When Aggregate AUROC Ranks the Shortcut: Trivial-Axis Headroom in Vision-Language Hallucination Detection

Anonymous authors
Paper under double-blind review

Abstract

Hallucination detectors for vision-language generators draw on four signal families (cross-modal alignment, self-consistency, an external classifier, token confidence) and are compared by aggregate AUROC. Proposing no detector, we introduce two diagnostics of what that comparison measures: *headroom* over a declared *trivial axis* — a label-free shortcut that never sees the image — and an exact decomposition of AUROC into the pairs that axis orders by itself and the rest. On chest X-ray report generation, our label rule marks asserted-but-unsupported claims hallucinated and denied-but-uncontradicted ones grounded, so reading silence as absence couples the label to claim polarity. A trivial axis reading only polarity reaches 0.84–0.87 AUROC on IU X-ray and MIMIC-CXR, leaving the best family a headroom of only 0.013–0.026 (one interval covering zero), whereas on COCO the best family clears a caption mention-count axis by 0.171–0.177. Because each axis lower-bounds what label-free scores can reach, these headrooms are upper bounds. Implementation choices move each family enough to reverse their ordering, yet at their best the three externally grounded ones converge on IU X-ray. That convergence is not agreement: on chest X-rays, polarity alone orders 70–74% of hallucinated–grounded pairs; cross-modal alignment gets all of them right yet is weakest or tied-weakest among claims denying a finding; and on MIMIC-CXR a classifier trained on neither corpus ranks last in aggregate (0.474, below chance) but first among those claims, and ordering claims by polarity and then by the classifier beats every family's best aggregate. Surface scores also depart from chance on three benchmarks labelled by others — label-free on POPE and M-HalDetect, label-informed on AMBER — even sign-flipped on POPE splits drawing negatives from frequent or co-occurring objects. Where labels are partly decidable from surface form, we recommend reporting headroom over a declared trivial axis, and stratified comparisons, alongside any ranking.

1 Introduction

Detecting when a vision-language generator states something the image does not support has converged on four families of signal: cross-modal grounding between the image and the claim, self-consistency of the claim under resampling, an external discriminative classifier queried about the claim, and the generator's own token probabilities. A steady stream of work reports which family detects hallucinations best, and practitioners choose accordingly. In a specialist domain such as chest X-ray report generation, that choice decides which generated claims a reader is warned about, so what the literature supplies is the comparison between families rather than any one detector.

Such a report always fixes an implementation: how many samples self-consistency draws, whether a resampled generation re-asserts a claim under cosine similarity or entailment, which checkpoint a classifier loads. These choices are seldom stated and rarely justified against the alternatives, so a reported ordering may be a property of the configurations compared rather than of the families. Declaring them is necessary but not sufficient, because a second question stays open even when every configuration is on the table: whether the benchmark's labels leave a comparison anything to resolve. Neither quantity is routinely measured.

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.

This paper measures both. We treat the implementation as a free coordinate and report each family's *implementation interval*: the detection AUROC it spans as its configuration varies over choices that are either attested in published work or equally natural to an implementer who has not seen the labels, every one fixed before its AUROC is computed. Against that we set what the benchmark gives away before any detector runs. A *trivial axis* is a score computable from the claim's surface form alone — no image, no detector, nothing fitted to labels — and the *headroom* of a family is what its best admitted configuration buys over the strongest such score. The specialist labelling rule marks a claim hallucinated when it asserts a finding the report does not support and grounded when it denies one the report does not assert; because the rule reads silence as absence, the label is coupled to the claim's *polarity*, and a score reading nothing but polarity is a trivial axis that never opens the image. Aggregate AUROC then decomposes exactly into the pairs polarity alone settles and the pairs it does not, and comparing families within one level of that axis, where it is constant by construction, separates how well a score encodes the axis from how well it detects hallucination. We run this on five corpus–generator cells over three corpora and two domains.

In summary, this paper makes the following three contributions.

- We define the *trivial axis* of a hallucination benchmark and the *headroom* a signal family holds over it, and measure both in five corpus–generator cells across two domains: specialist headroom is 0.013, 0.025 and 0.026 on two corpora, general-domain headroom 0.171 and 0.177, a contrast that survives a label-informed text-only baseline (Section 4.4).
- We decompose aggregate AUROC exactly, by pair type as in subgroup (Borkan et al., 2019) and cross-group (Kallus & Zhou, 2019) AUC, into polarity handling and within-stratum detection, show that 70–74% of its pairs are decided by polarity alone, and show on MIMIC-CXR, with a checkpoint trained on neither corpus, that the aggregate and stratified verdicts on the discriminative classifier are opposite, while a label-free composite — polarity, then the classifier — exceeds every family's best aggregate (Section 4.4).
- We formalise the *implementation interval* of a signal family under label-blind admissibility criteria, show the intervals overlap, so the ordering can reverse, for 24 of 27 family pairs (all three exceptions on COCO), and trace the discriminative family's width on IU to checkpoint provenance (Sections 3 and 4.2, Appendix E).

2 Related work

Signal families for hallucination detection. Detectors for vision-language generators draw on four recurring sources of evidence. Cross-modal grounding scores a claim by its alignment with the image, typically through a contrastively trained encoder. Self-consistency resamples the generator and flags claims that are not re-asserted, following SelfCheckGPT (Manakul et al., 2023) and the semantic-entropy formulation of Kuhn et al. (2023). External discriminative signals query a separately trained classifier or open-vocabulary detector about the claim. Generator-internal signals read the model's own token distributions, which Zhao et al. (2024) show already carry enough information at the first generated token to tell whether the model should answer. Recent work continues to add signals of each kind, including probes constrained for stability (Luan et al., 2026), cross-modal attention patterns (Zhang et al., 2025b), and adaptive token selection (Niu et al., 2025). These papers introduce one signal and compare it against baselines they implement themselves, as do detectors built for the specialist domain we study (Zhang et al., 2025a; Hardy et al., 2025). Our object is not a fifth signal but the comparison itself: we ask what such a ranking identifies once each family's implementation is treated as a free coordinate.

Critiques of hallucination evaluation protocols. A parallel line questions the benchmarks rather than the detectors. CHAIR (Rohrbach et al., 2018) and POPE (Li et al., 2023) remain the standard protocols, and both have drawn scrutiny: Chen et al. (2024a) show that scoring the presence of a single object class per image misses hallucination at the level of individual entities, and Tu et al. (2025) replace static question sets with a dynamically generated open-set protocol to avoid contamination. These critiques target the label side — what is asked, and what counts as ground truth. We find the label side matters too, and quantify it: swapping the object lexicon that decides whether a caption mentions a COCO category moves the measured hallucination rate and changes which claims exist

at all (Appendix C). But our central object is the complementary one, the *scoring* side, where a family's configuration is left free and unreported.

Implementation sensitivity and fair comparison. We are not the first to argue that implementation choices can drive published conclusions: Oliver et al. (2018) for tuned semi-supervised baselines, Cinà et al. (2025) for adversarial attacks under incompatible setups, Ding et al. (2022) for hyperparameter sensitivity in outlier detection, and, closest to our measurement, Radosavovic et al. (2019), who argue architectures should be compared as *distributions* over a design space rather than as point estimates. Our free coordinate is a set of *post-hoc scoring decisions* requiring no retraining, which makes the interval cheap to measure.

Trivial baselines and annotation artifacts. The observation that a benchmark can be largely solved without its intended input has a long history in NLP. Gururangan et al. (2018) and Poliak et al. (2018) independently showed that natural-language inference datasets can be predicted from the hypothesis alone, and Geirhos et al. (2020) give the general account of models exploiting such shortcuts. Chen et al. (2026) bring the argument to evaluation itself, finding that surface-only predictors reproduce much of several public LLM-judge label sets, and Santilli et al. (2025) prove that a bias shared by correctness labels and uncertainty methods skews AUROC rankings. Our trivial axis is that diagnostic applied to hallucination detection, with two differences: the trivial score is not trained, and what is audited is not one method's accuracy but the *comparison* between methods — which a large trivial axis does not by itself invalidate, if the room above it is large.

Disagreement and what labels mean. Annotator disagreement is well studied as a modelling target (Guo et al., 2023; Karimijafarbigloo et al., 2026; Kraus & Kroll, 2025). Stutz et al. (2023) carry label ambiguity into a downstream guarantee, but within classification, where the label space is fixed and annotators apply one definition. The specialist domain we study differs in kind: the disagreement is between the *corpora* that supervised each checkpoint, so that nominally identical pathology names denote different constructs; Appendix E measures this disagreement directly.

3 Measuring what a signal comparison actually identifies

Work on hallucination detection is organised around four *signal families*, three consulting evidence external to the generator and one reading its own state. We show that a comparison between families is meaningful only relative to an implementation, and give a measurement that makes this explicit.

3.1 Families, configurations, and the implementation interval

A signal family $\mathcal{F}$ is a scoring recipe that is fixed only up to a configuration $\theta \in \Theta_{\mathcal{F}}$. Instantiating it yields a concrete score $s_{\mathcal{F},\theta} : \mathcal{X} \to \mathbb{R}$, higher meaning more likely hallucinated, where $\mathcal{X}$ is the set of claims to be judged. Self-consistency, for example, leaves open the sample count, the re-assertion test (cosine similarity or entailment) and the aggregation (thresholded count or continuous); each is a coordinate of θ.

Given an evaluation set D with hallucination labels, write

$$\mathrm{A}(\mathcal{F}, \theta) \;=\; \mathrm{AUROC}(s_{\mathcal{F},\theta}, D) \tag{1}$$

for the detection performance of one instantiation. A paper reports a single number $\mathrm{A}(\mathcal{F}, \hat{\theta}_{\mathcal{F}})$ at the configuration its authors chose; $\hat{\theta}_{\mathcal{F}}$ is rarely stated in full and seldom justified against alternatives.

Definition 1 (Implementation interval)**.** *The implementation interval of family $\mathcal{F}$ over an admissible configuration set $\Theta_{\mathcal{F}}$ is*

$$\mathcal{I}(\mathcal{F}) \;=\; \Big[\; \min_{\theta \in \Theta_{\mathcal{F}}} \mathrm{A}(\mathcal{F}, \theta),\;\; \max_{\theta \in \Theta_{\mathcal{F}}} \mathrm{A}(\mathcal{F}, \theta) \;\Big]. \tag{2}$$

Definition 2 (Configuration-dependent comparison)**.** *Two families $\mathcal{F}_1, \mathcal{F}_2$ compare* configuration-dependently *on D if $\mathcal{I}(\mathcal{F}_1) \cap \mathcal{I}(\mathcal{F}_2) \neq \emptyset$: whether "$\mathcal{F}_1$ detects hallucinations better than $\mathcal{F}_2$" holds is then not a property of the families but of the configurations compared.*

Definition 2 is deliberately weak — overlapping intervals only say the ordering *can* be reversed. The sharper quantity contrasts the configurations in common use with the families' best:

$$\Delta_{\text{ref}} = \max_{\mathcal{F}} \mathrm{A}(\mathcal{F}, \hat{\theta}_{\mathcal{F}}) - \min_{\mathcal{F}} \mathrm{A}(\mathcal{F}, \hat{\theta}_{\mathcal{F}}), \qquad \Delta^* = \max_{\mathcal{F}} \mathrm{A}^*(\mathcal{F}) - \min_{\mathcal{F}} \mathrm{A}^*(\mathcal{F}), \tag{3}$$

with $\mathrm{A}^*(\mathcal{F}) = \max_\theta \mathrm{A}(\mathcal{F}, \theta)$. We call $1 - \Delta^*/\Delta_{\text{ref}}$ the *compression* of the between-family effect. Compression near 1 means the apparent superiority of one family does not survive repairing the others — either because implementation produced it, or because all of them meet a ceiling the benchmark imposes, which Section 4.4 shows is what happens here; compression near 0 means the families differ for reasons implementation cannot repair.

3.2 What counts as an admissible configuration

Definition 1 is vacuous if $\Theta_{\mathcal{F}}$ is allowed to contain deliberately crippled instantiations: any family can be made arbitrarily bad. We therefore admit a configuration only when it is attested or equally natural and label-blind, and we fix $\Theta_{\mathcal{F}}$ before measuring its AUROCs.

A configuration is admitted if it appears in published work on the family, or is a choice a competent implementer would plausibly make without seeing the labels. Cosine similarity with a threshold and NLI entailment are both attested ways to decide whether a resampled caption re-asserts a claim; a similarity threshold of 0.4 versus 0.6 is unattested but equally natural, since no published work justifies its value. The configuration must also be fixed without consulting the evaluation labels. Taking the best member of that fixed set still selects on AUROC, so A^* is optimistic; this can only inflate the headroom we report. Both endpoints of $\mathcal{I}(\mathcal{F})$ are therefore reachable by an honest implementer; Section 4.1 lists the coordinates and Appendix K all IU and COCO results.

Coordinates inherited from the released model and never examined are configurations too. Qwen2-VL ships `top_k=1`, `top_p=0.001`, so K "stochastic" generations are K copies of the greedy output and self-consistency is a constant; CheXagent ships `num_beams=5`, so a nominally greedy call is beam search without per-token probabilities. Neither warns. We override decoding for every generator (`num_beams=1`, T=0.7, `top_k=50`, `top_p=1.0`) and treat finding such defaults as part of the admissibility audit.

3.3 Trivial axes and headroom

An implementation interval says how much a family's measured performance can move. It does not say whether that performance measures the detection problem at all. A label rule induces its own regularities, and a score can exploit them without ever consulting the image.

Definition 3 (Trivial axis)**.** *A scoring function $t : \mathcal{X} \to \mathbb{R}$ is* trivial *for an evaluation set D if it is computable from the surface form of the claim alone — without the image, without any learned model beyond the parser the label rule itself applies to the claim, and without consulting the evaluation labels. Write $\mathrm{A}^{\text{triv}}(D) = \max_t \mathrm{A}(t, D)$ over a declared set of trivial scores.*

Definition 4 (Headroom)**.** *The headroom of family $\mathcal{F}$ on D is $H(\mathcal{F}) = \mathrm{A}^*(\mathcal{F}) - \mathrm{A}^{\text{triv}}(D)$: what the family's best admitted configuration buys over the best trivial score. The headroom of the benchmark is $H(D) = \max_{\mathcal{F}} H(\mathcal{F})$; a family with $H(\mathcal{F}) < 0$ is outranked by a score that never sees the image.*

Headroom bounds what an *aggregate* comparison on D can resolve. If H is small for every family, the between-family spread Δ^* is a difference between quantities that are themselves nearly determined by a label-rule regularity, and the ordering it induces carries little information about detection. Two families can then tie in aggregate for entirely different reasons — one recovering evidence, the other recovering the trivial axis — which no aggregate statistic distinguishes. We therefore report, for every cell, the strongest trivial axis alongside the family intervals, and we test convergence *conditionally*: within each level of the trivial axis, where it is constant and cannot contribute.

Declaring the trivial set is a modelling choice like any other. In both domains we admit only label-free surface properties of the claim and report the maximum (Section 4.4), with a stricter, label-informed text-only baseline reported alongside. A trivial axis is not a flaw in a benchmark — most label rules admit one — but the room above it decides whether aggregate AUROC can rank them.

4 EXPERIMENTS

4.1 SETUP

We instantiate the same four signal families in a specialist and a general domain, holding the measurement protocol fixed across both.

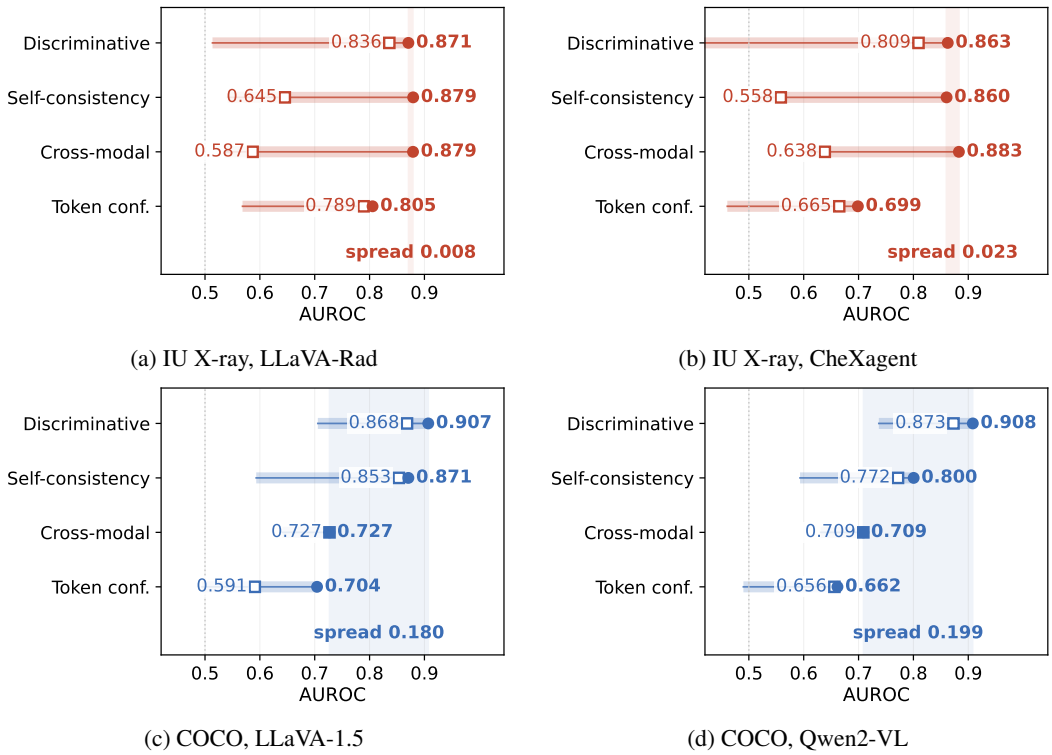
Specialist domain. LLaVA-Rad (Zambrano Chaves et al., 2025) greedily generates findings sections for Indiana University chest X-ray (IU) studies (Demner-Fushman et al., 2016). A generated sentence is a claim; it is labelled by comparing its RadGraph (Jain et al., 2021) pathology assertions against the gold report under a rule that reads silence as absence: a claim asserting a finding the report does not support is hallucinated, and a claim denying a finding the report does not assert is grounded. This yields 2,222 labelled claims, 157 hallucinated; 1,286 further sentences carry no mappable finding and are excluded. As a check against an independent labeller, a strong language model rated 215 LLaVA-Rad claims stratified across label-disagreement cells and agreed with the rule on 90.2%; Appendix J gives the rule step by step, this check, an audit of the MIMIC-CXR reference parse (Appendix J.9), the statistical procedures, data access and compute. A second specialist generator, CheXagent-8b (Chen et al., 2024b) (Q-Former + Mistral-7B, architecturally unrelated to LLaVA-Rad), generates for the same studies under identical decoding and is labelled by the same pipeline: 2,389 claims, 241 hallucinated, over 951 studies. Both generators write about one corpus, so we add a second: LLaVA-Rad under the same greedy decoding on MIMIC-CXR (Johnson et al., 2019a;b) (patient shard p10), labelled by the identical pipeline against each study's own findings section, giving 2,884 claims, 517 hallucinated, over 1,259 frontal studies (3,399 further sentences carry no mappable finding). CheXagent was also run on MIMIC-CXR, but that run inherited its checkpoint's beam-search default (Section 3.2), so we exclude it rather than mix decoding protocols. It is an inpatient corpus: 646 of those studies are portable AP films, and its hallucination rate under the label rule is 17.9% against 7.1% on IU.

General domain. LLaVA-1.5-7B (Liu et al., 2024) captions COCO val2017 images (Lin et al., 2014) under the CHAIR prompt (Rohrbach et al., 2018). A mentioned object is a claim, labelled hallucinated when it is absent from the image's human COCO annotation. Labels are therefore objective and need no judge arbitration. This yields 2,731 (image, object) claims, 779 hallucinated — five times the positive count of the specialist domain. A second general-domain generator, Qwen2-VL-7B (Wang et al., 2024), captions the same 822 images under identical decoding and yields 2,003 claims, 341 hallucinated.

Families and their configuration coordinates. Under the admissibility criteria of Section 3.2, *cross-modal alignment* (BiomedCLIP (Zhang et al., 2024); CLIP (Radford et al., 2021)) varies over the text template and over single image–text similarity versus a contrastive present-versus-absent posterior; in the specialist domain the template is filled from the RadGraph (finding, polarity) parse that forms the claim, so this family consumes the parse but not the label. *External discriminative* scoring (TorchXRayVision (Cohen et al., 2022); OWL-ViT (Minderer et al., 2022)) varies over checkpoint, probability or logit, and aggregation across a claim's targets; the TorchXRayVision checkpoint behind every best discriminative configuration (`all`) was trained on data that include both IU (as OpenI) and MIMIC-CXR, and Appendix F repeats the key comparison with three checkpoints trained on neither. *Self-consistency* varies over the number of samples K, the re-assertion rule (thresholded cosine, mean cosine, or entailment under DeBERTa-large-MNLI (He et al., 2021; Williams et al., 2018)) and the threshold. *Token confidence* varies over the aggregation of per-token probabilities and their attribution to a claim. Every configuration is fixed before its AUROC is computed, and every IU and COCO configuration appears in Appendix K. Confidence intervals are bootstrap percentiles; headroom and differences between families resample studies or images, paired, and single AUROCs resample claims.

4.2 EVERY FAMILY SPANS A WIDE IMPLEMENTATION INTERVAL

Figure 1 plots the intervals (endpoints: Appendix Table 4); their widths are 0.237–0.452 AUROC across the four families on the two IU generators and 0.078–0.277 on the two COCO generators; on MIMIC-CXR, where token confidence was not swept, the three externally-grounded families span 0.147, 0.212 and 0.399. No family is implementation-inert in either domain, and Definition 2 holds for 24 of the 27 family pairs we measure — every pair on all three chest X-ray cells. The



(a) IU X-ray, LLaVA-Rad

(b) IU X-ray, CheXagent

(c) COCO, LLaVA-1.5

(d) COCO, Qwen2-VL

Figure 1: Implementation intervals per signal family (endpoints: Appendix Table 4). Bars span the admitted configurations; open square: the reference configuration in common use (reconstructed, not quoted); filled circle: the best admitted. Band: the three externally-grounded families' spread at their best — on IU a trivial-axis ceiling (Section 4.4), on COCO not.

three exceptions are on COCO, where the open-vocabulary detector's interval lies clear of token confidence under both generators and of cross-modal alignment under Qwen2-VL (Section 4.3). At the reference configurations the discriminative classifier leads on both IU generators; at the best ones it is overtaken by cross-modal alignment (Figure 1).

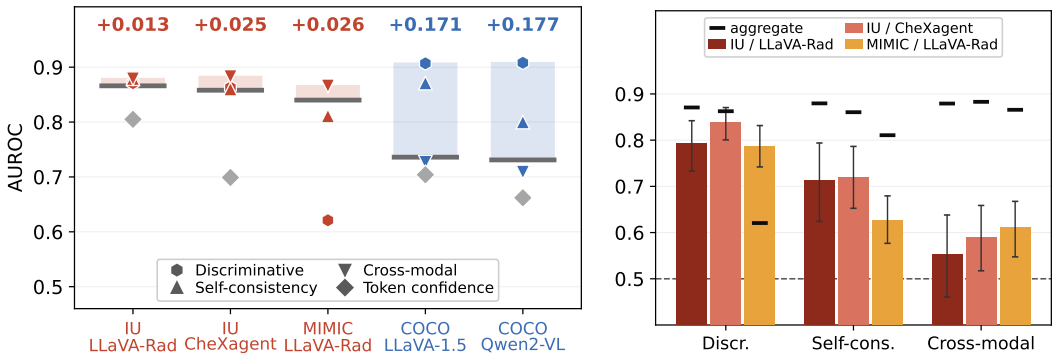
The coordinate that matters most is not the one a reader would guess. For self-consistency on IU, varying one coordinate at a time, the decision rule carries $+0.156$ AUROC (mean cosine $\rightarrow$ entailment, $K{=}10$), the aggregation $+0.067$ (thresholded count $\rightarrow$ continuous), and the sample count only $+0.014$ ($K{=}5 \rightarrow 10$) — the coordinate papers report most, K, matters least. For the discriminative family, swapping the checkpoint in common use moves AUROC by 0.322 and 0.398 on the two IU generators ($0.836 \rightarrow 0.514$; $0.809 \rightarrow 0.411$, below chance) but only 0.035–0.039 in the general domain, a gap Appendix E traces to checkpoint provenance; on MIMIC-CXR the whole discriminative interval is 0.147, narrower than either COCO detector's.

4.3 Repairing implementations makes the IU families converge

Three of the four families consult evidence external to the generator: cross-modal alignment, the discriminative classifier, and self-consistency judged by an external entailment model, which reads text rather than the image. On both IU generators these three finish within 0.008 and 0.023 of one another once each is well implemented, compressions of 97% and 91% from reference spreads of 0.249 and 0.251. On MIMIC-CXR repair instead widens the spread, from 0.123 at the reference configurations to 0.245, as on COCO: the discriminative classifier's best admitted configuration reaches 0.621 against 0.811 for self-consistency and 0.866 for cross-modal alignment. On both general generators they do not converge; repairing their implementations *widens* the gap, from 0.141 to 0.180 and from 0.164 to 0.199, because the open-vocabulary detector gains while cross-modal alignment is already at its best (Appendix Figure 10b). Token confidence, which reads the genera-

Table 1: Headroom over the strongest trivial axis (Definition 3: claim polarity on chest X-rays, caption mention count on COCO). Family columns give the best admitted configuration; headroom is the best family minus the axis, with a paired bootstrap interval over studies or images. Last column: a stricter, label-informed text-only baseline (the hallucination rate of the claim's finding and polarity, or category, on other studies or images), which uses no image but is not trivial.

		Best family				Text
Corpus / generator	Trivial	Discr.	Self-cons.	Cross-modal	Headroom [95% CI]	+ labels
IU / LLaVA-Rad	0.866	0.871	0.879	0.879	$+\mathbf{0.013}$ $[-0.019, +0.046]$	0.908
IU / CheXagent	0.858	0.863	0.860	0.883	$+\mathbf{0.025}$ $[+0.006, +0.046]$	0.884
MIMIC / LLaVA-Rad	0.840	0.621	0.811	0.866	$+\mathbf{0.026}$ $[+0.013, +0.039]$	0.878
COCO / LLaVA-1.5	0.736	0.907	0.871	0.727	$+\mathbf{0.171}$ $[+0.149, +0.192]$	0.669
COCO / Qwen2-VL	0.731	0.908	0.800	0.709	$+\mathbf{0.177}$ $[+0.149, +0.205]$	0.741



(a) Headroom over the trivial axis (grey)

(b) Absent stratum vs aggregate

Figure 2: (a) Each family's best configuration against the strongest trivial axis (grey); the shaded band is the headroom. Token confidence: not swept on MIMIC-CXR. (b) Within the absent-polarity stratum (95% CI; ticks: aggregate AUROC) the discriminative classifier keeps its signal while cross-modal alignment falls towards chance; on MIMIC-CXR the classifier goes from last in aggregate to first. Its checkpoint saw both corpora; Table 5 repeats (b) with checkpoints that saw neither.

tor's own state, follows the generator rather than the domain (ceilings 0.662–0.805), so we make no four-family claim (Appendix Figure 10a).

4.4 The IU convergence is a trivial axis, not agreement

The IU convergence of Section 4.3 is not agreement, and the room above the trivial axis is about the same size on both chest X-ray corpora.

An image-free score reaches the same ceiling. The specialist label rule marks a claim hallucinated when it asserts a finding the gold report does not support, and grounded when it denies one the gold report does not assert; because it reads silence as absence, the rule couples the label to the claim's *polarity*. It does not require references to omit their negatives — on MIMIC-CXR 66% of grounded denials deny one the reference explicitly negates — but runs through the rule's other side, where 86% of hallucinated assertions state a finding the reference leaves unmentioned. On IU, 117 of the 143 claims from LLaVA-Rad asserting a finding (81.8%) are labelled hallucinated, against 40 of the 2,079 (1.9%) denying one, and 176/206 against 65/2,183 for CheXagent. A score that reads only polarity — no image, no labels, no model beyond the rule's own parse — therefore attains AUROC 0.866 and 0.858 (Definition 3), and the best family clears it by 0.013 $[-0.019, 0.046]$ and 0.025 $[0.006, 0.046]$ (Table 1, Figure 2a), the first interval covering zero. MIMIC-CXR differs from IU in every quantity that fixes a binary score's AUROC: claims asserting a finding are 22.9% of the labelled set rather than 6.4%, 61.6% of them are hallucinated rather than 81.8%, and 4.9% of the

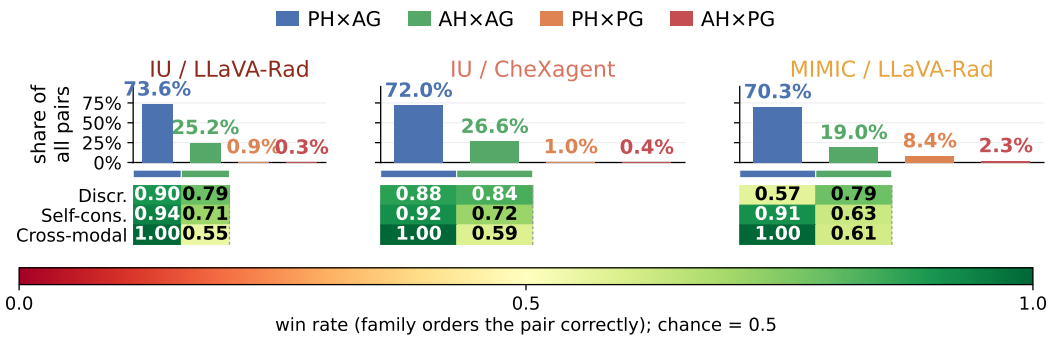


Figure 3: Aggregate AUROC is a fixed mixture. Bars: each pair type's share of all hallucinated-versus-grounded pairs (P/A: the claim asserts or denies a finding; H/G: hallucinated or grounded). Grid: the fraction each family orders correctly on PH×AG (left column) and AH×AG (right), the two types carrying 99% of the mass on IU and 89% on MIMIC-CXR, one scale centred at chance. One type carries about 0.7 of the mass; cross-modal alignment wins essentially all of it and is weakest on the absent-stratum type carrying most of the rest.

rest rather than 1.9%. Polarity still reaches 0.840, the best family clears it by 0.026 $[0.013, 0.039]$, and the best discriminative and self-consistency configurations fall below it.

The general domain gets the same check. For COCO we admit the label-free surface properties we could construct without the image, category prior frequency and mention count; mention count is the stronger, at 0.736 and 0.731, and the best family clears it by 0.171 and 0.177, taking 65% and 66% of the room above the axis against 10%–18% on chest X-rays. A stricter text-only baseline allowed to learn from labels (last column of Table 1) matches or exceeds every family on chest X-rays and falls far short of the best family on COCO.

Where aggregate AUROC gets its number. The ceiling has an exact accounting. Every pair AUROC averages over sets a hallucinated claim against a grounded one, and 73.6%, 72.0% and 70.3% of those pairs set a hallucinated claim asserting a finding against a grounded claim denying one, which polarity alone orders correctly. Aggregate AUROC is therefore a fixed mixture: about 0.7 parts how well a score encodes polarity, 0.19–0.27 parts how well it detects hallucination among claims that deny a finding, and remaining terms worth 1% of the pairs on IU but 11% on MIMIC-CXR. Cross-modal alignment wins nearly every polarity-decided pair in each cell — its template is filled from the claim's parsed polarity (Section 4.1), so as a score it orders claims by polarity with AUROC 1.000 — and that term supplies most of its aggregate (Figure 3; all terms: Appendix G). The families converge on IU because all three encode polarity well there (score-to-polarity AUROC 0.86–1.00), not because they agree; on MIMIC-CXR the classifier does not (0.54) and is last.

Under the ceiling, the families are doing different things. Within the absent-polarity stratum, where polarity is constant, cross-modal alignment falls to 0.552 $[0.461, 0.638]$, 0.589 $[0.517, 0.659]$ and 0.611 $[0.547, 0.668]$, while the discriminative classifier retains 0.793, 0.839 and 0.787 and leads it in every cell with an interval excluding zero (Figure 2b). Because that checkpoint was trained on data including both corpora, we repeat the comparison with three checkpoints trained on neither (Table 2); no IU cell separates it from self-consistency. Two caveats: the absent stratum holds only 21%–27% of the hallucinations (in the present stratum no family reaches 0.71), and it has label-free shortcuts too — on MIMIC-CXR the film's AP-view flag reaches 0.606 within it. Families are stratified at aggregate-selected configurations.

On the second corpus, the aggregate verdict is reversed. MIMIC-CXR shows the consequence with a checkpoint that never saw it. Scored by the CheXpert checkpoint, the discriminative classifier is last in aggregate on the claims it can score, at 0.474 — below chance, because it orders the polarity-decided pairs worse than chance (0.363) — against 0.808 and 0.862 for self-consistency and cross-modal alignment; within the absent-polarity stratum it is first, at 0.760 against 0.627 and 0.610, both intervals excluding zero. With all six checkpoints at the common configuration of

Table 2: Checkpoints trained on neither corpus: aggregate AUROC, AUROC within the absent-polarity stratum, and the stratum lead over cross-modal alignment (*: paired bootstrap 95% interval excludes zero). Six of nine leads exclude zero, all share a sign; all six checkpoints and the leads over self-consistency: Table 5.

	IU / LLaVA-Rad			IU / CheXagent			MIMIC / LLaVA-Rad		
Checkpoint	Aggr.	Stratum	$-$CM	Aggr.	Stratum	$-$CM	Aggr.	Stratum	$-$CM
NIH	0.770	0.722	$+0.031$	0.800	0.783	$+0.089$	0.595	0.704	$+0.113^*$
PadChest	0.712	0.825	$+0.164^*$	0.725	0.869	$+0.176^*$	0.577	0.655	$+0.081$
CheXpert	0.650	0.708	$+0.156^*$	0.683	0.709	$+0.119^*$	0.474	0.760	$+0.149^*$

Table 3: Surface shortcuts on three benchmarks we did not build (no image, no model; label-free on POPE and M-HalDetect, label-informed on AMBER); every interval excludes chance.

Split	n	AUROC [95% CI]	$\lvert\text{AUROC} - 0.5\rvert$
POPE (Li et al., 2023) — axis: the named object's COCO category prior			
random	3,000	0.769 [0.753, 0.785]	**0.269**
popular	3,000	0.277 [0.256, 0.297]	**0.223**
adversarial	3,000	0.416 [0.395, 0.437]	**0.084**
AMBER (Wang et al., 2023) — axis: the object's in-sample truth rate (label-informed)			
generative	10,523	0.837 [0.830, 0.844]	**0.337**
M-HalDetect (Gunjal et al., 2024) — axis: the span's position in the response			
validation	16,450	0.674 [0.665, 0.683]	**0.174**
training	57,083	0.675 [0.670, 0.679]	**0.175**

Table 5, the MIMIC-CXR aggregate is at most 0.60 (0.621 at its best admitted one) and the classifier is first within the stratum, though for the two clean checkpoints that score fewer claims not every lead excludes zero (Table 5). The signal is recoverable label-free: ordering claims by polarity, then by any discriminative checkpoint's score, beats polarity alone by 0.03–0.09 AUROC and, for every checkpoint, every family's best aggregate on the same claims (clean CheXpert on MIMIC-CXR: 0.894 vs 0.862; *all*: 0.941, 0.950, 0.902, Table 6).

Surface shortcuts are not peculiar to our labelling. Every axis above uses a label rule this paper built; on three benchmarks fixed by other authors, each probed by a different surface feature, a surface score is exploitable anyway, sign-flipped on POPE's popular and adversarial splits (Table 3; Appendix B). These measure a shortcut, not the headroom, which needs a generator's own outputs and all four families on those images (future work).

5 Conclusion

Aggregate AUROC here mostly ranks how well a score encodes a label-free axis: an AUROC of 0.88 on the specialist benchmark sounds like a working detector; it is 0.01–0.03 above a rule that never sees the image. The axis follows from a label rule that reads silence as absence; references that state their negatives do not remove it (MIMIC-CXR's mostly do; its axis still reaches 0.840), and balanced benchmarks pairing present and absent objects keep a surface shortcut too, sign-flipped (Table 3). Small headroom does not mean the benchmark cannot separate detectors — within a polarity stratum it can — but it rules out ranking them by aggregate AUROC. We therefore recommend reporting, alongside any ranking, headroom over a declared trivial axis, stratified comparisons, $\mathcal{I}(\mathcal{F})$ over a declared admissible set, and a medical classifier's checkpoint and training data (Appendix E). Stratifying by polarity removes the axis but has a floor of its own (0.606 from the AP-view flag in the MIMIC-CXR absent stratum) to declare.

Configurations transfer between the chest X-ray corpora (within 0.004 AUROC); levels do not ($0.871 \rightarrow 0.617$ for the discriminative classifier, Appendix J.10). Our scope is limited: the second chest X-ray corpus has one generator and no token-confidence sweep; our best discriminative checkpoint saw both corpora in training (Appendix F); the trivial axis is a lower bound, so our headroom is an upper bound for the admitted configurations; COCO differs in claim unit, label source and prevalence (a descriptive contrast); and the specialist labels were checked by a language model on IU only (and 200 MIMIC-CXR reference parses, Appendix J.9), never by radiologists.

AI USE STATEMENT

In this work, we used generative AI tools for language editing and to improve the clarity and readability of the manuscript. As part of the methodology, we used a large language model, Claude, to independently score the claims in Section 4.1 and the two readers described in Section 1.9, with full details provided in the preregistration. We did not use generative AI to generate synthetic datasets or to generate the prompts used in the study.

ETHICS STATEMENT

This work analyses existing data and collects none from human subjects. IU X-ray is a de-identified public collection, MIMIC-CXR was accessed through PhysioNet credentialed access under its data use agreement, and COCO is public. The findings concern how hallucination detectors for radiology report generation are evaluated: a detector whose aggregate AUROC looks strong on these benchmarks may add little over a rule that never sees the image, which is a caution against deploying such detectors on the strength of aggregate scores rather than a new capability.

REPRODUCIBILITY STATEMENT

The configuration coordinates are listed in Section 4.1, decoding settings in Section 3.2, and the result of every IU and COCO configuration in Appendix K; the clean-checkpoint comparison is in Appendix F. All generators, detectors, encoders and the entailment model are public checkpoints, and all three corpora are public or available under credentialed access. Every number in the paper is produced by a script from stored per-claim or per-configuration results; the analysis code, the labelling module of Appendix J.1, every aggregate result file and the COCO per-claim scores are provided as anonymized supplementary material; other per-claim and generated-text files are withheld (MIMIC-CXR's under its data use agreement) and regenerate from the included scripts, MIMIC-CXR's under credentialed access, except the readers' labels of Appendix J.9, which regenerate by re-running two readers with the instruction included in the package; decoding settings and prompts are given in Section 3.2 and Appendix J.7.

REFERENCES

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, pp. 491–500, 2019. doi: 10.1145/3308560.3317593.

Jianlin Chen, Wenhui Chen, Ziyao Lin, and Chi Man Vong. A judge should know what changed: Construct validity for LLM-as-a-judge evaluation. *arXiv preprint arXiv:2608.24419*, 2026.

Xuweiyi Chen, Ziqiao Ma, Xuejun Zhang, Sihan Xu, Jianing Yang, David F. Fouhey, Joyce Chai, and Shengyi Qian. Multi-object hallucination in vision language models. In *Advances in Neural Information Processing Systems*, 2024a. doi: 10.52202/079017-1409.

Zhihong Chen, Maya Varma, Justin Xu, Magdalini Paschali, Dave Van Veen, Andrew Johnston, Alaa Youssef, Louis Blankemeier, Christian Bluethgen, Stephan Altmayer, et al. A vision-language foundation model to enhance efficiency of chest x-ray interpretation. *arXiv preprint arXiv:2401.12208*, 2024b. doi: 10.48550/arXiv.2401.12208.

Antonio Emanuele Cinà, Jérôme Rony, Maura Pintor, Luca Demetrio, Ambra Demontis, Battista Biggio, Ismail Ben Ayed, and Fabio Roli. AttackBench: Evaluating gradient-based attacks for adversarial examples. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 2600–2608, 2025. doi: 10.1609/aaai.v39i3.32263.

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

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. doi: 10.1093/jamia/ocv080.

Xueying Ding, Lingxiao Zhao, and Leman Akoglu. Hyperparameter sensitivity in deep outlier detection: Analysis and a scalable hyper-ensemble solution. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. doi: 10.1038/s42256-020-00257-z.

Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 18135–18143, 2024. doi: 10.1609/aaai.v38i16.29771.

Hui Guo, Boyu Wang, and Grace Yi. Label correction of crowdsourced noisy annotations with an instance-dependent noise transition model. In *Advances in Neural Information Processing Systems*, 2023. doi: 10.52202/075280-0018.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2018. doi: 10.18653/v1/N18-2017.

Romain Hardy, Sung Eun Kim, Du Hyun Ro, and Pranav Rajpurkar. ReXTrust: A model for fine-grained hallucination detection in AI-generated radiology reports. In *Proceedings of The First AAAI Bridge Program on AI for Medicine and Healthcare*, volume 281 of *Proceedings of Machine Learning Research*, pp. 173–182. PMLR, 2025.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations*, 2021.

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

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 *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, volume 1, 2021.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-Ying Deng, Roger G. Mark, and Steven Horng. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6:317, 2019a. doi: 10.1038/s41597-019-0322-0.

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

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

Sanaz Karimijafarbigloo, Armin Khosravi, Alireza Kheyrkhah, Reza Azad, Mauricio Reyes, and Dorit Merhof. Harmonized feature conditioning and frequency-prompt personalization for multi-rater medical segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Kelsey Kraus and Margaret Kroll. Maximizing signal in human-model preference alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025. doi: 10.1609/aaai.v39i26.34950.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*, 2023.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, Singapore, 2023. doi: 10.18653/v1/2023.emnlp-main.20.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755, 2014.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. doi: 10.1109/cvpr52733.2024.02484.

Bozhi Luan, Gen Li, Yalan Qin, Jifeng Guo, Yun Zhou, Faguo Wu, Hongwei Zheng, Wenjun Wu, and Zhaoxin Fan. Lyapunov probes for hallucination detection in large foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017, Singapore, 2023. doi: 10.18653/v1/2023.emnlp-main.557.

Matthias Minderer, Alexey A. Gritsenko, Austin V. Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. Simple open-vocabulary object detection with vision transformers. In *Computer Vision – ECCV 2022*. Springer, 2022. doi: 10.1007/978-3-031-20080-9_42.

Mengjia Niu, Hamed Haddadi, and Guansong Pang. Robust hallucination detection in LLMs via adaptive token selection. In *Advances in Neural Information Processing Systems*, 2025. doi: 10.52202/085713-4213.

Avital Oliver, Augustus Odena, Colin Raffel, Ekin D. Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, 2018. doi: 10.18653/v1/S18-2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Ilija Radosavovic, Justin Johnson, Saining Xie, Wan-Yen Lo, and Piotr Dollár. On network design spaces for visual recognition. In *IEEE/CVF International Conference on Computer Vision*, pp. 1882–1890, 2019. doi: 10.1109/iccv.2019.00197.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4035–4045, Brussels, Belgium, 2018. doi: 10.18653/v1/D18-1437.

Andrea Santilli, Adam Golinski, Michael Kirchhof, Federico Danieli, Arno Blaas, Miao Xiong, Luca Zappella, and Sinead Williamson. Revisiting uncertainty quantification evaluation in language models: Spurious interactions with response length bias results. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 743–759, Vienna, Austria, 2025. doi: 10.18653/v1/2025.acl-short.60.

David Stutz, Abhijit Guha Roy, Tatiana Matejovicova, Patricia H. Strachan, Ali Taylan Cemgil, and Arnaud Doucet. Conformal prediction under ambiguous ground truth. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=CAd6V2qXxc.

Yahan Tu, Rui Hu, and Jitao Sang. ODE: Open-set evaluation of hallucinations in multimodal large language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. doi: 10.1109/cvpr52734.2025.01847.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. AMBER: An LLM-free multi-dimensional benchmark for MLLMs hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. doi: 10.48550/arXiv.2409.12191.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. ChestX-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2097–2106, 2017. doi: 10.1109/CVPR.2017.369.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2018. doi: 10.18653/v1/N18-1101.

Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Sheng Zhang, Fei Wang, Yujia Xie, Mahmoud Khademi, Ziyi Yang, et al. A clinically accessible small multimodal radiology model and evaluation metric for chest x-ray findings. *Nature Communications*, 16:3108, 2025. doi: 10.1038/s41467-025-58344-x.

Serena Zhang, Sraavya Sambara, Oishi Banerjee, Julian N. Acosta, L. John Fahrner, and Pranav Rajpurkar. RadFlag: A black-box hallucination detection method for medical vision language models. In *Proceedings of the 4th Machine Learning for Health Symposium*, volume 259 of *Proceedings of Machine Learning Research*, pp. 1087–1103. PMLR, 2025a.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. A multimodal biomedical foundation model trained from fifteen million image–text pairs. *NEJM AI*, 2024. doi: 10.1056/AIoa2400640.

Yudong Zhang, Ruobing Xie, Xingwu Sun, Yiqing Huang, Jiansheng Chen, Zhanhui Kang, Di Wang, and Yu Wang. DHCP: Detecting hallucinations by cross-modal attention pattern in large vision-language models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025b. doi: 10.1145/3746027.3755118.

Qinyu Zhao, Ming Xu, Kartik Gupta, Akshay Asthana, Liang Zheng, and Stephen Gould. The first to know: How token distributions reveal hidden knowledge in large vision-language models? In *Computer Vision – ECCV 2024*, Lecture Notes in Computer Science. Springer, 2024. doi: 10.1007/978-3-031-73195-2_8.

A EXACT IMPLEMENTATION INTERVALS

Figure 1 in the body plots the intervals; their endpoints and spreads, including the MIMIC-CXR cell the figure omits, are tabulated below.

Table 4: Implementation intervals per signal family. "Reference" is the configuration in common use, which we reconstruct rather than quote; the interval runs over all admitted configurations. Δ_{ref} and Δ^* are the spreads among the three externally-grounded families at the reference and best settings; token confidence is excluded (Section 4.3).

Domain	Signal family	Reference	Interval	Span
Specialist (IU / LLaVA-Rad)	Discriminative (TXRV)	0.836	[0.514, 0.871]	0.357
	Self-consistency	0.645	[0.541, 0.879]	0.338
	Cross-modal (BiomedCLIP)	0.587	[0.536, 0.879]	0.343
	Token confidence	0.789	[0.568, 0.805]	0.237
	$\Delta_{\mathrm{ref}} = 0.249 \rightarrow \Delta^* = 0.008$	(97% compression)		
Specialist (IU / CheXagent)	Discriminative (TXRV)	0.809	[0.411, 0.863]	0.452
	Self-consistency	0.558	[0.513, 0.860]	0.347
	Cross-modal (BiomedCLIP)	0.638	[0.603, 0.883]	0.280
	Token confidence	0.665	[0.461, 0.699]	0.238
	$\Delta_{\mathrm{ref}} = 0.251 \rightarrow \Delta^* = 0.023$	(91% compression)		
Specialist (MIMIC / LLaVA-Rad)	Discriminative (TXRV)	0.596	[0.474, 0.621]	0.147
	Self-consistency	0.615	[0.599, 0.811]	0.212
	Cross-modal (BiomedCLIP)	0.492	[0.467, 0.866]	0.399
	Token confidence	not swept		
	$\Delta_{\mathrm{ref}} = 0.123 \rightarrow \Delta^* = 0.245$	(widens)		
General (COCO / LLaVA-1.5)	Discriminative (OWL-ViT)	0.868	[0.706, 0.907]	0.201
	Self-consistency	0.853	[0.593, 0.871]	0.277
	Cross-modal (CLIP)	0.727	[0.641, 0.727]	0.086
	Token confidence	0.591	[0.495, 0.704]	0.210
	$\Delta_{\mathrm{ref}} = 0.141 \rightarrow \Delta^* = 0.180$	(widens)		
General (COCO / Qwen2-VL)	Discriminative (OWL-ViT)	0.873	[0.737, 0.908]	0.171
	Self-consistency	0.772	[0.593, 0.800]	0.207
	Cross-modal (CLIP)	0.709	[0.631, 0.709]	0.078
	Token confidence	0.656	[0.490, 0.662]	0.171
	$\Delta_{\mathrm{ref}} = 0.164 \rightarrow \Delta^* = 0.199$	(widens)		

B SHORTCUTS ON THREE BENCHMARKS WE DID NOT CONSTRUCT

Every trivial axis in the body is measured against a label rule this paper built. The strongest objection to that is that the shortcut is a property of our construction, not of hallucination benchmarks generally. Three existing benchmarks — POPE, AMBER and M-HalDetect — let us check without building anything: their images, prompts and labels are all fixed by other authors, and each is checked here by a trivial score built from a different surface feature, so this is three independent tries at the same question rather than one construction repeated. **On every one, a surface score with no image and no model is exploitable (label-free on POPE and M-HalDetect, label-informed on AMBER); on POPE it survives even the two splits whose negatives are drawn from frequent or co-occurring objects, with its sign flipped.** Table 3 in the body collects every number; the paragraphs below give each benchmark's construction and the caveats that go with it.

POPE (Li et al., 2023): object identity, and a sign flip under adversarial sampling. POPE samples its negatives three ways (*random*, *popular*, *adversarial*), each 3,000 yes/no questions of the form "Is there a {object} in the image?", exactly balanced. The COCO category-prior score already used for CHAIR (Section 4.4) transfers directly: the marginal frequency with which {object} appears in any COCO image, read off the object named in the question, with no image opened and no

model run. On *random* this reaches AUROC 0.769 $[0.753, 0.785]$ — the same order as the specialist polarity axis (0.840–0.866). On *popular* and *adversarial* it reaches 0.277 $[0.256, 0.297]$ and 0.416 $[0.395, 0.437]$. **Below chance is not absent: all three splits are exploitable, the harder two with the same feature's sign flipped, not removed.** POPE's harder splits sample negatives from categories that are themselves frequent (popular) or that co-occur with what the image contains (adversarial), which inflates the prior of the negatives past that of the positives and flips which answer the score favours; $|\mathrm{AUROC} - 0.5|$ — 0.269, 0.223 and 0.084, all three intervals excluding 0 — is the right measure of exploitability, not the AUROC's raw value or direction. Casting this into the paper's own headroom terms would need a generator's actual answers and the four families' scores, both of which need images POPE specifies from COCO's 2014 split; only 74 of the 500 referenced images are among the $5{,}000$ this paper holds (2017 split), ruling out a fuller replication.

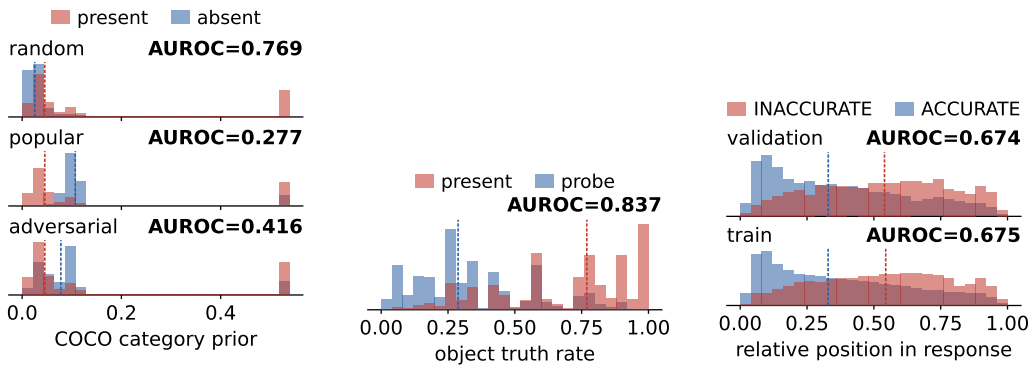
AMBER (Wang et al., 2023): object identity, no sign flip. AMBER's own discriminative-existence query set cannot be used the same way — every one of its $4{,}924$ ground-truth labels is "no" by construction (it is a fixed hallucination probe set; the hallucination is a property of a model's answer, which we do not have). Its *generative*-type annotations do work: one entry per image ($1{,}004$ images) giving a `truth` list (objects actually present) and a `hallu` list (plausible-but-absent objects AMBER's authors chose as probes). Pooling every (image, object) pair implied by `truth` $\cup$ `hallu` recovers POPE's shape, and each object's own global truth rate across all $1{,}004$ images — how often, across the dataset, a candidate named by that word turns out to be real rather than a probe — is the AMBER-native analogue of the COCO prior, again with no image and no model. It is not, however, label-free in the sense of Definition 3: unlike POPE's COCO category prior, which is an external corpus frequency, this rate is estimated in-sample from the very labels it is then scored against, so objects appearing once take their own label as their score. It is therefore label-informed, of a piece with the text-only baseline in the last column of Table 1 rather than with the trivial axes, and the weakest of the three checks here. Read with that caveat, over $10{,}523$ such pairs it reaches AUROC $\mathbf{0.837}$ $[0.830, 0.844]$, in the same direction throughout and with no sign-flip complication.

M-HalDetect (Gunjal et al., 2024): a different mechanism, in free text in place of a fixed template. Both checks above rely on a fixed question template naming one object; M-HalDetect has none — Scale AI annotators labelled sub-sentence spans of InstructBLIP-generated captions of COCO images as `ACCURATE`, `INACCURATE`, `ANALYSIS` or `UNSURE`, so its labels are human judgements of free text, closer in shape to this paper's own claims than to a yes/no template. That shape rules out an object-identity prior but opens a different trivial axis: a span's own position in the response, divided by the response's length, dropping `ANALYSIS`/`UNSURE` spans (not a presence judgement). Over $16{,}450$ `ACCURATE`/`INACCURATE` spans in the validation split this reaches AUROC 0.674 $[0.665, 0.683]$, and 0.675 $[0.670, 0.679]$ over $57{,}083$ spans in the training split — weaker than the other two, but the two splits agree closely on the largest sample used in this appendix, and both intervals exclude chance by a wide margin. Hallucinated spans are not randomly placed within a response; they cluster later, a pattern that needs no knowledge of what the response is about, no category vocabulary, and no image.

What three tries establish. None of the three replicates this paper's full headroom analysis, which needs a generator's own answers and all four families' scores on each benchmark's images; that is future work. What they do establish is that a surface-feature shortcut is not a property of the labelling rule audited in the body — the score reading that feature being label-free on POPE and M-HalDetect and label-informed on AMBER: it recurs across three benchmarks with three different label sources (POPE's yes/no questions under three negative-sampling schemes, AMBER's per-image object annotations, and free text judged by paid human annotators) and three different surface scores (two keyed to object identity, one to response position). On POPE's two splits whose negatives are drawn from frequent or co-occurring objects, the shortcut shrinks to $|\mathrm{AUROC} - 0.5| \in \{0.223, 0.084\}$ rather than to zero. Figure 4 shows the three surface scores by label class; on POPE's popular and adversarial splits the negatives sit to the right of the positives, which is the sign flip.

C LABEL DEFINITIONS MOVE THE MEASUREMENT TOO

The implementation interval has a counterpart on the label side. In the general domain, swapping the object lexicon that decides whether a caption "mentions" a COCO category — from the con-



(a) POPE: COCO category prior (b) AMBER: object truth rate (c) M-HalDetect: span position

Figure 4: The surface scores of Table 3, by label class (dashed lines: class medians; script `r087`). (a) POPE: the prior frequency of the named object's COCO category, for objects present ("yes") and absent ("no"); the tall bar at 0.53 is *person*. On the random split present objects have the higher prior; on the popular and adversarial splits, whose negatives are drawn from frequent or co-occurring categories, absent objects do, and AUROC falls below 0.5. (b) AMBER: the in-sample truth rate of the object, for present objects and probes. (c) M-HalDetect: a span's relative position in the response; inaccurate spans cluster later, on both splits.

ventional list with broad synonyms to a strict name-plus-plural list — changes the claim count from 2,731 to 2,217 and the measured CHAIR hallucination rate from 28.5% to 30.4%, with 18.8% of mentions existing under only one lexicon. Both lexicons are defensible and neither is standard practice to report. The standard lexicon also manufactures hallucinations outright — the COCO category *orange* is matched by the colour adjective, so "orange flowers" is scored as a hallucinated fruit — and does so at a rate that depends on the generator, so we exclude the category for both (Appendix D). The specialist domain shows the same pattern more severely. On a 443-claim sample from LLaVA-Rad on IU, replacing the RadGraph labelling with one based on semantic similarity to the reference sentence moves the discriminative signal's AUROC from 0.828 to 0.541, and the two labellings agree on only 67.7% of claims. The independent ratings described in Section 4.1 agreed with the RadGraph labels 90.2% of the time and with the semantic labels 34.9%, which is why we use the former throughout.

Figure 5 makes the specialist rule concrete: one real generated sentence from each of the four (polarity, label) quadrants it produces, on IU X-ray.

D THE COCO *orange* CATEGORY

The object lexicon conventionally used with the CHAIR protocol matches the COCO category *orange* against the colour adjective, so a caption reading "orange flowers" is scored as hallucinating a fruit. The effect is large enough to change a between-generator comparison. It accounts for 59% of Qwen2-VL's and 67% of LLaVA-1.5's *orange* hallucinations, and because Qwen2-VL describes colour more often, for 9% of *all* its positives against 1% for LLaVA-1.5. A reader comparing the two generators' hallucination rates under this lexicon would be reading, in part, a difference in how often each mentions colour. All general-domain numbers in the body exclude the category.

E WHY CHECKPOINT CHOICE MATTERS MORE ON IU X-RAY

A family's configuration frequently includes *which pretrained checkpoint to use*. When the available checkpoints were trained on a common corpus under a common target definition, this coordinate of θ is nearly inert; when their provenance differs, swapping the checkpoint silently swaps the quantity being measured. The available checkpoints differ more in the specialist domain. General-domain open-vocabulary detectors are released as sizes and recipes trained toward one objective over one

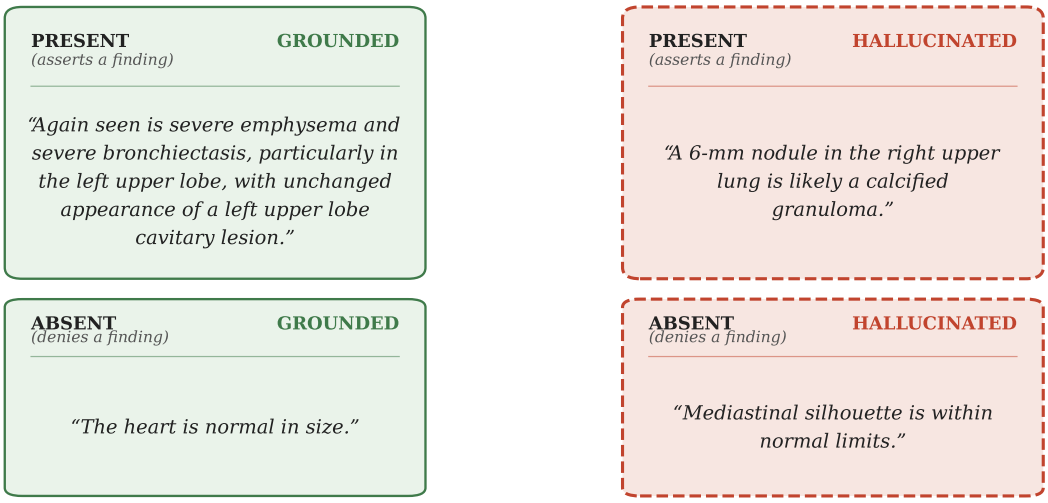


Figure 5: One real example claim per (polarity, label) quadrant, from LLaVA-Rad on IU X-ray. Whether a claim is hallucinated tracks whether it asserts or denies a finding — these four are illustrative, not a sample, but the same coupling holds in aggregate (Section 4.4).

object inventory, whereas medical classifiers are released per corpus, and those corpora differ both in their images and in the tools used to extract labels from free-text reports.

This yields a falsifiable prediction, tested below: if checkpoint choice dominates because provenance differs, then *the checkpoints themselves must disagree*, and must disagree more in the specialist domain. We measure pairwise rank correlation between checkpoints' scores on identical claims, and state in advance that agreement comparable to the general domain would refute the explanation. This mechanism concerns the *width* of one family's interval; it does not explain where the families' best configurations end up, which Section 4.4 addresses.

Test. The account predicts that specialist checkpoints disagree with one another more than general-domain detectors do, and they do (Figure 11): mean pairwise Spearman correlation between per-claim scores is 0.486 across six medical checkpoints (15 pairs, range 0.152–0.828) against 0.663 across three open-vocabulary detectors (0.629–0.701). The structure locates the cause, and not where we expected. The one pair that shares a corpus and differs only in the labeller that read its reports — the two MIMIC-CXR (Johnson et al., 2019a) checkpoints, supervised by NegBio and by CheXpert annotations — agrees at 0.828, the highest of the fifteen. Every cross-corpus pair agrees less, from 0.738 (CheXpert (Irvin et al., 2019) against MIMIC-CXR) down to 0.152 (NIH ChestX-ray14 (Wang et al., 2017) against MIMIC-CXR). Provenance dominates, but through the training corpus rather than the text-extraction rule, which refines the hypothesis rather than confirming it.

Scope of this account. It is an account of IU X-ray. On MIMIC-CXR the discriminative interval spans only 0.147, narrower than either COCO detector's, and the two checkpoints trained on MIMIC-CXR itself score 0.503 and 0.553 there at the reference configuration. The `all` checkpoint, behind the best discriminative configurations on both corpora, was trained on data that include both; Appendix F repeats the within-stratum comparison without it.

F The within-stratum comparison with checkpoints trained on neither corpus

Every best discriminative configuration in the paper uses TorchXRayVision's `all` checkpoint, trained on data that include IU X-ray (as OpenI) and MIMIC-CXR. Table 5 repeats the absent-polarity comparison of Section 4.4 with each of the six checkpoints at the one configuration swept for all of them, on the claims each checkpoint can score. The discriminative classifier's lead over cross-modal alignment does not depend on the overlap; its lead over self-consistency does not sur-

vive it on IU. The overlap is visible in aggregate instead: on IU the `all` checkpoint orders the polarity-decided pairs better than any checkpoint trained on neither corpus, so IU aggregate numbers for the discriminative family are likely optimistic.

Table 5: The absent-polarity comparison with every TorchXRayVision checkpoint, at maximum aggregation, probability, centre-cropped, on the claims each checkpoint can score; self-consistency (SC) and cross-modal alignment (CM) are evaluated on the same claims. †: trained on neither IU X-ray nor MIMIC-CXR. Differences are study-clustered paired bootstrap means with 95% intervals; *: interval excludes zero. The *all* row is at this common configuration, not its best admitted one, so its stratum values (0.787, 0.793, 0.795) differ from the best-configuration ones quoted in Section 4.4.

				Absent stratum		Discr. minus	
Checkpoint	Claims	Hall.	Aggregate	Discr.	SC / CM	SC	CM
IU / LLaVA-Rad							
NIH†	1,546	22	0.770	0.722	0.780 / 0.682	-0.060 [-0.16,$+0.03$]	$+0.031$ [-0.10,$+0.15$]
PadChest†	1,687	22	0.712	0.825	0.758 / 0.659	$+0.067$ [-0.02,$+0.17$]	$+0.164$ [$+0.06$,$+0.27$]*
CheXpert†	2,205	40	0.650	0.708	0.712 / 0.552	-0.003 [-0.13,$+0.12$]	$+0.156$ [$+0.07$,$+0.24$]*
MIMIC (NegBio)	2,205	40	0.525	0.720	0.712 / 0.552	$+0.010$ [-0.11,$+0.14$]	$+0.168$ [$+0.08$,$+0.25$]*
MIMIC (CheXpert)	2,205	40	0.514	0.700	0.712 / 0.552	-0.014 [-0.14,$+0.11$]	$+0.147$ [$+0.06$,$+0.24$]*
all	2,222	40	0.836	0.787	0.712 / 0.552	$+0.075$ [-0.02,$+0.17$]	$+0.233$ [$+0.14$,$+0.31$]*
IU / CheXagent							
NIH†	1,560	27	0.800	0.783	0.804 / 0.692	-0.021 [-0.13,$+0.09$]	$+0.089$ [-0.05,$+0.21$]
PadChest†	1,561	27	0.725	0.869	0.804 / 0.692	$+0.064$ [-0.03,$+0.17$]	$+0.176$ [$+0.05$,$+0.30$]*
CheXpert†	2,331	65	0.683	0.709	0.719 / 0.589	-0.010 [-0.10,$+0.09$]	$+0.119$ [$+0.05$,$+0.19$]*
MIMIC (NegBio)	2,331	65	0.532	0.716	0.719 / 0.589	-0.003 [-0.09,$+0.08$]	$+0.127$ [$+0.06$,$+0.20$]*
MIMIC (CheXpert)	2,331	65	0.411	0.641	0.719 / 0.589	-0.078 [-0.16,$+0.01$]	$+0.053$ [-0.03,$+0.13$]
all	2,389	65	0.809	0.793	0.719 / 0.589	$+0.076$ [-0.00,$+0.16$]	$+0.203$ [$+0.13$,$+0.28$]*
MIMIC / LLaVA-Rad							
NIH†	2,056	54	0.595	0.704	0.611 / 0.590	$+0.092$ [-0.01,$+0.18$]	$+0.113$ [$+0.02$,$+0.20$]*
PadChest†	2,177	54	0.577	0.655	0.586 / 0.573	$+0.066$ [-0.03,$+0.15$]	$+0.081$ [-0.01,$+0.17$]
CheXpert†	2,847	110	0.474	0.760	0.627 / 0.610	$+0.134$ [$+0.08$,$+0.19$]*	$+0.149$ [$+0.09$,$+0.21$]*
MIMIC (NegBio)	2,847	110	0.503	0.764	0.627 / 0.610	$+0.137$ [$+0.08$,$+0.20$]*	$+0.154$ [$+0.10$,$+0.21$]*
MIMIC (CheXpert)	2,847	110	0.553	0.733	0.627 / 0.610	$+0.104$ [$+0.05$,$+0.16$]*	$+0.121$ [$+0.06$,$+0.18$]*
all	2,884	110	0.596	0.795	0.627 / 0.611	$+0.167$ [$+0.12$,$+0.22$]*	$+0.185$ [$+0.13$,$+0.24$]*

G ALL TERMS OF THE PAIR DECOMPOSITION

Section 4.4 quotes the dominant terms of the decomposition; Table 6 gives all four for every cell. It is an accounting identity over the labelled pairs, not a causal attribution: it says which pairs a score orders correctly, not why. Polarity ties every pair within a stratum, so its within-stratum entries are 0.5 by construction.

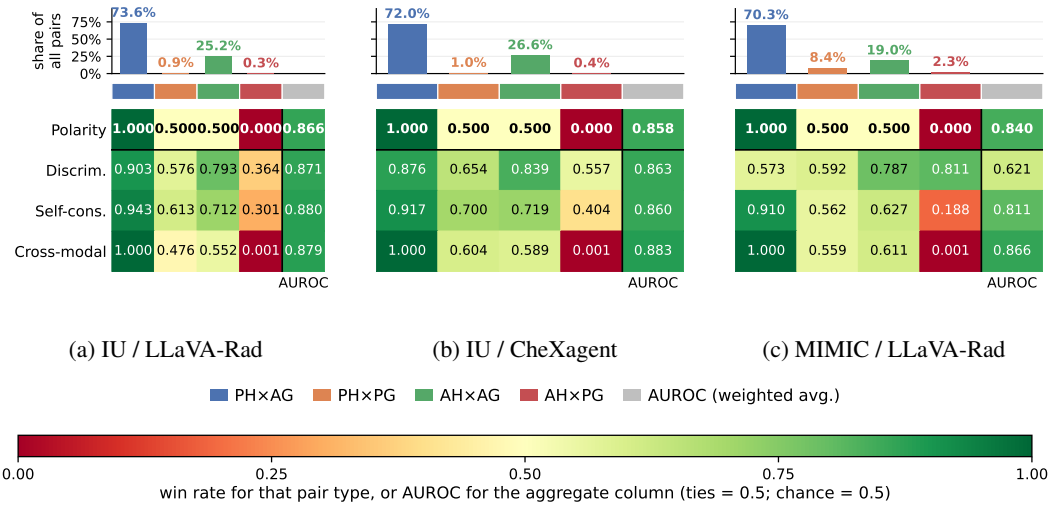
Figure 6 shows the same identity visually: the top row is the weight each pair type carries, the heatmap is polarity's and each family's win rate on that pair type, and the rightmost column is the resulting AUROC on the same colour scale. PH$\times$AG (blue) is most of the mass and is near-solved by every score except the discriminative classifier on MIMIC-CXR (0.573); the two absent-stratum types (green, red) are a fifth to a quarter of the mass and are where the families visibly disagree, most sharply for the discriminative classifier on MIMIC-CXR.

H WHAT THE LABEL RULE COMPARES, ON REAL STUDIES

Figures 7, 8 and 9 show the comparison the label rules actually make, on real images. Every study and image is picked by a stated rule rather than by hand. For each IU generator, the absent-stratum panel is the median study, among those holding at least one grounded and one hallucinated claim in that stratum and an image (26 studies for LLaVA-Rad, 47 for CheXagent), by the gap between its hallucinated and its grounded claims' discriminative percentile; the polarity panel is the study containing the asserted claim of Figure 5 for LLaVA-Rad, and for CheXagent the median of the 40 qualifying studies by the discriminative percentile of its asserted hallucinated claims. The COCO panels are the median image per generator by the gap between its grounded and hallucinated claims' detector probability. Each claim carries the finding the rule extracted from it and each panel the finding set extracted from the gold report, so what is compared is visible rather than implied. MIMIC-

Table 6: Every term of the pair decomposition of aggregate AUROC. A pair sets a hallucinated (H) claim against a grounded (G) one; P and A mark a claim that asserts or denies a finding. The weight row gives each pair type's share of all pairs; the other rows give the fraction of pairs of that type a score orders correctly (ties count one half), so each AUROC is the weighted sum of its row, which reproduces it exactly. Families are at their aggregate-selected configurations; "polarity, then score" orders claims by polarity and breaks ties by the family's score. AUROCs are recomputed from the stored per-claim scores; IU / LLaVA-Rad self-consistency prints 0.880 here and 0.879 in Table 1 because the recomputed and sweep values differ by 10^{-4} (0.87957 vs 0.87949).

	PH×AG	PH×PG	AH×AG	AH×PG	AUROC	Polarity, then score
IU / LLaVA-Rad						
weight	73.6%	0.9%	25.2%	0.3%		
polarity	1.000	0.500	0.500	0.000	0.866	–
Discriminative	0.903	0.576	0.793	0.364	0.871	0.941
Self-consistency	0.943	0.613	0.712	0.301	0.880	0.921
Cross-modal	1.000	0.476	0.552	0.001	0.879	0.879
IU / CheXagent						
weight	72.0%	1.0%	26.6%	0.4%		
polarity	1.000	0.500	0.500	0.000	0.858	–
Discriminative	0.876	0.654	0.839	0.557	0.863	0.950
Self-consistency	0.917	0.700	0.719	0.404	0.860	0.918
Cross-modal	1.000	0.604	0.589	0.001	0.883	0.883
MIMIC / LLaVA-Rad						
weight	70.3%	8.4%	19.0%	2.3%		
polarity	1.000	0.500	0.500	0.000	0.840	–
Discriminative	0.573	0.592	0.787	0.811	0.621	0.902
Self-consistency	0.910	0.562	0.627	0.188	0.811	0.869
Cross-modal	1.000	0.559	0.611	0.001	0.866	0.866



(a) IU / LLaVA-Rad (b) IU / CheXagent (c) MIMIC / LLaVA-Rad

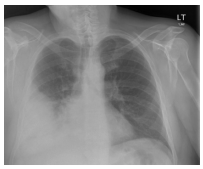
Figure 6: The pair decomposition of Table 6, visualised. Top: each pair type's share of all pairs. Heatmap: win rate on that pair type for polarity and each family (rows), and the resulting AUROC (rightmost column), one diverging colour scale (green = 1, red = 0, centred at 0.5).

CXR appears in Figure 9 as two studies shown for illustration; all other MIMIC-CXR evidence stays aggregate (Figure 2b, Table 5); no MIMIC-CXR record-level file is in the supplementary material.

Both median absent-stratum cases are laterality failures, not a sampling accident. In Figure 7a a claim that the *left* lung is clear is scored against an opacity the report places in the *right*

lower and middle lobes; in (c) the report contains the sentence *"Right lung remains clear."* and the claim says exactly that, yet the claim is labelled hallucinated because the report also describes a left upper lobe consolidation. The rule matches finding names and carries no laterality, so both are false positives. The selection rule surfaces them because it ranks by discriminative percentile and the classifier, being whole-image, fires hardest on precisely these claims. Figure 7b shows a second failure mode: the extractor has no entry for *granuloma*, so the gold parse is silent on *Nodule* even though the report states a calcified granuloma, so asserting a nodule is unsupported.

How much of the within-stratum result rests on that blind spot. The discriminative classifier shares the rule's laterality blindness, so its lead could in principle come from agreeing with the rule on claims both mislocalise. Claims naming a side are indeed concentrated among the hallucinated: $4/40$, $7/65$ and $29/110$ of the hallucinated absent-stratum claims in the three specialist cells name one, against 0.6%, 0.8% and 3.0% of the grounded ones. Dropping every claim that names a side nevertheless leaves the comparison intact: the discriminative AUROC within the stratum moves from 0.793, 0.839 and 0.787 to 0.791, 0.837 and 0.740, and its lead over cross-modal alignment from $+0.240$, $+0.249$ and $+0.177$ to $+0.239$, $+0.247$ and $+0.180$ — at most 0.004 in any cell. On MIMIC-CXR the levels fall together rather than apart. The rule's laterality blindness is real and costs it false positives; it is not what produces the within-stratum ordering.



(a) LLaVA-Rad, absent stratum, one study

`gold: Cardiomegaly: abs.; Fracture: abs.; Lung Opacity: pres.; Pneumothorax: abs.`

[HALLUCINATED] *denies*
"The left lung is clear."
`claim: Lung Opacity: abs.`
Discr. 97
Self-cons. 81
Cross-modal 71

[GROUNDED] *denies*
"There is no pneumothorax."
`claim: Pneumothorax: abs.`
Discr. 91
Self-cons. 44
Cross-modal 39

[GROUNDED] *denies*
"The mediastinal contours are normal."
`claim: Enlarged Cardiomediastinum: abs.`
Discr. 74
Self-cons. 37
Cross-modal 42

FINDINGS: "There is a large airspace opacity in the right lower and middle lobes. There is no pneumothorax. Heart size is normal. Soft tissue and bony structures unremarkable."



(b) LLaVA-Rad, polarity, one study

`gold: Consolidation: abs.; Effusion: abs.; Pneumothorax: abs.`

[HALLUCINATED] *asserts*
"A 6-mm nodule in the right upper lung is likely a calcified granuloma."
`claim: Nodule: pres.`
Discr. 88
Self-cons. 94
Cross-modal 100

[GROUNDED] *denies*
"The lungs are well expanded and clear without focal consolidation, pleural effusion or pneumothorax."
`claim: Consolidation: abs.; Effusion: abs.; Lung Opacity: abs.; Pneumothorax: abs.`
Discr. 61
Self-cons. 94
Cross-modal 51

[GROUNDED] *denies*
"Heart size is normal."
`claim: Cardiomegaly: abs.`
Discr. 14
Self-cons. 80
Cross-modal 19

FINDINGS: "The XXXX examination consists of frontal and lateral radiographs of the chest. The cardiac silhouette is not enlarged. There has been apparent interval increase in low density convexity at the left cardiophrenic XXXX. Calcified granuloma is again seen in the right upper lobe. There is no consolidation, pleural effusion or pneumothorax."



(c) CheXagent, absent stratum, one study

`gold: Cardiomegaly: abs.; Consolidation: pres.; Lung Opacity: pres.`

[HALLUCINATED] *denies*
"The right lung is clear."
`claim: Lung Opacity: abs.`
Discr. 92
Self-cons. 80
Cross-modal 84

[GROUNDED] *denies*
"The cardiomediastinal silhouette is normal."
`claim: Cardiomegaly: abs.; Enlarged Cardiomediastinum: abs.`
Discr. 67
Self-cons. 9
Cross-modal 64

[GROUNDED] *denies*
"There is no pleural effusion or pneumothorax."
`claim: Effusion: abs.; Pneumothorax: abs.`
Discr. 83
Self-cons. 32
Cross-modal 76

FINDINGS: "In the interval, consolidation has developed in the left upper lobe. Also, anterior segment XXXX opacity is present. Right lung remains clear. Heart size is normal."



(d) CheXagent, polarity, one study

`gold: Cardiomegaly: abs.; Effusion: abs.; Enlarged Cardiomediastinum: abs.; Lung Opacity: abs.`

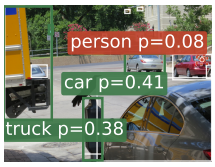
[HALLUCINATED] *asserts*
"A 1.5 cm nodule is seen in the left upper lobe."
`claim: Nodule: pres.`
Discr. 91
Self-cons. 100
Cross-modal 98

[GROUNDED] *denies*
"The lungs are otherwise clear."
`claim: Lung Opacity: abs.`
Discr. 39
Self-cons. 53
Cross-modal 38

[GROUNDED] *denies*
"The cardiomediastinal silhouette is normal."
`claim: Cardiomegaly: abs.; Enlarged Cardiomediastinum: abs.`
Discr. 10
Self-cons. 64
Cross-modal 23

FINDINGS: "Heart size and mediastinal contour are normal. Pulmonary vascularity is normal. Lungs are clear. No pleural effusions or pneumothoraces."

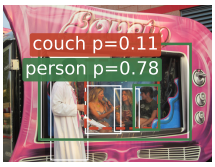
Figure 7: The label rule and the three families on real IU X-ray studies, for both generators (selection rules in Appendix H). Each claim shows its tag, the finding the rule extracted (`claim:`) and each family's score as a percentile among all claims of that cell (higher = scored as more likely hallucinated); each panel shows the gold report's extracted finding set (`gold:`) and its verbatim findings section. In (a) and (c) polarity is held fixed — every claim denies a finding — and the discriminative classifier separates the hallucinated one; in (b) and (d) the asserted claim is scored high by every family, cross-modal alignment included, which is the polarity term of Section 4.4. The hallucinated claims in (a) and (c) are both laterality failures of the rule, discussed above.



(a) COCO, LLaVA-1.5

"The image features a black crow perched on a parking meter on a city street. The crow is sitting on the parking meter, which is located near the center of the scene. ..."

[GROUNDED] car p=0.41
[GROUNDED] parking meter p=0.44
[HALLUCINATED] person p=0.08
[GROUNDED] traffic light p=0.50
[GROUNDED] truck p=0.38

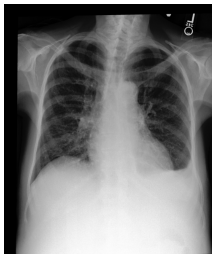


(b) COCO, Qwen2-VL

"The image depicts a lively scene featuring a pink, retro-style television set that appears to be a mobile installation or exhibit. The television set is prominently displayed, with a large, rectangular screen that occupies most of the image. ..."

[HALLUCINATED] couch p=0.11
[GROUNDED] person p=0.78
[GROUNDED] tv p=0.23

Figure 8: The same treatment on COCO, one image per generator. Solid green boxes are the open-vocabulary detector's best box for a grounded object, dashed red for a hallucinated one (ringed where the box is only a few pixels across), thin white the COCO ground truth. Here the comparison is settled on the image: the detector finds a box for the objects the caption gets right and not for the one it invents. Labels for three grounded objects are omitted where no collision-free slot exists; all claims are listed beside the image.



(a) Absent stratum, one study

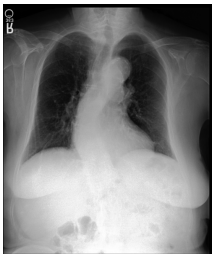
`gold: Cardiomegaly: abs.; Consolidation: abs.; Effusion: pres.; Enlarged Cardiomediastinum: abs.; Lung Opacity: abs.; Pneumonia: abs.; Pneumothorax: abs.`

[HALLUCINATED] *denies*
"There is no right pleural effusion."
`claim: Effusion: abs.`
Discr. 99
Self-cons. 61
Cross-modal 16

[GROUNDED] *denies*
"No definite focal consolidation is seen."
`claim: Consolidation: abs.`
Discr. 94
Self-cons. 21
Cross-modal 2

[GROUNDED] *denies*
"No pneumothorax is seen."
`claim: Pneumothorax: abs.`
Discr. 82
Self-cons. 31
Cross-modal 5

FINDINGS: "Interval removal of the left PICC line. The previously seen right lung pneumonia has resolved. No new consolidation. The hila and pulmonary vasculature are normal. The left pleural effusion is unchanged since ___. No pneumothorax. The cardiomediastinal silhouette is stable. Vertebrae appear diffusely sclerotic, consistent with history of myelodysplastic syndrome."



(b) Polarity, one study

`gold: Cardiomegaly: abs.; Consolidation: abs.; Effusion: abs.; Enlarged Cardiomediastinum: abs.; Lung Opacity: abs.`

[HALLUCINATED] *asserts*
"There is minimal left base atelectasis."
`claim: Atelectasis: pres.`
Discr. 68
Self-cons. 84
Cross-modal 97

[GROUNDED] *denies*
"No focal consolidation, pleural effusion, or evidence of pneumothorax is seen."
`claim: Effusion: abs.; Pneumothorax: abs.`
Discr. 31
Self-cons. 48
Cross-modal 33

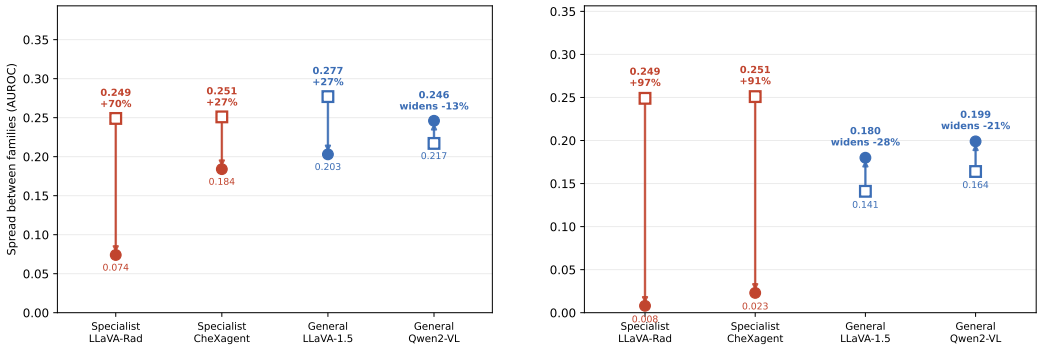
[GROUNDED] *denies*
"No overt pulmonary edema is seen."
`claim: Edema: abs.`
Discr. 55
Self-cons. 20
Cross-modal 14

FINDINGS: "The lungs are clear of consolidation, effusion, or vascular congestion. Cardiomediastinal silhouette is within normal limits. Tortuous thoracic aorta is noted with atherosclerotic calcifications at the arch. Mid to lower thoracic dextroscoliosis is again noted. Accentuated kyphosis is seen."

Figure 9: The same treatment on MIMIC-CXR / LLaVA-Rad, one study per panel, laid out as Figure 7. Studies are selected by the same median rules, restricted to the candidates whose gold parse the first reader of Appendix J.9, who also read all 244, confirmed on every slot the rule could touch (67 of 244). (a) Absent stratum: the discriminative classifier ranks the hallucinated denial near the top (99th percentile) and cross-modal alignment near the bottom (16th); the denial is, once more, a laterality failure of the rule, since the report describes a *left* effusion. (b) Polarity: the asserted finding (a basal atelectasis the report never mentions) is scored above both of the report's denials by every family, cross-modal alignment highest, which is the polarity term of Section 4.4.

I ADDITIONAL FIGURES

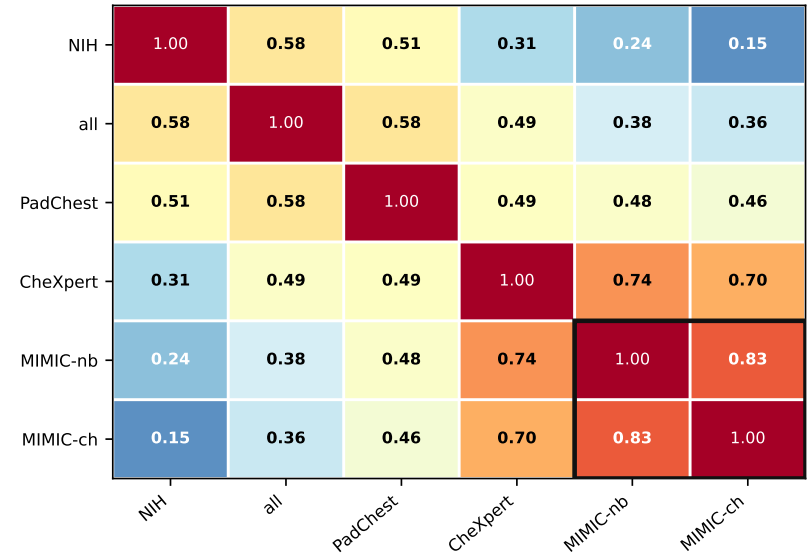
Figures 10 and 11 support results stated in the body: the first shows that the between-family spread behaves consistently across generators within each domain, the second that the discriminative family's wide interval tracks disagreement between the available pretrained checkpoints.



(a) All four families

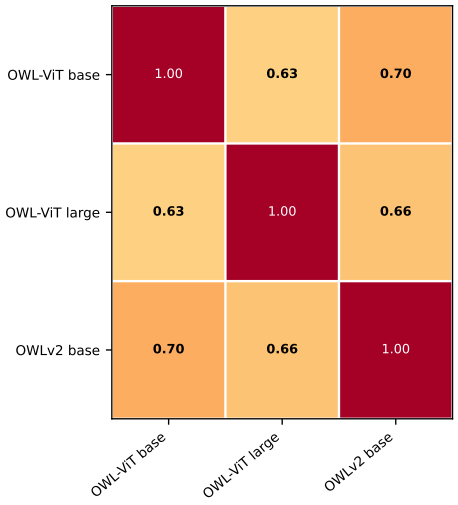
(b) Externally-grounded families only

Figure 10: Does compression follow the domain or the generator? Spread between families at reference (open) and best (filled) configurations, two generators per domain on IU X-ray and COCO. (a) Over all four families the specialist result is generator-dependent (70% vs. 27%). (b) Over the three externally-grounded families it is not: the spread collapses to $\leq$ 0.023 on both specialist generators and widens to $\geq$ 0.180 on both general generators.

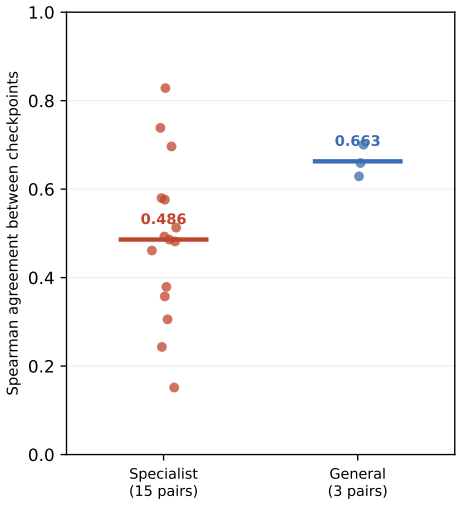


Boxed pair in (a): the two MIMIC-CXR checkpoints share a corpus and differ only in the labeler that read the reports (NegBio vs CheXpert). They agree at 0.83, the highest of the fifteen pairs, and every cross-corpus pair falls below them -- so what moves a checkpoint's scores is the training corpus, not the text-extraction rule. NIH ChestX-ray14, different in both, is the most distant at 0.15-0.31.

(a) Specialist: TXRV checkpoints, mean pairwise agreement 0.486



(b) General: open-vocabulary detectors, mean pairwise agreement 0.663



(c) Specialist checkpoints disagree, and disagree unevenly

Figure 11: Why checkpoint choice dominates in the specialist domain. (a) Specialist checkpoints agree only weakly, and the structure is set by the training corpus: the two MIMIC-CXR checkpoints, which differ only in the labeller that read the reports, are the most-agreeing pair of the fifteen, and every cross-corpus pair falls below them. (b) General-domain detectors agree uniformly. (c) Pairwise agreement, both domains.

J PROTOCOL DETAILS

J.1 THE SPECIALIST LABEL RULE

Both the generated claim and the reference findings section are parsed with RadGraph (Jain et al., 2021), and each parse is mapped to a fixed set of pathology names with a status, *present* or *absent*. An observation marked definitely present or definitely absent maps to its pathology; a normal descriptor ("clear", "normal") attached to an anatomical term marks that anatomy's characteristic pathology absent; if one claim yields both statuses for a pathology, *present* wins. A claim is then labelled as follows.

1. If the claim maps to no pathology, it is excluded (1,286 LLaVA-Rad sentences on IU, 3,399 on MIMIC-CXR).
2. For each pathology the claim names: if the reference states the opposite status, the claim is *hallucinated*; if the reference is silent and the claim asserts the pathology present, the claim is *hallucinated*; if the reference is silent and the claim denies it, the pathology is passed.
3. A claim with no hallucinated pathology is *grounded*.

Step 2 is where silence is read as absence: a denied finding the reference never mentions passes, an asserted one fails. The rule compares pathology names only. It carries no laterality ("the left lung is clear" against a right-sided consolidation is a contradiction) and inherits the extractor's vocabulary (a calcified granuloma is not a *Nodule*); Appendix H shows both failures and bounds the first.

J.2 INDEPENDENT LANGUAGE-MODEL CHECK

The 215 IU claims of Section 4.1 (LLaVA-Rad) were drawn by stratified sampling that over-represents claims on which the RadGraph labelling and a semantic-similarity labelling disagree, so the sample is not representative. Each was rated *hallucinated* or *grounded* by one of four independent instances of a strong language model (Claude), each instance receiving a disjoint quarter of the claims (54, 54, 54, 53) and, per claim, only the claim and the reference findings section: no signal scores and no rule labels. Every rating is binary, every claim received one, and each came with a one-line reason. An author read all 215 ratings and their reasons against the reference reports; no rating was changed. Of the 215 claims, 208 are in the final 2,222-claim label set (the other seven predate a re-run of the labelling pipeline); on those 208 the rule labels are unchanged and agreement is $189/208$ (90.9%). Against the RadGraph rule the ratings agree on 194 of 215 claims (90.2%, Cohen's $\kappa = 0.59$); the rule labels 13 claims hallucinated that the rater calls grounded and 8 the other way. Against the semantic labelling agreement is 34.9%. The rating instruction was issued inline and its exact wording was not archived; what is archived, and included in the supplementary material (`llm_check/`), is the protocol above, the 215 inputs, the four output files with their reasons, and the rule labels they are compared against. This is a strong-model adjudication read by an author, not radiologist ground truth, and it covers one generator on one corpus.

J.3 WHICH SIDE OF THE RULE COUPLES LABEL TO POLARITY

Table 7 gives the counts behind the 66% and 86% of Section 4.4. The reference reports are not terse about negatives: they state 3.56 explicit negatives on average and 86.2% state at least one, and two thirds of grounded denials deny a finding the reference explicitly negates. The coupling comes from the rule's other side: 352 of 407 hallucinated assertions name a finding the reference never mentions.

J.4 STATISTICAL PROCEDURES

Every AUROC is the rank (Mann–Whitney) estimator with tied scores credited one half, which is what makes the pair decomposition of Appendix G exact. Every interval is a 95% bootstrap percentile interval (2.5th to 97.5th percentile), with seed 0. The resampling unit follows the claim: a single AUROC resamples claims; headroom and every difference between families resample studies (chest X-rays) or images (COCO), paired, so that both terms of a difference are computed on the same resample and claims from one report move together. Each analysis script uses either 1,000 or 2,000 resamples and states its count; the COCO headroom, the pair decomposition, the MIMIC-CXR reversal checks and the third-party axes use 2,000. Selecting each family's best admitted

Table 7: MIMIC-CXR / LLaVA-Rad: what the reference says about each claim's findings (from script `r084`, over the stored per-claim parses).

Claims	Reference, for the claim's findings	n	share
Grounded denials (2,113)	explicitly negates all of them	1,396	66.1%
	explicitly negates some	188	8.9%
	silent on all of them	529	25.0%
Hallucinated assertions (407)	explicitly contradicts	55	13.5%
	silent	352	86.5%
Reference reports (1,259)	explicit negatives per report (mean)		3.56
	explicit positives per report (mean)		0.96
	at least one explicit negative		86.2%

configuration by aggregate AUROC and then reporting that AUROC is optimistic (Section 3.2); no interval corrects for it, and the direction of the bias can only enlarge the headroom we report.

J.5 Trivial-axis candidates on COCO

On COCO we tried every label-free surface property of a claim we could construct for object mentions, and report the maximum (Table 8). A label-based leave-one-image-out hallucination rate is shown for reference only; it consults labels and is not trivial. We did not try the response-position axis that is exploitable on M-HalDetect (Appendix B), because a COCO claim here is one object mention rather than a free-text span; a larger declared trivial set can only raise the axis and so lower the COCO headroom, which would narrow, not reverse, the domain contrast.

Table 8: Candidate trivial axes on COCO (AUROC, script `r049`); bold marks the trivial axis, the best label-free candidate.

Candidate	Kind	LLaVA-1.5	Qwen2-VL
COCO category prior frequency	label-free	0.480	0.525
Mention count in the caption	label-free	**0.736**	**0.731**
Leave-one-image-out hallucination rate	label-based	0.669	0.741
Best family (for reference)	—	0.907	0.908

J.6 Data sources and access

Table 9: Data used, and on what terms. No new data were collected.

Source	Used for	Access
IU X-ray (Demner-Fushman et al., 2016)	images, reports	public, de-identified (Open-i)
MIMIC-CXR (Johnson et al., 2019a)	images, reports	PhysioNet credentialed, DUA
COCO	images, object annotations	public
POPE (Li et al., 2023)	question files	MIT licence
AMBER (Wang et al., 2023)	generative annotations	Apache-2.0
M-HalDetect (Gunjal et al., 2024)	span annotations	CC BY-NC 4.0

No MIMIC-CXR record-level file is redistributed: the supplementary material contains aggregate results for MIMIC-CXR, and every withheld per-claim file is regenerated by its script, with the labelling module of Appendix J.1 (`label_rule/`), under a reader's own credentialed access.

J.7 Models, prompts and decoding

Cosine re-assertion rules embed sentences with the `all-MiniLM-L6-v2` sentence-transformer; entailment uses DeBERTa-large-MNLI (He et al., 2021); claims and references are parsed with

RadGraph (Jain et al., 2021). Generation prompts are "Provide a description of the findings in the radiology image." (LLaVA-Rad, CheXagent) and "Describe this image in detail." (LLaVA-1.5, Qwen2-VL). The text that is scored is the greedy output; the resamples that self-consistency consumes use temperature 0.7, `top_k=50`, `top_p=1.0`, with $K{=}10$ on chest X-rays and $K{=}5$ on COCO. No generation seed was fixed; the stored resamples are re-scored, not regenerated.

J.8 Order of decisions

One admissibility choice consulted labels of a sample other than the one it was applied to. COCO's resample count $K{=}5$ was set after the IU sweep had shown $K{=}5 \rightarrow 10$ worth about $+0.01$ AUROC there; it affects only self-consistency, which is not the best family on COCO, so the COCO headroom is unchanged by it. More generally, every configuration set was fixed before its own sweep was scored, but after earlier sweeps on other cells had been seen. Where such choices favoured what had scored well elsewhere they can only raise a family's best admitted value, the direction the upper-bound reading of Section 3.2 already allows for; COCO's $K{=}5$ went the other way, for a family that is neither best nor worst there, so neither the COCO headroom nor Δ^* moves.

J.9 How often the gold parse misreads the report

The specialist label is a function of the RadGraph parse of the reference report, not of its text, so a parse that inverts or misses a finding mislabels every claim that touches it. To measure this on MIMIC-CXR we drew 200 of the 1,259 labelled studies at random (seed 0) and had two independent readers (instances of Claude, blind to the parse and to each other) label each of the 16 pathologies in each reference as present, absent, uncertain or unmentioned, from the text alone (scripts `r088`, `r089`). The readers agree on 96.7% of the 3,200 slots ($\kappa = 0.94$); the 3,094 slots they agree on are compared with the parse. Where the readers call a finding present (288 slots) the parse marks it absent in 25 (8.7%) and is silent in 91 (31.6%); where they call it absent (768) the parse marks it present in 4 (0.5%). Of the 194 findings the parse marks present, 22 (11.3%) are not confirmed by the readers; of the 744 it marks absent, 25 (3.4%) are read as present. Lung Opacity accounts for 14 of the 25 inversions and 34 of the 91 silences, because the readers were told to count atelectasis, consolidation, pneumonia and nodules as opacities, unlike the word map.

Re-running the label rule on the sample's 485 claims with the readers' consensus overriding the parse changes 39 labels (8.0%); polarity's AUROC on the sample falls from 0.855 to 0.718, the discriminative family rises from 0.627 to 0.718 and cross-modal alignment falls from 0.872 to 0.770. Most of that movement is the Lung Opacity definition: overriding the other 15 pathologies only changes 16 labels (3.3%) and moves polarity to 0.819 and the best family to 0.848, while overriding Lung Opacity alone changes 23 and moves polarity to 0.759. Status errors of the parse are therefore a small lever on the headline; what a benchmark counts as an opacity is a larger one, and it is a labelling choice, not a parsing error. Neither reader is a radiologist, both are instances of one model given one instruction (so the κ above is an upper bound on reliability), and the audit covers one corpus. The first reader also labelled the 244 candidate studies of Figure 9; a candidate counts as confirmed when that reader finds no pathology present that the parse marks absent or omits, and none of the parse's present findings unconfirmed.

J.10 Configuration transfer between chest X-ray corpora

Table 10 backs the conclusion's transfer claim: each family's best IU configuration, applied unchanged to MIMIC-CXR, lands within 0.004 AUROC of the best configuration selected on MIMIC-CXR itself, while the discriminative level falls.

Table 10: LLaVA-Rad: each family's IU-selected best configuration scored on MIMIC-CXR, against the MIMIC-CXR-selected best (from the stored sweeps, scripts `r053` and `r055`).

Family	IU-best configuration	On IU	On MIMIC-CXR	MIMIC best
Discriminative (TXRV)	all / mean / logit / crop	0.871	0.617	0.621
Self-consistency	$K{=}10$, NLI entailment	0.879	0.811	0.811
Cross-modal (BiomedCLIP)	contrastive, t_0, $T{=}1$	0.879	0.866	0.866

J.11 COMPUTE

No model is trained; every generator, detector, encoder and entailment model is a public checkpoint used for inference only. The most expensive step, drawing $K{=}10$ resampled reports for each of 1,259 MIMIC-CXR studies with LLaVA-Rad, took 17.9 hours on a single workstation GPU (NVIDIA RTX 4000 Ada, 20 GB).

K ALL ADMITTED CONFIGURATIONS

Every configuration admitted under the admissibility criteria of Section 3.2, with bootstrap percentile intervals, for each (domain, generator) pair. Rows are sorted by AUROC within each family; the extremes of each block are the endpoints of the implementation interval reported in Table 4.

Table 11: Specialist domain, LLaVA-Rad (IU X-ray): all admitted configurations.

Configuration	AUROC	95% CI	n
Discriminative – TXRV			
TXRV / all / mean / logit / crop	0.871	[0.849, 0.891]	2222
TXRV / all / mean / prob / crop	0.867	[0.843, 0.888]	2222
TXRV / all / mean / logit / nocrop	0.866	[0.843, 0.887]	2222
TXRV / all / min / prob / crop	0.866	[0.842, 0.887]	2222
TXRV / all / min / logit / crop	0.866	[0.842, 0.887]	2222
TXRV / all / mean / prob / nocrop	0.863	[0.839, 0.884]	2222
TXRV / all / min / prob / nocrop	0.861	[0.835, 0.883]	2222
TXRV / all / min / logit / nocrop	0.861	[0.835, 0.883]	2222
TXRV / all / max / prob / crop	0.836	[0.811, 0.859]	2222
TXRV / all / max / logit / crop	0.836	[0.811, 0.859]	2222
TXRV / all / max / prob / nocrop	0.833	[0.808, 0.856]	2222
TXRV / all / max / logit / nocrop	0.833	[0.808, 0.856]	2222
TXRV / nih / max / prob / crop	0.770	[0.733, 0.802]	1546
TXRV / pc / max / prob / crop	0.712	[0.675, 0.747]	1687
TXRV / chex / max / prob / crop	0.650	[0.611, 0.685]	2205
TXRV / mimic_nb / max / prob / crop	0.525	[0.477, 0.568]	2205
TXRV / mimic_ch / max / prob / crop	0.514	[0.470, 0.555]	2205
Cross-modal – BiomedCLIP			
CLIP / contrastive t0 T=1	0.879	[0.841, 0.917]	2222
CLIP / contrastive t0 T=100	0.879	[0.841, 0.917]	2222
CLIP / contrastive t1 T=1	0.720	[0.684, 0.759]	2222
CLIP / contrastive t1 T=100	0.720	[0.684, 0.759]	2222
CLIP / contrastive t2 T=1	0.708	[0.669, 0.750]	2222
CLIP / contrastive t2 T=100	0.708	[0.669, 0.750]	2222
CLIP / single-text prompt 'a chest x-ray showing'	0.653	[0.605, 0.699]	2222
CLIP / raw-claim-sim (as published)	0.587	[0.541, 0.632]	2222
CLIP / single-text prompt 'findings:'	0.536	[0.484, 0.585]	2222
Self-consistency			
V3_K10_NLI_entail	0.879	[0.849, 0.909]	–
V2b_K10_continuous	0.723	[0.688, 0.758]	–
V1_K5_continuous	0.709	[0.674, 0.745]	–
V4_K5_thresh0.7	0.701	[0.659, 0.741]	–
V4_K5_thresh0.6	0.672	[0.631, 0.713]	–
V2a_K10_thresh0.55	0.656	[0.611, 0.699]	–
V0_K5_thresh0.55	0.645	[0.603, 0.689]	–
V4_K5_thresh0.55	0.645	[0.603, 0.689]	–
V4_K5_thresh0.5	0.614	[0.571, 0.657]	–
V4_K5_thresh0.4	0.541	[0.503, 0.583]	–
Token confidence			

Configuration	AUROC	95% CI	n
TokenConf / exact / min	0.805	[0.777, 0.832]	2176
TokenConf / exact / geomean	0.804	[0.776, 0.831]	2176
TokenConf / exact / mean	0.803	[0.774, 0.830]	2176
TokenConf / r010 / min	0.797	[0.768, 0.826]	2176
TokenConf / r010 / geomean	0.792	[0.763, 0.820]	2176
TokenConf / r010 / mean	0.789	[0.760, 0.818]	2176
TokenConf / exact / median	0.756	[0.714, 0.795]	2176
TokenConf / r010 / median	0.744	[0.704, 0.782]	2176
TokenConf / exact / first	0.717	[0.681, 0.751]	2176
TokenConf / exact / max	0.619	[0.572, 0.666]	2176
TokenConf / r010 / max	0.602	[0.555, 0.647]	2176
TokenConf / r010 / first	0.568	[0.525, 0.612]	2176

Table 12: Specialist domain, CheXagent (IU X-ray): all admitted configurations.

Configuration	AUROC	95% CI	n
Discriminative – TXRV			
TXRV / all / min / prob / nocrop	0.863	[0.846, 0.879]	2389
TXRV / all / min / logit / nocrop	0.863	[0.846, 0.879]	2389
TXRV / all / min / prob / crop	0.859	[0.841, 0.876]	2389
TXRV / all / min / logit / crop	0.859	[0.841, 0.876]	2389
TXRV / all / mean / logit / nocrop	0.856	[0.838, 0.873]	2389
TXRV / all / mean / logit / crop	0.853	[0.836, 0.870]	2389
TXRV / all / mean / prob / nocrop	0.852	[0.834, 0.870]	2389
TXRV / all / mean / prob / crop	0.850	[0.832, 0.867]	2389
TXRV / all / max / prob / crop	0.809	[0.789, 0.828]	2389
TXRV / all / max / logit / crop	0.809	[0.789, 0.828]	2389
TXRV / all / max / prob / nocrop	0.807	[0.786, 0.826]	2389
TXRV / all / max / logit / nocrop	0.807	[0.786, 0.826]	2389
TXRV / nih / max / prob / crop	0.800	[0.773, 0.825]	1560
TXRV / pc / max / prob / crop	0.725	[0.696, 0.754]	1561
TXRV / chex / max / prob / crop	0.683	[0.647, 0.720]	2331
TXRV / mimic_nb / max / prob / crop	0.532	[0.490, 0.576]	2331
TXRV / mimic_ch / max / prob / crop	0.411	[0.368, 0.452]	2331
Self-consistency			
V3_K10_NLI_entail	0.860	[0.834, 0.884]	2389
V2b_K10_continuous	0.665	[0.632, 0.698]	2389
V4_K5_thresh0.7	0.644	[0.609, 0.680]	2389
V1_K5_continuous	0.636	[0.605, 0.667]	2389
V2a_K10_thresh0.55	0.582	[0.543, 0.619]	2389
V4_K5_thresh0.6	0.572	[0.535, 0.611]	2389
V0_K5_thresh0.55	0.558	[0.522, 0.593]	2389
V4_K5_thresh0.55	0.558	[0.522, 0.593]	2389
V4_K5_thresh0.5	0.515	[0.479, 0.552]	2389
V4_K5_thresh0.4	0.513	[0.479, 0.549]	2389
Cross-modal – BiomedCLIP			
CLIP / contrastive t0 T=1	0.883	[0.854, 0.911]	2389
CLIP / contrastive t0 T=100	0.883	[0.854, 0.911]	2389
CLIP / contrastive t1 T=1	0.822	[0.793, 0.850]	2389
CLIP / contrastive t1 T=100	0.822	[0.793, 0.850]	2389
CLIP / contrastive t2 T=1	0.794	[0.766, 0.824]	2389
CLIP / contrastive t2 T=100	0.794	[0.766, 0.824]	2389
CLIP / raw-claim-sim (as published)	0.638	[0.604, 0.673]	2389

Configuration	AUROC	95% CI	n
CLIP / single-text prompt 'a chest x-ray showing'	0.604	[0.570, 0.638]	2389
CLIP / single-text prompt 'findings:'	0.603	[0.566, 0.639]	2389
Token confidence			
TokenConf / exact / min	0.699	[0.654, 0.742]	2389
TokenConf / r010 / min	0.699	[0.654, 0.742]	2389
TokenConf / exact / geomean	0.684	[0.640, 0.727]	2389
TokenConf / r010 / geomean	0.684	[0.640, 0.727]	2389
TokenConf / exact / median	0.667	[0.626, 0.706]	2389
TokenConf / r010 / median	0.667	[0.626, 0.706]	2389
TokenConf / exact / mean	0.665	[0.624, 0.705]	2389
TokenConf / r010 / mean	0.665	[0.624, 0.705]	2389
TokenConf / exact / first	0.627	[0.587, 0.666]	2389
TokenConf / r010 / first	0.627	[0.587, 0.666]	2389
TokenConf / exact / max	0.461	[0.421, 0.500]	2389
TokenConf / r010 / max	0.461	[0.421, 0.500]	2389

Table 13: General domain, LLaVA-1.5 (COCO val2017): all admitted configurations.

Configuration	AUROC	95% CI	n
Discriminative – open-vocabulary detector			
Detector / owlvit-large / prob	0.907	[0.895, 0.918]	2731
Detector / owlvit-large / logit	0.907	[0.895, 0.918]	2731
Detector / owlv2-base / prob	0.883	[0.869, 0.896]	2731
Detector / owlv2-base / logit	0.883	[0.869, 0.896]	2731
Detector / owlvit-base / prob	0.868	[0.854, 0.882]	2731
Detector / owlvit-base / logit	0.868	[0.854, 0.882]	2731
Detector / owlvit-large / binary@0.3	0.822	[0.805, 0.836]	2731
Detector / owlvit-base / binary@0.1	0.772	[0.754, 0.790]	2731
Detector / owlv2-base / binary@0.3	0.752	[0.737, 0.767]	2731
Detector / owlv2-base / binary@0.1	0.741	[0.721, 0.758]	2731
Detector / owlvit-base / binary@0.3	0.728	[0.714, 0.742]	2731
Detector / owlvit-large / binary@0.1	0.706	[0.687, 0.724]	2731
Self-consistency			
SelfCons / support+mentions	0.871	[0.856, 0.885]	2731
SelfCons / count/K (discrete, K=5)	0.853	[0.838, 0.869]	2731
SelfCons / binary thr$\geq$3	0.761	[0.743, 0.778]	2731
SelfCons / binary thr$\geq$2	0.683	[0.665, 0.701]	2731
SelfCons / binary thr$\geq$1	0.593	[0.580, 0.607]	2731
Cross-modal – CLIP			
CLIP / single-text / photo	0.727	[0.708, 0.747]	2731
CLIP / single-text / bare	0.717	[0.697, 0.736]	2731
CLIP / single-text / scene	0.713	[0.692, 0.733]	2731
CLIP / contrastive / photo+empty	0.695	[0.673, 0.717]	2731
CLIP / contrastive / scene+empty	0.683	[0.660, 0.705]	2731
CLIP / contrastive / photo+noneg	0.680	[0.658, 0.702]	2731
CLIP / contrastive / scene+noneg	0.678	[0.654, 0.700]	2731
CLIP / contrastive / bare+noneg	0.644	[0.621, 0.666]	2731
CLIP / contrastive / bare+empty	0.641	[0.618, 0.664]	2731
Token confidence			
TokenConf / sentence-exact / first	0.704	[0.683, 0.724]	2731
TokenConf / word / geomean	0.622	[0.599, 0.645]	2731
TokenConf / word / min	0.620	[0.596, 0.643]	2731

Configuration	AUROC	95% CI	n
TokenConf / word / mean	0.609	[0.585, 0.633]	2731
TokenConf / sentence-r010 / min	0.607	[0.585, 0.629]	2731
TokenConf / sentence-exact / min	0.607	[0.585, 0.629]	2731
TokenConf / sentence-r010 / geomean	0.607	[0.585, 0.630]	2731
TokenConf / sentence-exact / geomean	0.606	[0.585, 0.629]	2731
TokenConf / sentence-exact / mean	0.592	[0.569, 0.614]	2731
TokenConf / sentence-r010 / mean	0.591	[0.568, 0.614]	2731
TokenConf / word / median	0.588	[0.564, 0.611]	2731
TokenConf / sentence-r010 / first	0.585	[0.559, 0.609]	2731
TokenConf / sentence-exact / median	0.576	[0.552, 0.599]	2731
TokenConf / sentence-r010 / median	0.574	[0.552, 0.600]	2731
TokenConf / word / max	0.535	[0.512, 0.560]	2731
TokenConf / word / first	0.527	[0.502, 0.550]	2731
TokenConf / sentence-exact / max	0.498	[0.475, 0.521]	2731
TokenConf / sentence-r010 / max	0.495	[0.473, 0.517]	2731

Table 14: General domain, Qwen2-VL (COCO val2017): all admitted configurations.

Configuration	AUROC	95% CI	n
Discriminative – open-vocabulary detector			
Detector / owlvit-large / prob	0.908	[0.891, 0.923]	2003
Detector / owlvit-large / logit	0.908	[0.891, 0.923]	2003
Detector / owlv2-base / prob	0.885	[0.864, 0.905]	2003
Detector / owlv2-base / logit	0.885	[0.864, 0.905]	2003
Detector / owlvit-base / prob	0.873	[0.853, 0.893]	2003
Detector / owlvit-base / logit	0.873	[0.853, 0.893]	2003
Detector / owlvit-large / binary@0.3	0.819	[0.797, 0.840]	2003
Detector / owlv2-base / binary@0.3	0.768	[0.744, 0.788]	2003
Detector / owlvit-base / binary@0.1	0.767	[0.740, 0.793]	2003
Detector / owlvit-base / binary@0.3	0.743	[0.724, 0.760]	2003
Detector / owlv2-base / binary@0.1	0.741	[0.715, 0.766]	2003
Detector / owlvit-large / binary@0.1	0.737	[0.708, 0.764]	2003
Self-consistency			
SelfCons / support+mentions	0.800	[0.771, 0.828]	2003
SelfCons / count/K (discrete, K=5)	0.772	[0.744, 0.802]	2003
SelfCons / binary thr$\geq$3	0.695	[0.665, 0.724]	2003
SelfCons / binary thr$\geq$2	0.648	[0.621, 0.676]	2003
SelfCons / binary thr$\geq$1	0.593	[0.572, 0.616]	2003
Cross-modal – CLIP			
CLIP / single-text / photo	0.709	[0.680, 0.736]	2003
CLIP / single-text / scene	0.703	[0.674, 0.731]	2003
CLIP / single-text / bare	0.695	[0.665, 0.723]	2003
CLIP / contrastive / scene+noneg	0.683	[0.649, 0.714]	2003
CLIP / contrastive / photo+empty	0.681	[0.652, 0.711]	2003
CLIP / contrastive / scene+empty	0.677	[0.645, 0.710]	2003
CLIP / contrastive / photo+noneg	0.671	[0.639, 0.702]	2003
CLIP / contrastive / bare+noneg	0.639	[0.606, 0.671]	2003
CLIP / contrastive / bare+empty	0.631	[0.600, 0.664]	2003
Token confidence			
TokenConf / sentence-exact / geomean	0.662	[0.631, 0.693]	2003
TokenConf / sentence-r010 / geomean	0.661	[0.630, 0.692]	2003
TokenConf / sentence-r010 / mean	0.656	[0.625, 0.688]	2003
TokenConf / sentence-exact / mean	0.656	[0.625, 0.688]	2003

Configuration	AUROC	95% CI	n
TokenConf / sentence-r010 / median	0.635	[0.601, 0.667]	2003
TokenConf / sentence-exact / min	0.633	[0.601, 0.666]	2003
TokenConf / sentence-exact / median	0.631	[0.597, 0.662]	2003
TokenConf / word / geomean	0.630	[0.594, 0.661]	2003
TokenConf / sentence-r010 / min	0.630	[0.596, 0.663]	2003
TokenConf / word / mean	0.625	[0.589, 0.658]	2003
TokenConf / word / min	0.607	[0.576, 0.640]	2003
TokenConf / word / median	0.594	[0.557, 0.630]	2003
TokenConf / word / max	0.594	[0.558, 0.630]	2003
TokenConf / word / first	0.575	[0.541, 0.609]	2003
TokenConf / sentence-exact / first	0.547	[0.511, 0.585]	2003
TokenConf / sentence-r010 / first	0.528	[0.492, 0.564]	2003
TokenConf / sentence-r010 / max	0.493	[0.460, 0.527]	2003
TokenConf / sentence-exact / max	0.490	[0.458, 0.525]	2003