answer, it is a selection rate and not a transition rate, and its records are not independent studies or patients.

Table 12 reports the three frontier systems under their served identifiers, claude-opus-5, gpt-5.5, and gemini-3.6-flash, on the paired control and negated answer sets. All three gain accuracy under the negated set because recoveries outnumber reversals, and all three still convert between 7.14% and 14.52% of their control-correct answers into No-X, with study-cluster intervals that exclude zero. Table 3(b) adds the reversal range under the other explicit wordings and the fraction of answers that change when the image is blank. Each system has its own control-correct denominator, and percentages with different denominators are not compared as if they were one metric; the table is a breadth check, not a leaderboard. API model versions are recorded at run time but cannot be pinned by us, so these rows are tied to the recorded identifiers.

Finally, we place the targeted polarity error in a broader clinical error taxonomy and document the artifacts required for reproduction.

G Clinical Error Audit and Reproducibility

Negated-option attraction captures one directional failure rather than the full error space of chest X-ray question answering. The clinical error audit places that failure in a broader taxonomy, while the reproducibility subsection documents the artifacts required to reconstruct the protocols and analyses.

Table 13 partitions 100 sampled MedGemma errors from a ReXVQA slice into mutually exclusive, radiologist-facing categories. Presence-versus-absence confusion is the largest single category with 31 cases, and negation or question-scope errors add 18, so polarity-related mistakes account for about half of the audited errors. The remaining errors concern finding identity, differential diagnosis, location or laterality, severity or interval change, granularity, and measurement or geometry. Polarity errors are therefore important but coexist with several other error types, consistent with clinically informed evaluations in which radiologists score distinct error categories, such as false findings, omissions, and incorrect location or severity, according to their clinical significance rather than under one aggregate score (Yu et al., 2023). CXR-ContraBench is not a complete clinical error taxonomy; it targets the directional failure that a paired answer-set intervention can isolate.

G.1 COMPUTE AND REPRODUCIBILITY

The released artifact contains protocol manifests, item and image identifiers, raw-output hashes, analysis scripts, and generated tables and figures. Exact numerical reproduction additionally requires the listed checkpoints, prompts, decoders, source data, and runtime versions. All reported bootstrap intervals use 2,000 study-cluster resamples.

Data and model licenses. ReXVQA is distributed under a non-commercial data access and use agreement (Pal et al., 2025); CheXStruct, used through the CXReasonBench release, is distributed under CC BY 4.0 (Lee et al., 2025); CheXpert and OpenI retain their original dataset terms (Irvin et al., 2019; Demner-Fushman et al., 2016); and MedGemma is provided under the Health AI Developer Foundations license (Sellergren et al., 2025). The release references these sources rather than redistributing raw images, and all raw data remain subject to the licenses and access controls of the original datasets.

Taken together, the appendix separates construction validity, statistical uncertainty, intervention sensitivity, linguistic breadth, structural shortcuts, external transfer, and broader clinical error context rather than treating them as interchangeable evidence.