EVIDENCE THAT CAN CONTRADICT: WHAT REPORT-AUDITING METRICS ACTUALLY MEASURE IN MEDICAL VISION LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

As vision-language models (VLMs) begin drafting radiology reports in clinical settings, their safety relies on audit signals that flag which statements require physician review. The community typically evaluates these signals by pooling all generated claims into a single ranking and reporting the AUROC. We show that this standard metric is dominated by a prior that requires no visual information at all. Pooled claim-level AUROC decomposes exactly into within-concept and between-concept comparisons; on MIMIC-CXR, 90% to 94% of the pairs it scores compare *different* concepts, where the ordering is largely settled by knowing which findings are generally stated correctly. We derive a *prior ceiling*—the maximum score achievable by any predictor constant within a (concept, polarity) group, computable directly from claim counts without a model or an image. Across five generators from 61M to 4.3B parameters, this ceiling exceeds the strongest visual audit signals, including a recently published black-box auditor, by 0.11 to 0.22 AUROC. On the standard CheXbert-14 evaluation, the between-condition share is 93.4% to 93.6% and the ceiling is 0.771 to 0.792, so reported pooled AUROCs should be read against this ceiling, not 0.5. Because the ceiling is an oracle, we also rank claims by training base rates alone; the visual audit signals beat this realisable baseline, but by a surprisingly narrow margin. We propose a stratified metric that removes this prior, under which the base-rate predictor scores exactly 0.500, and show that it reorders current auditing methods.

1 INTRODUCTION

Radiology report generation is rapidly moving from research to practice; in fact, radiology accounts for roughly three-quarters of AI-enabled medical devices authorized by the US FDA (U.S. Food and Drug Administration, 2026). The clinical viability of these generative models hinges on safety. Because vision-language models (VLMs) are prone to hallucination, a drafted report is only useful if accompanied by a reliable audit signal indicating which statements a physician actually needs to verify. This clinical imperative has driven a surge of research into confidence and audit mechanisms for generated claims—ranging from image-text similarity in concept-bottleneck models (Han et al., 2026) to calibrated uncertainty estimation (Bani-Harouni et al., 2026; Wang et al., 2025) and sampling-based hallucination detection (Zhang et al., 2024).

Despite the architectural diversity of these approaches, the community has largely converged on a single evaluation protocol to determine their clinical readiness: pooling every claim generated by the model into a unified ranking, ordered by the audit's confidence, and reporting the Area Under the Receiver Operating Characteristic curve (AUROC). If this metric is flawed, we risk deploying auditing tools that fail to flag actual clinical errors, directly compromising patient safety.

Here, we ask what this pooled AUROC actually measures. Decomposing it into its within- and between-concept components reveals that it is dominated by cross-concept comparisons, which can be ordered without looking at the image at all. We make this consequence precise by deriving a *prior ceiling*: the highest score reachable by any predictor that remains constant within a (concept, polarity) group. This ceiling is computable from claim counts alone, and on every generator we study, it comfortably exceeds the strongest audit signal we test.

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.

We need to be careful about what this implies. The ceiling is an oracle—it requires knowing the per-concept error counts on the test set—so it bounds what the metric *can* reward, rather than describing a baseline someone could actually deploy. Its realisable counterpart, which ranks claims by training base rates alone, is weaker, and the audit signals beat it. The point is not that visual evidence is worthless, but that the pooled metric inflates its apparent value by heavily rewarding dataset priors.

A brief note on scope: apart from RadFlag (Zhang et al., 2024), a published black-box auditor we run on all five generators, the audit signals we evaluate are baselines we implemented ourselves—two similarity variants, a supervised probe, and the world-model instrument. Our primary object of study is the evaluation protocol, and our claims focus on what that protocol measures across these signals.

Core Contributions.

1. **The prior ceiling, and what it does and does not bound.** We apply a standard within/between partition of AUROC to claim-level auditing and derive the *prior ceiling*: the maximum score reachable by any predictor constant within a (concept, polarity) group. Because it is computable from claim counts alone, it can be checked for any published setup whose claim distribution is known. It exceeds every audit signal we test by 0.11 to 0.22 across all five generators. We carefully separate this oracle bound from its realisable counterpart, and we instantiate the ceiling on the standard CheXbert-14 evaluation so it can be applied outside our specific concept bank (Section 4.1).
2. **A utility check and a corrected metric.** Across five generators spanning 61M to 4.3B parameters, we measure claim triage—the fraction of a generator's errors caught in its least-confident $k\%$ of claims—against the base-rate predictor. On four of the generators, the audit signals fail to beat this image-free baseline at any review fraction that saves work. We then introduce a stratified metric that removes the prior term, under which the base-rate predictor scores exactly 0.500, and report the corrected method ordering, which differs from the pooled ordering on every generator (Sections 4.2 and 4.4).
3. **Evidence that the metric is blind to spatial attribution.** By extending CheXWorld's action space from spatial transforms to clinical statements, we create an instrument whose region-dependence can be varied while holding its accuracy fixed. Two configurations reach indistinguishable AUROCs while differing roughly five-fold in how much they depend on the finding's actual anatomy. Furthermore, the similarity signal used by concept-bottleneck auditors shows no measurable dependence on the anatomy at all. We report this to support our argument about the metric's blind spots, rather than as a separate claim about attribution (Section 4.3).

2 Related Work

Auditing Generated Medical Reports. Concept-bottleneck systems for radiology score a report's claims by measuring the similarity between the image and a concept name (Han et al., 2026). ConRad (Bani-Harouni et al., 2026) trains a model to verbalize calibrated confidence per sentence; CURV (Wang et al., 2025) couples uncertainty to a reasoning chain; RadFlag (Zhang et al., 2024) flags likely hallucinations from sampling behaviour, and MAIRA-2 (Bannur et al., 2024) grounds a model's own output as it is produced. While these approaches differ architecturally, they are uniformly evaluated by pooling claims and reporting AUROC. That shared protocol is our focus.

Crucially, these systems operate under different constraints, which dictates which ones we can run. ConRad and CURV are *generation-time*: they train the generator itself to express uncertainty, meaning they cannot audit a fixed report from a third-party model—the exact deployment setting we study here. RadFlag is post-hoc and black-box, re-sampling the generator to see if a claim recurs, so it applies to any system that can be sampled; we evaluate it alongside our own signals. The rest are baselines we implement, which bounds our conclusions: we are evaluating what the pooled protocol measures on the signals we can run, not how every published system would rank under it.

Visual Grounding and Attribution. Recent work has begun questioning whether medical VLMs actually utilize visual features. Zafar et al. (2026) replace images with blank, shuffled, and hard-negative surrogates to quantify visual reliance. Sadanandan & Behzadan (2026) occlude 16×16 patches and compare the resulting causal maps against attention weights using shifted and random control boxes. While both studies ask whether a model's stated explanation is faithful, we use a

similar family of perturbations in Section 4.3 for a different purpose: to show that two configurations of the same evidence model can be equally accurate while differing several-fold in region dependence, proving that a reported accuracy does not constrain where the signal reads.

Negation in Vision-Language Models. Alhamoud et al. (2025) demonstrate that VLMs systematically fail to understand negation, often performing at chance. This is highly relevant to radiology, where the majority of statements are negative (e.g., "no pleural effusion"). This phenomenon explains why raw cosine similarity performs poorly in Table 4. We leverage this insight to construct a fairer, differential similarity baseline rather than evaluating against a known weak formulation.

Prior-Dominated Metrics and Grouped Ranking. The observation that a pooled metric can be carried by a prior is not new, nor is the partition we use to show it. In visual question answering, question-only models that never see the image score close to full systems, which motivated rebalanced benchmarks and prior-robust splits (Goyal et al., 2017; Agrawal et al., 2018); the structure of the problem—a strong label prior conditioned on the textual side—is identical to what we find here. On the metric side, AUROC is a rank statistic over pairs, so partitioning those pairs by group and separating within-group from cross-group terms is standard practice. Kallus & Zhou (2019) decompose bipartite ranking loss in exactly this way to expose cross-group disparities that a pooled AUC hides. Equation (6) is simply that partition applied to claim-level auditing. What we add is the consequence for this specific setting: the *prior ceiling* of Proposition 2, which is computable from claim counts alone and can be evaluated for a published result without access to its model, alongside its instantiation on the CheXbert-14 protocol that report-generation papers already use.

Evidence Models and Evaluation. Our world-model evidence extends CheXWorld (Yue et al., 2025), which belongs to the Joint-Embedding Predictive Architecture (JEPA) family (Assran et al., 2023; Garrido et al., 2024), adapting it from spatial actions to clinical statements. In this work, it serves as an analytical instrument rather than a proposed clinical system, specifically enabling the controlled experiments in Section 4.3. Reports are scored using RadGraph entity F_1 (Jain et al., 2021) and CheXbert labels (Smit et al., 2020). We explicitly avoid LLM-based metrics (Ostmeier et al., 2024) to ensure the evaluation target remains deterministic across hundreds of experimental runs.

3 METHODOLOGY

We first formalize the evaluation protocol and the evidence sources used to score claims. Crucially, we introduce a clinical-action world model designed specifically as an analytical instrument to control spatial attribution.

3.1 TASK FORMULATION AND EVALUATION PROTOCOL

Task Definition. As illustrated in Figure 1b, a generator produces a findings section for a given chest radiograph. We parse this output using RadGraph (Jain et al., 2021) into discrete claims (c, π), where c is a clinical concept from a fixed bank and $\pi \in \{+, -\}$ is the polarity (present or absent). A claim is deemed *correct* if the reference report asserts the identical polarity. Unmentioned reference concepts default to absent, and generated claims mapping outside the 66-concept bank are excluded. An audit mechanism assigns a confidence score to each claim; the metric under study ranks all claims by this confidence and reports the AUROC against correctness.

Concept Bank. The concept bank is mined from $60{,}000$ MIMIC-CXR training reports. Polarity is excluded from concept identity (e.g., "effusion" and "no effusion" map to the same c with opposite π). To ensure robust two-sided comparisons, a concept is included only if it is asserted in at least 50 training studies *and* negated in at least 50. This filtering removes findings that radiologists only state in the positive, resulting in $C = 66$ concepts derived from $541{,}051$ observation mentions. Notably, by removing concepts that carry a near 100% deterministic prior, this filtering represents a conservative choice that likely *understates* the true severity of the prior-domination effect.

Datasets. We utilize MIMIC-CXR (Johnson et al., 2019) under its official patient-level split, using one frontal image per study: $140{,}571$ training, $1{,}099$ validation, and $2{,}018$ test studies. Evidence models are fitted on a fixed $60{,}000$-study training subset. IU X-Ray (Demner-Fushman et al., 2016), under the split of Chen et al. (2020), is held out entirely to evaluate zero-shot generalization.

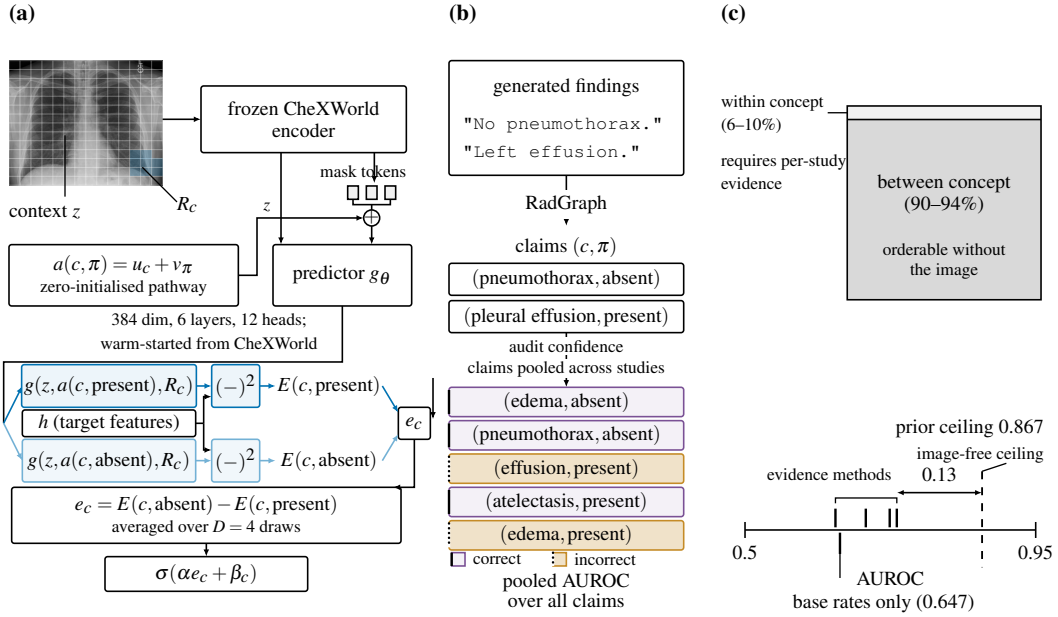

Figure 1: Overview of the evaluation paradigm and our analytical instrument. **(a)** The clinical-action world model, which predicts hidden region features conditioned on a clinical statement. The evidence e_c is the difference in prediction error between polarities; under a Gaussian model its calibration family is exact, and our evidence fits it closely (Proposition 1). **(b)** The standard evaluation protocol: generated findings are parsed into claims, pooled across all studies, and ranked by audit confidence to compute a single AUROC. **(c)** On DiA-CLEAR, 93% of the pooled weight compares different concepts, which can be ordered without the image. The dashed line is the concept-only ceiling (0.867); grouping by (concept, polarity) raises this ceiling to 0.938. Every signal falls below it, but it is an oracle bound, and a realisable base-rate predictor (0.647) is beaten (Table 1).

Generators. Table 3 details the five systems evaluated in this study. None are contributions of this paper, and all are audited without modification. Two are variational vision-language autoencoders trained from published recipes: DiA (Shaik et al., 2026), and DiA-CLEAR, which replaces DiA's alignment term with the concept-grounding objective of Han et al. (2026). The remaining three are public checkpoints used out-of-the-box. MedGemma-4b-it utilizes a narrower vocabulary (37 concepts vs. 50–54), which impacts the triage analysis in Section 4.2. Every generator receives a single frontal image and decodes greedily, with *no* indication section or prior study provided, preventing data leakage into the generation benchmark.

3.2 Evidence Sources and the World Model Instrument

We evaluate four distinct signals, all scoring the identical set of claims: 1) **Cosine similarity** between the image embedding (from the image and text towers of DiA-CLEAR's contrastive concept-grounding stage: an EfficientNet-B4 image encoder and a Transformer text encoder trained with a symmetric InfoNCE loss on the MIMIC-CXR training split) and the concept name (the incumbent standard). 2) A **differential cosine similarity** formulated as $\cos(v, c) - \cos(v, \text{“no } c\text{”})$, which recovers polarity information at the cost of one extra matrix product. 3) An **attentive probe** trained on frozen CheXWorld features, supervised directly on the evaluation labels, serving as a discriminative upper bound. 4) A **clinical-action world model**, constructed specifically to allow us to vary a signal's dependence on anatomical regions while holding accuracy fixed.

The Action Space of the World Model. As shown in Figure 1a, CheXWorld (Yue et al., 2025) predicts the latent features of a hidden region of a radiograph from visible context, conditioned on an *action* that explains how the two relate. We retain the predictor but replace the spatial action with a clinical statement about the hidden region: a pair (c, π) read as "concept c is present/absent."

Let z denote the encoder's context tokens, R_c the anatomical region occupied by concept c (mapped via fixed anatomical bounding boxes), $h \in \mathbb{R}^{N_t \times d}$ the frozen encoder's target features on that region, and $a(c, \pi) = u_c + v_\pi$ the action embedding. The predictor g_θ forecasts the region under the statement, and we score the statement by its prediction error:

$$E(c, \pi) \;=\; \tfrac{1}{N_t d} \left\| g_\theta(z,\, a(c, \pi),\, R_c) - h \right\|^2. \tag{1}$$

Because the action is fused into every mask token, the prediction is a direct function of the clinical statement.

Evidence as a Likelihood Ratio. The evidence for concept c is defined as the difference in prediction error between the two polarities, averaged over $D = 4$ deterministic spatial draws of the masked region:

$$e_c \;=\; \frac{1}{D} \sum_{d=1}^{D} \big[\, E_d(c, -) - E_d(c, +) \,\big]. \tag{2}$$

This isn't just an intuitive heuristic; as Proposition 1 shows, it emerges naturally from a Gaussian predictive model.

Proposition 1 (Evidence is a log Bayes factor)**.** *Assume the predictive distribution of the hidden region is Gaussian with isotropic residual variance τ^2, i.e., $h \mid z, (c, \pi) \sim \mathcal{N}\big(g_\theta(z, a(c, \pi), R_c),\, \tau^2 I\big)$. Then:*

$$e_c \;=\; \frac{2\tau^2}{N_t d} \log \mathrm{BF}_c, \qquad \mathrm{BF}_c \;=\; \frac{p(h \mid z, (c, +))}{p(h \mid z, (c, -))}, \tag{3}$$

meaning e_c is an affine function of the log Bayes factor. Consequently:

$$\Pr[(c, +) \mid z, h] \;=\; \sigma(\alpha\, e_c + \beta_c)\,, \qquad \alpha = \frac{N_t d}{2\tau^2}, \tag{4}$$

where β_c is the log prior odds of c. Under this model the sigmoid we fit is the exact posterior family; our evidence satisfies the model closely but not exactly (Section B).

The proof is provided in Section B. Crucially, because β_c acts as a per-concept intercept, the calibrated score explicitly carries the concept prior—the exact term that Section 4.1 demonstrates dominates the pooled metric. This calibration aligns with standard Platt scaling (Platt, 1999).

Training and Polarity Contrast. Training minimizes Equation (1) on the true statement. To prevent the predictor from ignoring the action (which drives e_c to zero), we introduce a polarity-contrastive term on the relative gap between the true statement and its polarity flip:

$$\mathcal{L} \;=\; E(c, \pi_{\text{true}}) \;+\; \lambda\, \mathrm{softplus}(-\kappa\, r)\,, \qquad r \;=\; \frac{E(c, \bar{\pi}) - E(c, \pi_{\text{true}})}{\mathrm{sg}[E(c, \pi_{\text{true}})]}, \tag{5}$$

where $\bar{\pi}$ is the flipped polarity, $\mathrm{sg}[\cdot]$ stops the gradient, and κ is an inverse temperature. The weight λ is the critical hyperparameter in this study: as shown in Section 4.3, varying λ allows us to drastically alter the model's reliance on anatomical regions without affecting its overall AUROC.

4 Experiments & Discussion

We now put these theoretical observations to the test. We first decompose the standard AUROC metric to reveal its reliance on dataset priors. We then evaluate the practical utility of these metrics via claim triage, demonstrate the spatial blindness of accuracy metrics using our world model instrument, and finally propose a stratified metric that removes the prior term.

4.1 Pooled AUROC is Dominated by Concept Priors

Let n_c^+ and n_c^- denote the number of correct and incorrect claims for concept c. Because AUROC represents the fraction of correctly ordered (correct, incorrect) pairs, we can partition these pairs based on whether the two claims share the same concept. This yields an exact identity:

$$\mathrm{AUROC}_{\text{pooled}} \;=\; w_{\text{in}} \cdot \mathrm{AUROC}_{\text{within}} \;+\; w_{\text{bet}} \cdot \mathrm{AUROC}_{\text{between}}, \qquad w_{\text{in}} = \frac{\sum_c n_c^+ n_c^-}{\big(\sum_c n_c^+\big)\big(\sum_c n_c^-\big)}. \tag{6}$$

Table 1: The decomposition of Equation (6), per generator, on MIMIC-CXR test. *Between* is w_{bet}. *Base-rate* ranks claims by training base rate, which depends on the stated polarity, so it is constant within (concept, polarity): the ceiling that bounds it is the right-hand one, and it may exceed the concept-only ceiling. The two ceilings are Proposition 2 under its two groupings. *Best* is the strongest of our signals under the pooled metric; the last column is its distance to the (concept, polarity) ceiling. The final row (bdry) evaluates DiA-CLEAR zero-shot on IU X-Ray, where the gap to the ceiling vanishes under distribution shift.

			ceiling (Proposition 2)			gap to
Generator	between	base-rate	concept	(conc.,pol.)	best	ceiling
DiA-CLEAR	93.0%	0.647	0.867	0.938	0.735	−0.203
DiA	90.2%	0.676	0.845	0.926	0.744	−0.182
ViT–BERT	94.4%	0.631	0.886	0.938	0.721	−0.217
CheXagent	94.1%	0.674	0.888	0.939	0.761	−0.178
MedGemma	93.0%	0.659	0.938	0.947	0.833	−0.114
DiA-CLEAR/IU *(bdry)*	93.8%	0.785	0.968	0.983	0.982	−0.001

The between-concept term can be maximized without any visual information; it merely requires knowing which concepts are generally stated correctly. Conversely, the within-(concept, polarity) term requires per-study visual evidence, as all claims within a specific concept and polarity share the same base rate. As summarized in Figure 1c, Equation (6) reveals that 90% to 94% of the metric's weight is placed on the between-concept term—a question that does not require an image to answer.

Proposition 2 (Prior Ceiling)**.** *Partition the claims into groups g, and let n_g^+, n_g^- be the number of correct and incorrect claims in group g. Among all confidence scores that are constant within each group, the maximum pooled AUROC is attained by ordering groups by their odds of correctness n_g^+/n_g^-, and equals*

$$\frac{\frac{1}{2}\sum_g n_g^+ n_g^- + \sum_{g \prec g'} n_g^+ n_{g'}^-}{\left(\sum_g n_g^+\right)\left(\sum_g n_g^-\right)}$$

where $\prec$ is that ordering. It depends exclusively on the per-group claim counts.

The argument is an exchange over the group ordering and is given in Section A. Two groupings matter here, and they differ in what a scorer is allowed to know. Grouping by concept alone gives the ceiling for a score that sees only *which* finding is asserted. Grouping by (concept, stated polarity) gives the ceiling for a score that also sees *how* it was asserted. Both are image-free—concept and polarity are read off the generated text, not the radiograph—and the second is the right comparator for any predictor built from the report alone, including the base-rate predictor we run, which is constant within (concept, polarity) rather than within concept. We report both and compare against the matched one. The practical content is unchanged: either bound needs only claim counts, so it can be computed for any published setup whose claim distribution is known, without access to the model.

Table 1 gives the result, and we must be careful to separate the theoretical limit from what is actually deployable. Conflating the two would overstate the case.

The ceiling bounds the metric. On every MIMIC generator it exceeds the best audit signal by 0.11 to 0.22 AUROC. It is an oracle: attaining it requires the per-group error counts, so this says what the metric *can* reward with no image, not what an available baseline scores. The realisable version is weaker. Ranking claims by training base rates alone gives 0.631 to 0.676, and the best signal beats that on all five generators—so we do not claim that visual evidence loses to a base-rate predictor. Part of that margin can come from calibration: our four signals' confidences include a per-concept intercept fitted on validation labels (Appendix D), which is itself a prior. The evidence that they read the image is the stratified metric, which that intercept cannot move: the contrastive world model scores 0.560 to 0.644 within (concept, polarity) on every generator, against 0.500 for the base-rate predictor (Table 5). What we claim is narrower and still awkward for the metric: this crude image-free score, using no test-set information at all, already outranks the supervised probe on three of five generators and the contrastive world model on all five, while sitting 0.11 to 0.22 below a ceiling that any sufficiently prior-fitted score approaches.

Table 2: The decomposition on the standard CheXbert-14 evaluation: 14 conditions × 2,018 MIMIC-CXR test studies, as the clinical-efficacy metric scores them. *Base-rate* ranks conditions by their positive rate in the *training* reports, using no test information and no image. Because it depends on the stated polarity it is constant within (condition, polarity), so it is bounded by the right-hand ceiling and can exceed the condition-only one, as it does on three generators. Nothing here uses an image, a model, or an audit score.

				ceiling (Proposition 2)	
Generator	claims	between	base-rate	condition	(cond., pol.)
DiA-CLEAR	28,252	93.6%	0.720	0.716	0.787
DiA	28,252	93.6%	0.729	0.713	0.792
ViT–BERT	28,252	93.4%	0.701	0.703	0.777
CheXagent	28,252	93.4%	0.715	0.707	0.786
MedGemma	28,252	93.4%	0.687	0.694	0.771

The pooled metric also reorders methods. Differential similarity is far the stronger concept-level discriminator ($+0.149$ over raw cosine) yet scores *worse* pooled, and on DiA-CLEAR the pooled metric ranks the reconstruction-only world model first while the stratified metric of Section 4.4 ranks it last (Table 5). The prior term does not merely inflate a real signal; it changes which method a reader would choose.

Is this just arithmetic in the number of concepts? One might wonder if this is just an artifact of having many concepts. It is, in part. With balanced counts the within share is exactly $1/C$, so a large bank guarantees a large w_{bet}; our counts are concentrated, so the measured w_{in} is 2.6–4.9× that prediction, but w_{bet} near 90% is still largely structural. The ceiling is not. Holding each concept's claim count fixed but giving every concept the same odds of correctness leaves C unchanged and w_{bet} at 77–88%, yet the ceiling collapses from 0.845–0.938 to 0.504–0.506. What makes it high is how unevenly correctness is spread across concepts—a property of the generator, not the bank size—which is why it must be computed per setup.

Validation on the Standard CheXbert-14 Benchmark. To confirm this isn't a quirk of our specific concept bank, we replicate the calculation on the standard CheXbert-14 clinical-efficacy evaluation (Smit et al., 2020). As shown in Table 2, the between-condition share is 93.4% to 93.6% and the text-only ceiling is 0.771 to 0.792. The realisable predictor, ranking conditions by their positive rate in the *training* reports, already reaches 0.687 to 0.729 without any test information. A reported pooled AUROC on this benchmark is therefore better read against those numbers than against 0.5.

A published auditor. Every signal above is one we implement, which leaves open whether the pattern is ours or the metric's. RadFlag (Zhang et al., 2024) settles it for one published method: it is black-box, trains nothing, and scores a claim by how often the generator repeats it across $K=5$ samples at temperature 1. On all five generators, RadFlag's pooled AUROC is *below* the base-rate predictor that never sees the image (0.589–0.663 against 0.631–0.676; Table 5). It is not a weak auditor. Under the stratified metric of Section 4.4 it scores 0.551–0.598, above the base-rate predictor's 0.500; and at the task it beats the base-rate predictor at 20% review on three of the five, including CheXagent, where none of our own signals does (Table 6).

One caveat bounds this comparison: RadFlag only votes on concepts some sample mentions, which covers 7 to 12% of (study, concept) cells; the rest carry no evidence and default to 0.5, so its pooled score is computed over a largely flat signal. However, this is not merely an artifact of heavy ties pulling the AUROC toward 0.5. Even when restricted strictly to the subset of claims RadFlag actually votes on, the base-rate predictor evaluated on those exact same claims still outperforms it (e.g., 0.681 against 0.645 on DiA-CLEAR; exact numbers vary by generator but the ordering holds). The pooled metric ranks a useful published auditor below a predictor that ignores the image.

4.2 Evaluating the Clinical Utility Proxy: Does Visual Evidence Aid Claim Triage?

Even if a metric is theoretically compromised, the underlying signals might still be useful in practice. We test this by evaluating claim triage: ranking each generator's claims by audit confidence and

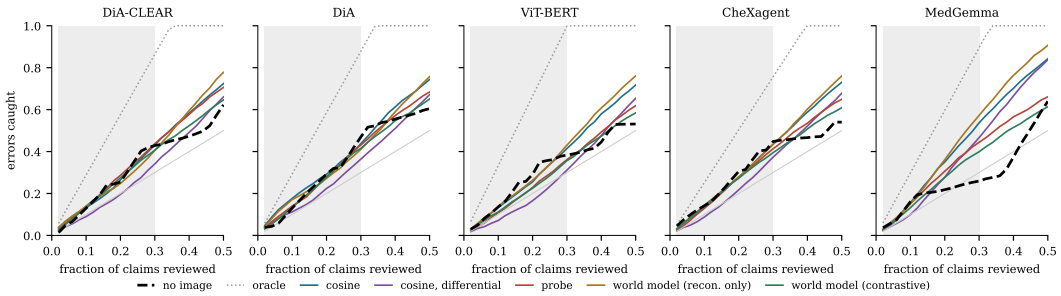

Figure 2: Errors caught against fraction of claims reviewed, for every method and generator. The grey band is the operational window $k \leq 0.30$; the thin diagonal is random review. On the first four generators no method separates from the base-rate predictor (black dashed) inside the band. MedGemma is the exception discussed in Section 4.2: there every method sits above the baseline throughout, because its baseline is the weakest of the five rather than as its methods are the strongest.

measuring the fraction of actual errors caught in the least-confident $k\%$. We compare the audit signals against the *base-rate predictor*, which ranks claims by the concept's training base rate and never sees the image. (Tied confidences, which occur within each (concept, polarity) group for the base-rate predictor and on RadFlag's unvoted claims, are broken by the fixed order in which claims were extracted, which does not use correctness).

Visual Evidence Fails to Outperform Base Rates on Four Generators. We summarise by the mean gap from the base-rate predictor over the operational window $k \leq 0.30$, rather than at a single review fraction, because the comparison changes sign with k and choosing one k would prove nothing. Of the 25 signal–generator combinations, 17 sit at or below the base-rate predictor and 12 are below it with a cluster-bootstrap interval excluding zero (Table 7). All 17 come from DiA-CLEAR, DiA, ViT–BERT and CheXagent. The other 8 are the five MedGemma signals, treated below, and three small gaps on DiA-CLEAR and DiA ($+0.003$ to $+0.019$), of which only cosine on DiA exceeds its own interval ($+0.019$, $[+0.011, +0.029]$). The differential similarity—the strongest similarity-based source in Table 4—is the worst triage signal on those four, trailing the predictor by 0.042 to 0.086. Full curves are in Figure 2, tabulated in Table 6.

The MedGemma Exception. MedGemma presents a notable exception: all five visual methods beat the base-rate baseline by $+0.038$ to $+0.104$. However, this is primarily because MedGemma's baseline is unusually weak (0.218 errors caught at 20% review, compared to 0.255–0.305 for other models). While we hypothesized this was due to MedGemma's concentrated vocabulary, restricting other generators to match this concentration did not replicate the effect. A plausible, though untested, alternative hypothesis is generation distribution shift: as a generalist model, MedGemma's error profile diverges from MIMIC-CXR training prevalences, rendering a MIMIC-calibrated base-rate predictor an artificially weak baseline. What it establishes is that visual evidence *can* aid triage; what the other four generators establish is that it does not do so reliably; and the pooled metric distinguishes neither case.

MedGemma gives the sharpest single illustration of the gap between the metric and the task. RadFlag has the *lowest* pooled AUROC of every signal on it (0.589, against 0.809 and 0.833 for cosine and the reconstruction-only world model), yet catches the *most* errors at 10% review (0.206, against at most 0.161 for every other signal and for the base-rate predictor; Table 6). A reader choosing an auditor by pooled AUROC would pick the worst one for the operating point that matters most.

4.3 A Reported Accuracy Does Not Constrain Where a Signal Reads

While AUROC isn't designed to measure attribution, one might hope that high accuracy forces a model to look at the right regions. It doesn't. We define *regionality* as the macro AUROC a signal loses when its read moves from the finding's own region to a size-matched disjoint control, and compare two world-model configurations differing only in the polarity weight λ (Table 10).

The two configurations reach 0.801 ± 0.019 and 0.806 ± 0.016—a gap several times smaller than the seed spread—while their regionality differs about five-fold. Because both rows run the same three seeds, we can pair them: the regionality difference is $+0.113$ and is positive on every seed, and the

ratio computed within each seed is 4.9, 5.0 and 4.5, so the effect is not an artefact of dividing two noisy means. We quote it as "about five-fold", since three seeds support no more precision.

The same holds for the signal the literature uses. Both configurations above are instruments we built, which is a fair objection to generalising from them. So we measure regionality for image–text similarity, the mechanism concept-bottleneck auditors actually use, by pooling the image encoder's spatial map over only the finding's region rather than the whole image. The towers are DiA-CLEAR's own (Section 3.2), so this tests the mechanism rather than a published checkpoint. Pooling over all cells reproduces the unoccluded scores exactly, which is the check that the regional read is the same computation. The result is stark: differential cosine scores 0.691 from the finding's own region and 0.687 from a disjoint control region, and its regionality interval includes zero on every split ($+0.004$ $[-0.011, +0.021]$ on MIMIC test). Whatever the similarity signal reads, it is not the finding. This could partly reflect feature delocalization in the encoder, but the same encoder and mask do register a drop for raw cosine ($+0.019$ $[+0.005, +0.040]$; Table 13), so the regional read is not blind by construction.

Of the five signals we measure (the probe, the world model at both λ, and the two similarity signals), neither ordering determines the other (Table 9 in Section C.6): the probe is most accurate, the $\lambda{=}1$ world model most regional, and differential cosine last on regionality (Table 13). The gap widens under institutional shift, as a co-occurrence account predicts, though one shift pair is weak evidence.

4.4 A Stratified Metric That Removes the Prior Term

The natural fix follows directly from Equation (6): isolate the within-group term. Stratifying by concept alone is not enough (macro-averaged concept-only stratification leaves every signal at 0.45–0.61, which is distinct from the pair-weighted within-concept term in Table 5). Adding polarity recovers 0.48–0.70.

Stratified by concept and polarity, the base-rate predictor scores exactly 0.500 (Table 8), so nothing a text-only score can exploit survives. That property validates the metric; the winner does not. The stratified metric selects the supervised probe on all five generators—at stratum floors of 5, 10, 20 and 30 alike (Section C.5)—but the probe is trained on the labels we evaluate against, so its selection is expected rather than informative.

More tellingly, if we exclude the oracle probe, the stratified metric selects the contrastive world model or RadFlag, and ranks the reconstruction-only variant last on four of five generators, and below every non-similarity signal on all five—exactly inverting the pooled ordering, which ranked it first on four. We recommend reporting within-(concept, polarity) AUROC alongside the ceiling for the setup.

5 Conclusion & Limitations

We have shown that the standard pooled AUROC for claim-level auditing is overwhelmingly driven by dataset priors rather than visual grounding. It spends 90–94% of its weight on comparisons that need no image, and a text-only (concept, polarity) ceiling exceeds every audit signal we test, a published one included, by 0.11–0.22. Its realisable counterpart is beaten, and the stratified metric shows the signals carry image information, so visual evidence is not worthless; the metric simply reports it at several times its weight and cannot say where it reads. We provide the ceiling and a stratified metric that removes the prior term.

Limitations. Regionality shows dependence on a region, not reading of it; under content substitution our evidence tracks a donor's label less closely than the probe. Triage has an unexplained counterexample in MedGemma. Of the published auditors we discuss, only RadFlag can audit a report it did not produce. Our protocol depresses the large VLMs, though weaker generators show no sign of inflating the ceiling (Section B). Crucially, the decomposition and ceiling depend only on claim counts, so they hold for whatever reports a generator produces—the protocol affects generator quality, not the mathematical validity of the metric analysis. Concept coverage (which shrinks from 66 to 14 concepts on MIMIC and 4 on IU under the stratified metric's stability floor) biases evaluation toward frequent conditions. Magnitudes are specific to chest radiography. Future work should focus on designing auditors that are verifiably region-aware and integrating the proposed stratified metric into broader clinical evaluation frameworks to ensure safety in automated reporting.

AI USE STATEMENT

Generative AI was used for manuscript preparation only: LaTeX conversion and formatting, bibliographic preparation, prose editing, and identification of internal inconsistencies across tables. It did not design experiments, generate or analyse data, or produce any empirical result reported here. All numbers are reported from successful experiments, and all claims were verified by the authors against those artifacts.

ETHICS STATEMENT

This work is a retrospective analysis of two public, de-identified chest radiograph corpora. MIMIC-CXR is accessed under the PhysioNet Credentialed Health Data License with the required data use agreement and CITI training completed; IU X-Ray is used under its public release terms. No patients were recruited and no new data were collected, so institutional review board approval was not required for secondary analysis of de-identified data. The audit scores we compute are research outputs and are not validated clinical decision rules; none of the generators or signals studied here is proposed for clinical deployment. Our central finding concerns a safety-relevant risk: an evaluation protocol that credits an audit signal for reproducing dataset priors may lead to tools that fail to flag the errors clinicians most need surfaced. Concept coverage, anatomical region defaults, and reference-label extraction all bound what the audit checks, and these limits are stated in § 3 and Appendix D.

REPRODUCIBILITY STATEMENT

Proofs of Propositions 1 and 2 are in Appendices A and B, with the shared-slope restriction tested empirically in Appendix B. Appendix D gives the dataset splits and study counts, concept bank construction including both count floors, world-model configurations and training cost, and the exact prompts and output-parsing rules for all five generators. Appendix C reports every signal under every metric, the triage tabulation, stratum-floor robustness, and the interventions that did not change the result. Every table in the paper is generated from stored result artifacts by a single script; no number is transcribed by hand. Code, configurations, the concept bank, and the ceiling computation will be released.

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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

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

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *ICCV*, 2023.

David Bani-Harouni, Chantal Pellegrini, Julian Lüers, Su Hwan Kim, Markus Baalmann, Benedikt Wiestler, Rickmer Braren, Nassir Navab, and Matthias Keicher. Calibrated confidence expression for radiology report generation. *arXiv preprint arXiv:2603.29492*, 2026.

Shruthi Bannur, Kenza Bouzid, Daniel C. Castro, Anton Schwaighofer, Anja Thieme, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, Felix Meissen, Mercy Ranjit, Shaury Srivastav, Julia Gong, Noel C. F. Codella, Fabian Falck, Ozan Oktay, Matthew P. Lungren, Maria Teodora Wetscherek, Javier Alvarez-Valle, and Stephanie L. Hyland. MAIRA-2: Grounded radiology report generation. *arXiv preprint arXiv:2406.04449*, 2024.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. In *EMNLP*, 2020.

Zhihong Chen, Maya Varma, Justin Xu, Magdalini Paschali, Dave Van Veen, Christian Bluethgen, Eduardo Pontes Reis, Sergios Gatidis, Curtis P. Langlotz, and Akshay S. Chaudhari. A vision-language foundation model to enhance efficiency of chest X-ray interpretation. *arXiv preprint arXiv:2401.12208*, 2024.

Jean-Benoit Delbrouck. chexpert-mimic-cxr-findings-baseline: a ViT–BERT encoder–decoder for findings generation. `https://huggingface.co/IAMJB/chexpert-mimic-cxr-findings-baseline`, 2023. Model card; accessed September 2026.

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.

Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

Quentin Garrido, Mahmoud Assran, Nicolas Ballas, Adrien Bardes, Laurent Najman, and Yann LeCun. Learning and leveraging world models in visual representation learning. *arXiv preprint arXiv:2403.00504*, 2024.

Google DeepMind. MedGemma-4b-it. `https://huggingface.co/google/medgemma-4b-it`, 2025. Model card; accessed September 2026.

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)*, 2017.

Tianyu Han, Riga Wu, Yu Tian, Firas Khader, Lisa C. Adams, Keno K. Bressem, Christos Davatzikos, Jakob Nikolas Kather, Li Shen, David A. Mankoff, Eduardo Mortani Barbosa, and Daniel Truhn. CLEAR: an auditable foundation model for radiology grounded in clinical concepts. *Nature Biomedical Engineering*, 2026. doi: 10.1038/s41551-026-01741-4.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P. Lungren, Andrew Y. Ng, Curtis P. Langlotz, and Pranav Rajpurkar. RadGraph: Extracting clinical entities and relations from radiology reports. In *NeurIPS Datasets and Benchmarks*, 2021.

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*, 2019.

Nathan Kallus and Angela Zhou. The fairness of risk scores beyond classification: Bipartite ranking and the xAUC metric. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. arXiv:1902.05826.

Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson, Michael Moseley, Curtis Langlotz, Akshay S. Chaudhari, and Jean-Benoit Delbrouck. GREEN: Generative radiology report evaluation and error notation. In *Findings of EMNLP*, 2024.

John Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in large margin classifiers*, 1999.

Binesh Sadanandan and Vahid Behzadan. Attention without grounding: Causal evaluation of visual explanations in medical VLMs. *arXiv preprint arXiv:2607.18577*, 2026.

Nagur Shareef Shaik, Teja Krishna Cherukuri, Adnan Masood, and Dong Hye Ye. DiA-gnostic VLVAE: Disentangled alignment-constrained vision language variational autoencoder for robust radiology reporting with missing modalities. In *Proceedings of the 40th AAAI Conference on Artificial Intelligence (AAAI-26)*, 2026. Oral presentation, main technical track.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y. Ng, and Matthew P. Lungren. CheXbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In *EMNLP*, 2020.

Stanford AIMI. CheXagent-2-3b. `https://huggingface.co/StanfordAIMI/CheXagent-2-3b`, 2025. Model card; accessed September 2026.

U.S. Food and Drug Administration. Artificial intelligence-enabled medical devices. `https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-enabled-medical-devices`, 2026.

Ziao Wang, Sixing Yan, Kejing Yin, Xiaofeng Zhang, and William K. Cheung. CURV: Coherent uncertainty-aware reasoning in vision-language models for x-ray report generation. In *NeurIPS*, 2025.

Yang Yue, Yulin Wang, Chenxin Tao, Pan Liu, Shiji Song, and Gao Huang. CheXWorld: Exploring image world modeling for radiograph representation learning. In *CVPR*, 2025.

Anas Zafar, Leema Krishna Murali, Siddhant Bharadwaj, Ashish Vashist, and Jia Wu. Do medical vision language models actually see? a counterfactual grounding framework and hard-negative contrastive training for visually-reliant medical VLMs. *arXiv preprint arXiv:2607.03647*, 2026.

Serena Zhang, Sraavya Sambara, Oishi Banerjee, Julian Acosta, L. John Fahrner, and Pranav Rajpurkar. RadFlag: A black-box hallucination detection method for medical vision language models. *arXiv preprint arXiv:2411.00299*, 2024.

A THE DECOMPOSITION AND THE PRIOR CEILING

A.1 THE IDENTITY

Let $P = \sum_c n_c^+$ and $Q = \sum_c n_c^-$ be the total correct and incorrect claims. AUROC is the fraction of the PQ ordered pairs (one correct, one incorrect) in which the correct claim receives the higher confidence, ties counting a half. Partitioning those pairs by whether the two claims share a concept gives disjoint, exhaustive sets of sizes $\sum_c n_c^+ n_c^-$ and $PQ - \sum_c n_c^+ n_c^-$, and Equation (6) follows immediately by grouping terms. The identity holds for any scoring function; nothing about the audit is assumed.

Empirically the reconstruction $w_{\text{in}} \cdot \text{AUROC}_{\text{within}} + w_{\text{bet}} \cdot \text{AUROC}_{\text{between}}$ matches the directly computed pooled AUROC to within 10^{-5} for every method and generator, which is how we verify the implementation.

A.2 PROOF OF PROPOSITION 2

Let s be any confidence score that is constant within each concept, s_g. Within-concept pairs then tie and contribute $\frac{1}{2}$ each. A between-concept pair, with the correct claim in g and the incorrect in g', contributes 1 if $s_g > s_{g'}$ and 0 if $s_g < s_{g'}$. So the objective depends on s only through the total order it induces on concepts, and equals

$$\frac{1}{PQ}\Big(\tfrac{1}{2}\sum_g n_g^+ n_g^- + \sum_{g \prec g'} n_g^+ n_{g'}^-\Big), \tag{7}$$

where $g \prec g'$ means g is ranked above g'. Consider two concepts adjacent in the order. Placing g above g' contributes $n_g^+ n_{g'}^-$; swapping them contributes $n_{g'}^+ n_g^-$. The first is larger exactly when $n_g^+/n_g^- > n_{g'}^+/n_{g'}^-$, and the comparison does not involve any other concept. An order is therefore optimal if and only if no adjacent swap improves it, which holds precisely when concepts are sorted by descending odds of correctness. Groups with zero incorrect claims ($n_g^- = 0$) represent infinite odds and naturally sort to the top of the ordering. Since the objective is a sum over pairs of a bounded quantity and the number of orders is finite, that order attains the maximum. □

Two things follow. The ceiling needs only the counts $\{(n_g^+, n_g^-)\}$, so it can be computed for any reported setup whose claim distribution is known. And it is an *oracle* over orderings: it assumes knowledge of which concepts are answered correctly on the evaluation set, which a deployed system would have to estimate. The fully realisable version, ordering by training base rates, is the *prior only* row of Table 8 and reaches 0.631 to 0.676—still above the pooled score of several methods.

B PROOF OF PROPOSITION 1

Under the stated model the hidden region's features are Gaussian about the predictor's output with isotropic variance τ^2, so for either polarity π,

$$\log p(h \mid z, (c, \pi)) \;=\; -\frac{1}{2\tau^2}\big\|g_\theta(z, a(c,\pi), R_c) - h\big\|^2 \;-\; \frac{N_t d}{2}\log\big(2\pi\tau^2\big)\,,$$

where the normalising term does not depend on π. Taking the difference between the two polarities, the normalisers cancel:

$$\log \text{BF}_c \;=\; \log\frac{p(h \mid z, (c,+))}{p(h \mid z, (c,-))} \;=\; \frac{1}{2\tau^2}\Big(\big\|g_\theta(z, a(c,-), R_c) - h\big\|^2 - \big\|g_\theta(z, a(c,+), R_c) - h\big\|^2\Big).$$

The bracket is exactly $N_t d\,[\,E(c,-) - E(c,+)\,]$ by the definition of E in Equation (1), which is $N_t d\, e_c$ in expectation over the draws in Equation (2). Substituting gives $\log \text{BF}_c = N_t d\, e_c/(2\tau^2)$, which is Equation (3).

For Equation (4), Bayes' rule on the two hypotheses gives

$$\Pr[(c,+) \mid z, h] \;=\; \sigma\!\left(\log \text{BF}_c + \log\frac{\Pr[(c,+) \mid z]}{\Pr[(c,-) \mid z]}\right) \;=\; \sigma\big(\tfrac{N_t d}{2\tau^2}\, e_c + \beta_c\big)\,,$$

with β_c the log prior odds. (This replaces the context-conditional prior with a marginal one, meaning the family is exact up to a context-independence assumption.) So the calibration family is a sigmoid of an affine function of e_c with a shared slope $\alpha = N_t d/(2\tau^2)$ (which presumes a uniform region size N_t across concepts) and a per-concept intercept, which is the family we fit. (Substituting the expected value e_c yields a sigmoid of the expected evidence, rather than the exact joint posterior).

What the statement does and does not claim. The Bayes factor is the one contributed by the hidden region *given the context* z; it is not the posterior odds of the finding in the study as a whole, and evidence visible only outside R_c does not enter it. Isotropy is an assumption about the residual, not about the features, and a fitted diagonal τ^2 per channel would give the same form with a reweighted norm. The empirical content of the proposition is the functional form of the calibration family; that the fitted slope is shared across concepts while the intercept is not is a testable restriction, which we test next.

Testing the shared slope. We fit both families on MIMIC validation—the shared slope of Equation (4), and a variant with a separate slope per concept—giving the per-concept slopes the same Gaussian shrinkage the production fit gives the intercepts, and compare them on held-out test data. Comparing in-sample likelihood would favour the larger model by construction.

	shared	per-concept slope	test log-loss	
Evidence	slope	(IQR, 11 concepts)	shared	per-concept
World model, λ=1	3.560	[3.509, 3.651]	0.2980	0.2965
World model, λ=10	4.283	[4.258, 4.327]	0.2647	0.2642
Probe	3.178	[3.041, 3.404]	0.2882	0.2843

For the world model the restriction holds closely but not exactly. Per-concept slopes fall within a few percent of the shared value, and freeing them improves held-out log-loss by 0.0015 at λ=1 and 0.0005 at λ=10, with no change in calibration error. The improvement is consistently positive, so "exact" describes the derivation, not these scores. The probe is a useful contrast: it has no Gaussian likelihood structure and Proposition 1 does not apply to it, and its slopes are two to five times more dispersed and gain most from being freed (0.0039). The model describes the evidence it was derived for better than one it was not. Only 11 concepts have enough validation data to estimate a slope of their own, so this is a check rather than a strong test.

Do weaker generators inflate the ceiling? A weak generator makes many errors, and one might expect that to hand the prior more to exploit. Across our five generators the text-only ceiling barely moves, 0.926 to 0.947, while claim accuracy spans 0.603 to 0.699 and RadGraph F_1 spans 0.195 to 0.228. Neither quality measure tracks the ceiling (Spearman $\rho = -0.05$ against claim accuracy, -0.36 against F_1). Five generators, all run under the same protocol, cannot rule the effect out; they give no sign of it. That protocol—one frontal image, greedy decoding, no indication—does depress the large VLMs below the scores they report under their own prompts, and we did not re-run them under those prompts, since doing so would change every number in the paper.

C Additional results

C.1 The audited generators

To ensure our findings are robust across different model architectures and scales, we evaluate five distinct generative systems. Table 3 details these models. We include two variational vision-language autoencoders we trained ourselves from published recipes, DiA (Shaik et al., 2026) and DiA-CLEAR, the latter substituting the concept-grounding objective of Han et al. (2026) for DiA's alignment term. To capture the spectrum of modern transformer-based architectures, we include ViT–BERT (a strong 61M parameter baseline) (Delbrouck, 2023), CheXagent-2-3b (a specialized 3B parameter VLM) (Chen et al., 2024), and MedGemma-4b-it (a 4.3B parameter generalist medical VLM) (Google DeepMind, 2025). Notably, the lightweight ViT–BERT baseline achieves the highest RadGraph F_1 score, underscoring that parameter count does not strictly correlate with clinical accuracy in this domain.

Table 3: The audited generators. All produce a findings section for the same 2,018 MIMIC-CXR test studies under the same protocol. DiA is the disentangled alignment-constrained VLVAE of Shaik et al. (2026); DiA-CLEAR is that architecture with the concept-grounding objective of CLEAR (Han et al., 2026) substituted for its alignment term, and is named for that pairing; ViT–BERT is Delbrouck (2023); CheXagent-2-3b is (Chen et al., 2024, model card: Stanford AIMI, 2025); MedGemma-4b-it is Google DeepMind (2025). The 61M ViT–BERT scores highest on the target metric, above a 3B chest-X-ray vision–language model.

Generator	Type	Params	% unique	RadGraph F_1
DiA-CLEAR	concept-grounded VLVAE	small	56.9%	0.210
DiA	disentangled-alignment VLVAE	small	51.4%	0.217
ViT–BERT	MIMIC-specific encoder–decoder	61M	80.7%	0.228
CheXagent-2-3b	CXR specialist VLM	3B	91.9%	0.220
MedGemma-4b-it	generalist medical VLM	4.3B	76.7%	0.195

C.2 Concept-level evidence quality

Before aggregating claims into a pooled metric, it is crucial to understand the raw discriminative power of each evidence source at the individual concept level. Table 4 provides the macro AUROC averaged over concepts that clear a strict count floor. We observe that the standard raw cosine similarity performs poorly (near chance), largely because it discards polarity information. The differential form recovers $+0.149$ macro AUROC on the MIMIC test set, making it the stronger concept-level discriminator by a wide margin. That advantage does not transfer: as shown in Section 4.1, it scores *worse* on the pooled claim-level metric than the raw form it improves on, which is direct evidence that the pooled metric is not measuring concept-level discrimination. The attentive probe serves as our supervised upper bound.

Table 4: Concept-level evidence quality: macro AUROC averaged over the concepts that clear a per-split count floor (6 on validation, 14 on test). The contrastive world-model row is seed 0 at $\lambda = 1$; across three seeds MIMIC test is 0.801 ± 0.019.

Evidence source	MIMIC val	MIMIC test	IU zero-shot	ECE (test)
Cosine similarity	0.543	0.536	0.610	0.126
Cosine, differential	0.843	0.685	0.728	0.061
World model, reconstruction only	0.681	0.646	0.647	0.112
World model, contrastive	0.864	0.779	0.815	0.018
Probe (discriminative bound)	0.908	0.848	0.887	0.018

C.3 The full method × metric grid

While Table 1 and Table 8 in the main text report only the top-selected methods for brevity, Table 5 provides the exhaustive evaluation of every evidence method under every metric. This full grid reveals a critical pathology of the pooled metric: it consistently prefers the reconstruction-only world model, a variant trained without polarity contrast that performs near chance at distinguishing a finding from its negation (0.544 polarity discrimination vs. 0.925 for the contrastive variant). Conversely, the contrastive variant—which is objectively the better visual evidence source—is ranked *last* by the pooled metric across all five generators. This inversion explicitly demonstrates that the pooled metric rewards prior memorization over actual visual grounding.

C.4 Claim triage, tabulated

To quantify the clinical utility proxy discussed in Section 4.2, Table 6 provides the exact numerical values for the claim triage curves presented in Figure 2. We report the fraction of actual errors caught at 10%, 20%, and 30% review thresholds. Across the first four generators no audit signal separates meaningfully from the base-rate ranking; MedGemma is the exception discussed in Section 4.2. The oracle row represents the theoretical maximum utility (perfectly ranking all errors first).

Table 5: Every signal under every metric, MIMIC-CXR test. *Pooled* is the standard metric; *within* is the within-concept term of Equation (6), pair-weighted so the identity closes; *strat.* is macro within-(concept, polarity) AUROC over strata with at least 10 claims of each outcome (Section 4.4). *Resid.* is the decomposition's reconstruction error. Bold is the best signal in each column. RadFlag is the published auditor; it votes only on concepts some sample mentions (10.1%, 10.9%, 11.6%, 10.6%, 7.3% of cells across generators) and defaults to 0.5 elsewhere.

Generator	Signal	pooled	within	strat.	resid.
DiA-CLEAR	base-rate predictor	0.647	0.500	0.500	—
	cosine	0.724	0.647	0.604	$< 10^{-5}$
	cosine, differential	0.653	0.593	0.596	$< 10^{-5}$
	probe	0.687	**0.685**	**0.671**	$< 10^{-5}$
	world model, reconstruction-only	**0.735**	0.651	0.547	$< 10^{-5}$
	world model, contrastive	0.640	0.677	0.636	$< 10^{-5}$
	RadFlag (published)	0.627	0.591	0.598	$< 10^{-5}$
DiA	base-rate predictor	0.676	0.500	0.500	—
	cosine	**0.744**	0.701	0.595	$< 10^{-5}$
	cosine, differential	0.664	0.641	0.601	$< 10^{-5}$
	probe	0.674	**0.716**	**0.702**	$< 10^{-5}$
	world model, reconstruction-only	0.739	0.709	0.564	$< 10^{-5}$
	world model, contrastive	0.639	0.697	0.644	$< 10^{-5}$
	RadFlag (published)	0.629	0.580	0.579	$< 10^{-5}$
ViT–BERT	base-rate predictor	0.631	0.500	0.500	—
	cosine	0.702	0.672	0.530	$< 10^{-5}$
	cosine, differential	0.625	0.571	0.571	$< 10^{-5}$
	probe	0.614	0.595	**0.622**	$< 10^{-5}$
	world model, reconstruction-only	**0.721**	**0.673**	0.517	$< 10^{-5}$
	world model, contrastive	0.573	0.611	0.581	$< 10^{-5}$
	RadFlag (published)	0.602	0.579	0.587	$< 10^{-5}$
CheXagent	base-rate predictor	0.674	0.500	0.500	—
	cosine	0.747	**0.729**	0.551	$< 10^{-5}$
	cosine, differential	0.681	0.599	0.523	$< 10^{-5}$
	probe	0.665	0.629	**0.574**	$< 10^{-5}$
	world model, reconstruction-only	**0.761**	0.705	0.484	$< 10^{-5}$
	world model, contrastive	0.617	0.637	0.567	$< 10^{-5}$
	RadFlag (published)	0.663	0.596	0.573	$< 10^{-5}$
MedGemma	base-rate predictor	0.659	0.500	0.500	—
	cosine	0.809	0.664	0.507	$< 10^{-5}$
	cosine, differential	0.767	0.662	0.571	$< 10^{-5}$
	probe	0.675	**0.695**	**0.659**	$< 10^{-5}$
	world model, reconstruction-only	**0.833**	0.650	0.542	$< 10^{-5}$
	world model, contrastive	0.598	0.652	0.560	$< 10^{-5}$
	RadFlag (published)	0.589	0.595	0.551	$< 10^{-5}$

C.5 Pooled versus stratified selection

Table 8 explicitly compares the method selected by the flawed pooled metric versus our proposed stratified metric. By stratifying across both concept and polarity, we neutralize the influence of dataset priors—evidenced by the fact that the image-free predictor drops to exactly 0.500 (chance) under stratification. With the prior removed, the stratified metric selects the attentive probe on all five generators where the pooled metric selects the reconstruction-only world model or cosine. As Section 4.4 notes, the probe is trained on the labels we evaluate against, so its selection is expected rather than informative; what validates the metric is the base-rate predictor falling to 0.500. Table 11 shows the selection does not depend on the stratum floor.

Table 6: Claim triage: fraction of each generator's erroneous claims found in its least-confident $k\%$, for every signal. *No image* ranks by concept base rate alone; *oracle* puts all errors first. **Bold** marks a value exceeding the *no image* row of the same block and column. The window summary in Section 4.2 is computed over our five signals; RadFlag is reported separately there.

Generator	ranking by	@10%	@20%	@30%
DiA-CLEAR	no image (base rates)	0.129	0.255	0.429
	cosine	**0.132**	**0.265**	**0.431**
	cosine, differential	0.090	0.196	0.341
	probe	**0.139**	**0.286**	**0.437**
	world model, reconstruction-only	**0.140**	0.248	0.405
	world model, contrastive	**0.133**	**0.275**	0.406
	RadFlag (published)	0.128	**0.274**	0.417
	oracle	*0.288*	*0.576*	*0.864*
DiA	no image (base rates)	0.130	0.292	0.470
	cosine	**0.176**	0.292	0.438
	cosine, differential	0.112	0.224	0.364
	probe	**0.140**	0.279	0.431
	world model, reconstruction-only	**0.171**	0.272	0.411
	world model, contrastive	**0.133**	0.278	0.415
	RadFlag (published)	**0.136**	0.290	0.433
	oracle	*0.293*	*0.586*	*0.879*
ViT–BERT	no image (base rates)	0.136	0.291	0.383
	cosine	0.134	0.255	**0.415**
	cosine, differential	0.071	0.166	0.312
	probe	0.109	0.229	0.362
	world model, reconstruction-only	**0.137**	0.263	**0.431**
	world model, contrastive	0.112	0.225	0.356
	RadFlag (published)	**0.142**	0.282	**0.433**
	oracle	*0.333*	*0.665*	*0.998*
CheXagent	no image (base rates)	0.142	0.305	0.448
	cosine	0.130	0.278	0.434
	cosine, differential	0.088	0.217	0.373
	probe	0.131	0.274	0.413
	world model, reconstruction-only	**0.143**	0.279	0.435
	world model, contrastive	**0.143**	0.273	0.394
	RadFlag (published)	**0.155**	**0.330**	**0.492**
	oracle	*0.252*	*0.504*	*0.756*
MedGemma	no image (base rates)	0.161	0.218	0.260
	cosine	0.156	**0.350**	**0.537**
	cosine, differential	0.099	**0.277**	**0.466**
	probe	0.152	**0.306**	**0.444**
	world model, reconstruction-only	0.150	**0.361**	**0.572**
	world model, contrastive	0.132	**0.277**	**0.399**
	RadFlag (published)	**0.206**	**0.312**	**0.396**
	oracle	*0.298*	*0.596*	*0.895*

C.6 Accuracy versus regionality

Table 9 places all five signals on both axes at once. Neither ordering determines the other. The most accurate signal, the supervised probe at 0.848, is second on regionality at $+0.042$, less than a third of the λ=1 world model's $+0.143$; the rank correlation across the five is $+0.6$, so the two are related but not interchangeable, and a reported accuracy leaves the regional read largely unconstrained.

The pair of similarity signals is the sharpest case, because they differ in one term rather than in architecture. The differential form adds 0.149 macro AUROC over raw cosine and is the better concept-level discriminator by a wide margin, yet its regionality falls from $+0.019$ to $+0.004$ and ceases to be distinguishable from zero on any split (Table 13). Subtracting $\cos(v, \text{“no } c\text{”})$ supplies polarity information while removing what little dependence on the finding's own region the raw score had. Both terms are computed from a global pooling of the image, so there is no reason the

Table 7: Mean gap in errors caught between each signal and the base-rate predictor, averaged over the operational window $k \leq 0.30$ (Section 4.2). Negative means the signal does worse than ignoring the image. * marks a 95% cluster-bootstrap interval (300 resamples over studies) excluding zero. Of the 25 entries, 17 are at or below zero and 12 are significantly below it.

Generator	cos.	diff.	probe	WM-rec	WM-con.
DiA-CLEAR	-0.004	-0.061^*	$+0.005$	-0.012^*	-0.006
DiA	$+0.019^*$	-0.042^*	-0.002	$+0.003$	-0.006
ViT–BERT	-0.013^*	-0.086^*	-0.043^*	-0.008	-0.042^*
CheXagent	-0.015^*	-0.064^*	-0.021^*	-0.011^*	-0.020^*
MedGemma	$+0.094^*$	$+0.040^*$	$+0.062^*$	$+0.104^*$	$+0.038^*$

Table 8: The pooled metric and the stratified metric select different methods, on every generator. *Prior only* is the image-free predictor, scored both ways; under stratification it is at chance by construction, which is the property we want. WM-rec is the reconstruction-only world model. Stratified scores are macro-averaged over (concept, polarity) strata holding at least 10 claims of each outcome.

		prior only		selected method	
Generator	claims	pooled	stratified	by pooled	by stratified
DiA-CLEAR	4718	0.647	0.500	WM-rec (0.735)	probe (0.671)
DiA	5401	0.676	0.500	cosine (0.744)	probe (0.702)
ViT–BERT	5841	0.631	0.500	WM-rec (0.721)	probe (0.622)
CheXagent	5320	0.674	0.500	WM-rec (0.761)	probe (0.574)
MedGemma	4472	0.659	0.500	WM-rec (0.833)	probe (0.659)

improvement should be localised, and it is not. For the similarity signals, the region masks are resampled from CheXWorld's 14×14 grid to the image encoder's 8×8 map by area overlap, and the read pools already-encoded features rather than occluding the input, as for the probe.

We read this as the appendix-level statement of § 4.3: the quantity the pooled metric rewards and the quantity an audit needs are measured on different axes, and improving one carries no guarantee about the other. It is not a claim that low regionality is always undesirable—a signal may legitimately combine evidence across the image—only that accuracy does not report it.

Table 9: Accuracy and regionality rank the five signals differently. The probe is the most accurate but exhibits low regionality, while the $\lambda = 1$ world model is highly regional but less accurate. Differential cosine is the least regional of all. (MIMIC test AUROC and regionality drops are aggregated from Table 4, Table 10, and Table 13).

Signal	Accuracy (AUROC)	Regionality
Probe	0.848	$+0.042$
World model, $\lambda = 10$	0.806	$+0.030$
World model, $\lambda = 1$	0.801	$+0.143$
Cosine, differential	0.685	$+0.004$
Cosine similarity	0.536	$+0.019$

C.7 World-model ablations

To ensure that the dissociation between accuracy and regionality observed in Section 4.3 is robust, Table 12 presents single-seed ablations varying the inverse temperature (κ), initialization strategy, and region budget. Across these variations, the model maintains high predictive accuracy (AUROC ~ 0.80) while exhibiting consistently low regionality. This confirms that the spatial blindness of the accuracy metric is a persistent structural issue, not an artifact of a specific hyperparameter combination.

Table 10: Accuracy does not constrain regionality. The first two rows are the same model on the same data, differing only in the polarity loss weight λ. Their accuracies are indistinguishable while their regionalities differ by a factor of about five, with non-overlapping ranges across seeds. Both rows use the same three seeds, so the comparison is paired: the regionality difference is $+0.113$ and is positive on every seed ($+0.085, +0.145, +0.108$), and the per-seed ratio is $4.9, 5.0, 4.5$, mean 4.8.

Configuration	MIMIC test AUROC	Regionality
World model, $\lambda = 1$ (seeds 0–2)	0.801 ± 0.019	$+0.143 \pm 0.037$
World model, $\lambda = 10$ (seeds 0–2)	0.806 ± 0.016	$+0.030 \pm 0.007$
Probe	0.848	$+0.042$

Table 11: The signal the stratified metric selects, at four stratum floors (minimum claims of each outcome per (concept, polarity) stratum), among our five signals. The selection is the same at every floor on every generator.

Generator	floor 5	floor 10	floor 20	floor 30
DiA-CLEAR	probe (0.700)	probe (0.671)	probe (0.700)	probe (0.723)
DiA	probe (0.685)	probe (0.702)	probe (0.736)	probe (0.755)
ViT–BERT	probe (0.618)	probe (0.622)	probe (0.649)	probe (0.658)
CheXagent	probe (0.564)	probe (0.574)	probe (0.609)	probe (0.609)
MedGemma	probe (0.630)	probe (0.659)	probe (0.647)	probe (0.647)

C.8 Occlusion with confidence intervals

Table 13 gives the statistical backing for the regionality drops visualized in Figure 3. We compute 95% confidence intervals using percentile bootstrap resampling over studies (Efron & Tibshirani, 1994). The paired contrast column confirms that the difference in spatial reliance between the world model and the probe is statistically significant on the MIMIC test set. On IU X-Ray, read zero-shot, the gap widens, which is what a co-occurrence account of low-regionality signals predicts—though a single shift pair is weak evidence for it.

Regionality is positive for 13 of 14 scored concepts on MIMIC test, 5 of 6 on validation and 4 of 4 on IU, so the aggregate is not driven by a few concepts.

C.9 What did not change the result

We report the interventions that failed, since each was an attempt to make our own evidence look better.

Estimator variance. The evidence averages D deterministic spatial draws of the hidden region. Raising D from 4 to 16, 32 and 64 moves MIMIC test AUROC by 0.006 in total ($0.779, 0.785, 0.784, 0.784$) and is flat from 16 onward. Sampling noise is not the limit.

More data. Training the world model on all $140{,}571$ training studies rather than the $60{,}000$ subset gives 0.804 against 0.815—no improvement.

The contrast weight. A sweep of $\lambda \in \{0.3, 1, 3, 10\}$ appeared to favour $\lambda = 10$ (0.820 against 0.779). Crossing λ with seeds dissolved the effect: 0.801 ± 0.019 at λ=1 against 0.806 ± 0.016 at λ=10, over seeds 0–2. The apparent gain was one lucky seed against one unlucky one. We report the crossed design, and the pre-registered criterion—that the world model come within 0.03 macro AUROC of the probe on validation—is not met at the registered setting.

Content substitution. Replacing a finding's own patch features with those of a donor study carrying the opposite label makes the probe follow the donor (0.760) far more than our evidence does (0.558). This is the opposite of what region dependence would predict, and we report it. The two readouts are not comparable under this perturbation: the probe reads those tokens directly, whereas the world model *predicts* them from surrounding context, so substituting the target degrades its signal rather than redirecting it. A same-label control quantifies that damage at $+0.040$ for the world model against -0.074 for the probe.

Table 12: Single-seed ablations, each λ=10 at seed 0 with one setting changed from the row above it in Table 10. Configurations are in Table 14. Because they share λ=10 and one seed, we read these as showing the accuracy/regionality dissociation survives other changes, not as five independent measurements.

Configuration	MIMIC test AUROC	Regionality
$\lambda = 10$, $\kappa = 10$	0.785	+0.061
$\lambda = 10$, $\kappa = 30$	0.804	+0.020
$\lambda = 10$, $\kappa = 300$	0.797	+0.020
$\lambda = 10$, no CheXWorld warm start	0.815	+0.020
$\lambda = 10$, half region budget	0.815	+0.014

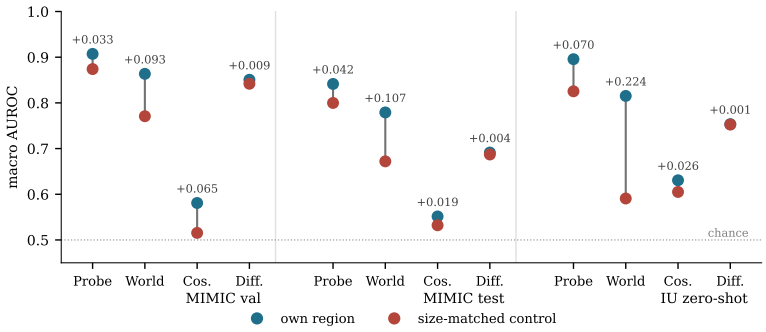

Figure 3: Regionality. Each pair is one evidence source scored from a finding's own anatomical region and from a size-matched, disjoint control region elsewhere in the image; the drop is the regionality. *World* is the contrastive world model at λ=1. The probe answers almost as well from the wrong region as the right one, on all three splits. The gap widens under institutional shift: on IU X-Ray, read zero-shot, the world model (seed 0) falls from 0.815 to 0.591, near chance, while the probe, reading only the finding's region, falls from 0.895 to 0.825.

C.10 Report-level confidence

A confidence head pooling claim evidence with the generator's own decoder state reaches 0.734 AUROC with real probe evidence, 0.733 with evidence *shuffled across reports*, and 0.726 with no evidence at all. Our own evidence is the worst of the four inputs at 0.713, below its own shuffled control at 0.722. We report this as a negative result: at report level this evidence adds nothing, most likely because report-level RadGraph F_1 is largely predictable from which findings were claimed at all.

D Reproducibility

Every table is generated from stored result artifacts by a single script; no number is transcribed by hand, and re-running it reproduces the tables byte for byte.

Splits. MIMIC-CXR under its official patient-level split, one frontal image per study with PA preferred and AP as fallback: 140,571 training, 1,099 validation and 2,018 test studies. Evidence models are fitted on a fixed 60,000-study subset of the training split. Evidence is extracted for the 63,117 MIMIC studies and 2,943 IU studies that carry both a frontal image and a parsable findings section. IU X-Ray uses the split of Chen et al. (2020), 2,060/294/589 studies, and is never fitted on.

Concept bank. Mined once from 60,000 MIMIC training reports, capped at 128 candidates (selected by highest mention frequency) and filtered by a two-sided floor: a concept is kept only if at least 50 studies assert it and at least 50 negate it. Sixty-six concepts survive. Concepts radiologists never negate are therefore excluded by construction. For the claim-level analyses a second floor applies at evaluation time: a (concept, polarity) stratum enters the stratified metric only with at least 10 correct and 10 incorrect claims, which leaves 14 concepts on MIMIC test and 4 on IU.

Table 13: Regionality with 95% paired cluster-bootstrap intervals over studies (1,000 resamples). * marks an interval excluding zero. *World model* is the contrastive model at λ=1. Three world-model regionality figures appear in the paper and measure different things: $+0.107$ is seed 0 alone (Figure 3), $+0.123$ here is the regionality of the evidence averaged over seeds 0–2, and $+0.143$ is the mean of the three per-seed regionalities (Table 10). Differential cosine is the only signal whose regionality is indistinguishable from zero on every split; raw cosine's excludes zero on both MIMIC splits but not on IU. The world-model minus probe contrast on MIMIC test is $+0.081$ $[+0.024, +0.127]$. Own-region scores restrict each signal to the finding's region, so the probe's differ from its whole-image scores in Table 4.

Signal	MIMIC val	MIMIC test	IU zero-shot
Probe	$+0.033$ $[+0.003, +0.052]^*$	$+0.042$ $[+0.018, +0.064]^*$	$+0.070$ $[+0.004, +0.126]^*$
World model	$+0.077$ $[+0.025, +0.144]^*$	$+0.123$ $[+0.072, +0.162]^*$	$+0.218$ $[+0.044, +0.330]^*$
Cosine	$+0.065$ $[+0.025, +0.080]^*$	$+0.019$ $[+0.005, +0.040]^*$	$+0.026$ $[-0.014, +0.091]$
Diff. cosine	$+0.009$ $[-0.010, +0.022]$	$+0.004$ $[-0.011, +0.021]$	$+0.001$ $[-0.031, +0.027]$

World-model configurations. All rows share the frozen CheXWorld encoder, the likelihood-ratio readout with $D = 4$ draws, and $\kappa = 100$ unless stated. Table 14 gives the settings behind every row of Table 10.

Table 14: Configurations behind Tables 10 and 12. λ is the weight on the polarity-contrastive term, κ its inverse temperature. *Warm start* initialises the predictor from CheXWorld; without it the same architecture trains from scratch. *Region budget* is the cap on hidden region tokens, 96 by default.

Configuration	λ	κ	warm start	region budget	seeds
World model, $\lambda = 1$	1	100	yes	96	$0, 1, 2$
World model, $\lambda = 10$	10	100	yes	96	$0, 1, 2$
$\lambda = 10$, $\kappa = 10$	10	10	yes	96	0
$\lambda = 10$, $\kappa = 30$	10	30	yes	96	0
$\lambda = 10$, $\kappa = 300$	10	300	yes	96	0
$\lambda = 10$, no warm start	10	100	no	96	0
$\lambda = 10$, half region budget	10	100	yes	48	0

Each world model takes 2.6 hours on one A40; reproducing every configuration in Table 10 takes about 22 GPU-hours. Calibration fits a shared slope and per-concept intercepts by L-BFGS on MIMIC validation with Gaussian shrinkage. Every claim-level metric uses the calibrated confidence $\sigma(\alpha e_c + \beta_c)$, or its complement for a negated claim: pooled AUROC, its within- and between-concept terms, and claim triage. Because β_c is constant within a concept, it cannot change any ordering inside a (concept, polarity) stratum, so the stratified metric (Tables 5, 8 and 11), per-concept macro AUROC (Table 4) and regionality (Tables 9, 10, and 13) are unaffected by the intercepts. It does change the ordering across concepts: the pooled and triage scores of our four signals therefore include a per-concept intercept fitted on validation labels, which the base-rate predictor and RadFlag (an uncalibrated vote fraction) do not carry.

Generator prompts. Every generator receives one frontal image and decodes greedily, with no indication section and no prior study. CheXagent-2-3b is prompted `"Describe the findings of the chest x-ray."` through its own chat template. MedGemma-4b-it is prompted `"Write the FINDINGS section of the radiology report for this chest X-ray. Use concise radiological sentences. Do not include a heading, preamble, impression, or recommendations."` The ViT–BERT baseline, DiA, and DiA-CLEAR are encoder–decoders with no text prompt. Each generator's output is inspected for model-specific markup before parsing: CheXagent emits bracketed section headers such as `[Breathing: Pleura]` that RadGraph would otherwise parse as clinical entities, and these are stripped before any claim is extracted.

Report-level confidence head. The report-level confidence head (Appendix C.10) is a 2-layer MLP trained to predict report-level RadGraph F_1 from the concatenated claim evidence and mean-pooled decoder hidden states, optimized via MSE on the validation set.