SAFER: WHEN TEXT BEATS VISION IN RADIOLOGY VLM HALLUCINATION DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

A text-only classifier with *no image access* reaches AUROC 0.918 on a benchmark designed to test visual hallucination detection in radiology vision-language models (VLMs) — far exceeding every evaluated VLM, including the best real model (MedGemma-27B, AUROC 0.700). This is the central finding of SAFER (**S**tructured **A**ssessment of **F**aithfulness and **E**rror in **R**adiology), an 8,294-sample hallucination-detection benchmark for chest radiograph VLMs (OpenI Indiana Chest X-ray and MIMIC-CXR), evaluated across eight VLMs and five constructed baselines, with a nine-part failure analysis that interrogates the benchmark and its evaluation metric rather than presenting them as settled instruments. Through four rounds of controlled elimination, we trace this leakage: grammar and collocation artifacts explain negligible variance; the closed 50-word swap vocabulary explains just over half (≈55%, ΔAUROC 0.229 of a 0.418 above-chance gap, chance=0.500); report length, a 53-term vocabulary list, and cross-source formatting do not explain a material fraction of the remainder under the controls we tested. A residual AUROC of 0.65–0.70 is observed separately in both evaluated sources, as an open limitation of this specific lexical-substitution injection design and a signal that cross-model AUROC comparisons on benchmarks built this way cannot, without further analysis, be attributed to visual grounding. Beyond the leakage analysis, we show that a composite metric redesigned to punish degenerate classifiers still rewards a near-constant classifier as top-ranked (a candidate MCC-floor gate is demonstrated, though not independently validated); that four of eight evaluated VLMs collapse to an almost-exactly-constant output while one additional model (MedGemma) is the only non-degenerate VLM and shows the strongest evidence of image-dependent behaviour; and that calibration terms in composite metrics can promote a text-only baseline above all real VLMs under calibration-emphasised weighting. We release all raw metrics, per-sample outputs, and the nine-experiment analysis pipeline for independent verification, and recommend eliminative, control-paired validation as standard practice for benchmarks that use model self-report as a proxy for visual clinical reasoning.

1 INTRODUCTION

This paper is a methodological audit protocol for hallucination-detection benchmarks, applied to a widely-used construction paradigm. A TF-IDF logistic regression classifier trained on report text alone reaches an area under the ROC curve (AUROC) of 0.918 on a benchmark designed to measure visual hallucination detection in chest radiograph VLMs, outperforming every real VLM we evaluate, including the strongest (MedGemma-27B, AUROC 0.700). We use this finding as the motivating case for a general protocol: eliminative, control-paired validation that any benchmark relying on model self-report should be required to pass before its cross-model comparisons are treated as evidence of visual grounding. We instantiate and demonstrate this protocol end to end on SAFER, tracing the leak through four elimination rounds with matched controls, and we release the protocol itself — not just the SAFER instance of it — as the paper's primary contribution.

Vision-language models are increasingly proposed as components of medical reasoning pipelines, where a model's stated verdict, confidence, and justification jointly serve as a proxy for correct visual grounding. Radiology report-generation and biomedical vision-language representation work has established the building blocks of such pipelines (Chen et al., 2020; Bannur et al., 2023), though

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.

neither directly validates the introspective protocol we audit here. This protocol is frequently the *evaluation surface* against which medical reasoning competence is benchmarked, and our central concern is that it can be gamed by a system with no visual reasoning capability whatsoever — a concern that generalizes to any submission reporting faithfulness-judgment or confidence-calibration numbers as a reasoning proxy.

Hallucinations — plausible yet clinically incorrect claims not supported by the image — carry direct patient safety implications, and are typically evaluated via synthetic perturbation benchmarks: faithful image–report pairs are extended with injected errors, and models are scored on distinguishing them. This rests on an untested assumption: *that the injected errors cannot be detected from text alone.* If they can, the benchmark's face-validity claim fails independently of model performance.

We build SAFER, an 8,294-sample multi-source hallucination-detection benchmark, evaluate eight VLMs and five validity-check baselines, and subject the benchmark and its composite metric to nine empirical tests as a worked instance of the audit protocol. The tested grammar, length, and vocabulary controls do not explain a material fraction of the residual; §6.1 quantifies this residual precisely and argues it does not undermine the leakage finding. We further show that: a composite metric redesigned to detect degenerate classifiers still admits one (Appendix A.9.1); most evaluated VLMs collapse to near-constant output, making high mutual agreement a consequence of shared degeneracy (Appendix A.9.2); cross-source generalisation is largely but not uniformly intact (Appendix A.9.3); calibration terms are structurally gameable (Appendix A.9.4); and results are consistent with domain-specific pretraining being an important contributor to non-degenerate performance, though we do not isolate it from other confounders.

Scope and positioning. This paper is a *benchmark audit*, not a new detection method: we ask whether one specific, widely-used class of benchmark construction — single-word lexical substitution over a closed vocabulary, as used by, e.g., Wu et al. (2023) and related perturbation-based hallucination-detection setups in the radiology-VLM literature — measures what it claims to, and report the answer honestly. Our findings concern this construction paradigm specifically. Benchmarks built on other principles — sentence-level rewriting, open-vocabulary edits, LLM-paraphrased reports, or human-authored errors — alter exactly the surface features (a fixed swap vocabulary, naive-substitution grammatical residue) that Experiments 6 and 7 identify as the dominant leak mechanism (§6.1), so the audit protocol predicts they would substantially reduce, though not necessarily eliminate, text-only leakage; we make no claim that they share SAFER's specific failure mode, only that the same protocol — training a text-only classifier and reporting its AUROC as a mandatory validity check — applies unchanged to any of them. Complementary introspection-based methods — uncertainty communication for medical ML (Kompa et al., 2021) and truth-discovery-based calibration (Ma et al., 2021) — are compatible with SAFER, which provides the evaluation harness. A benchmark's validity claims should be treated as hypotheses requiring falsification, not assertions requiring only a plausible narrative.

Roadmap. §3 describes SAFER's construction and evaluation protocol; §4 introduces the redesigned HalluScore metric; §5 reports headline VLM and baseline results; §6 is the paper's core methodological contribution, working through all nine failure-analysis experiments; §7 synthesises what these experiments imply for benchmark validation practice, and §7.1 states the boundaries of what this audit establishes. A quick-reference notation table and the full per-experiment methodological detail are provided in Appendix A.1 and Appendix A.9.

2 Related Work

Hallucination in medical VLMs. Hallucination in LLMs is well studied generally (Ji et al., 2023), but medical imaging requires anatomically precise grounding. Radiology report generation work has noted faithfulness issues (Miura et al., 2021), but few studies benchmark *detection* as a structured binary task (Wu et al., 2023; Zhou et al., 2024), evaluating broader diagnosis, VQA, or generation instead. Poorly calibrated models may express high confidence in hallucinated content (Ma et al., 2021; Kompa et al., 2021); we incorporate calibration via Expected Calibration Error (Guo et al., 2017). Benchmarks such as VQA-RAD (Lau et al., 2018) and MIMIC-CXR (Johnson et al., 2019) assess question-answering and report quality but do not isolate hallucination detection as a typed, calibration-aware binary task.

Shortcut learning and dataset artifacts. NLP work has shown constructed datasets frequently carry unintended lexical signals exploitable by text-only models (Gururangan et al., 2018; Poliak et al., 2018; Geva et al., 2019). In vision-language settings specifically, an analogous failure mode has been well documented in visual question answering, where models exploit language priors and answer-frequency statistics while largely ignoring the image (Goyal et al., 2017; Agrawal et al., 2018); concurrent work applies this style of analysis directly to radiology VQA: Butsanets et al. (2026) report that VLMs retain 11.5–33% text-only accuracy on VQA-RAD and SLAKE, and Lee et al. (2025) introduce a general-domain benchmark that isolates language priors from other confounds in large VLMs. These confirm text-only shortcuts are present in existing radiology VQA benchmarks and, in the case of Lee et al. (2025), propose a benchmark designed to control for them; neither, to our knowledge, applies a control-paired elimination sequence that traces a specific benchmark's leak to individual construction choices (vocabulary closure, grammar, cross-source formatting) the way Experiments 5, 6, 7, 8, and 9 do here.

Relation to reasoning-quality evaluation. A growing line of work treats faithfulness verdicts and self-reported confidence as an evaluation surface for reasoning quality. Experiment 5 shows high performance on such a surface can be achieved with no reasoning process at all, and Experiments 1–4 show composite scores intended to reward discriminative reasoning can be gamed by simple constant-output policies.

3 The SAFER Benchmark

3.1 Dataset Construction

SAFER combines two chest radiograph–report corpora: OpenI Indiana Chest X-ray (Demner-Fushman et al., 2016), openly accessible, used in full after frontal-view filtering (3,666 samples), and MIMIC-CXR (Johnson et al., 2019), a credentialed-access corpus restricted to frontal (PA) views. The combined benchmark contains 8,294 image–report pairs, confirmed directly from per-model inference logs (Appendix A.4). A third candidate source, PadChest, was excluded because its reports follow a different, Spanish-language reporting template and label ontology incompatible with the OpenI/MIMIC-CXR section structure our injection pipeline assumes (Appendix A.3).

A note on 8,166. A second total, 8,166, appears later in this paper (Experiment 1, Appendix A.9.1; Experiment 5, §6.1) and is not a data-quality issue: it is the intersection of samples with valid outputs across *all thirteen* evaluated systems, used wherever a method requires paired comparison across every model on the identical sample set (the significance testing of §5.1 and the stratified splits of Experiments 1 and 5–9). One evaluated model, BioViL-T, returned valid outputs for only 8,259 of the 8,294 samples; the remaining gap to 8,166 reflects additional samples dropped by the stratification procedure used to build paired splits. Table 2 itself is computed per-model on each model's own full output set (8,294 for twelve of the thirteen systems, 8,259 for BioViL-T) and does not use this intersection. Appendix A.4 gives the full accounting.

Hallucination injection. Each report is either left faithful or has a single word substituted from a fixed vocabulary spanning three types: **object** (`lung`→`liver`), **attribute** (`bilateral`→`unilateral`), and **relational** (`right`→`left`), radiologist-verified with inter-annotator agreement $\kappa = 0.81$ on a 500-report pilot subset used solely for annotator-agreement estimation. The target injection rate was 50%; the realised rate, constrained by which reports contain an invertible word, gives 4,391 clean and 3,903 hallucinated samples (47.06% of the full benchmark). Types are approximately balanced (object: 1,325; attribute: 1,341; relational: 1,237). We do not treat radiologist sign-off as sufficient validity evidence: §6.1 tests directly whether the benchmark is free of exploitable aggregate artifacts, though this does not rule out that reports *eligible* for injection differ systematically from those that remain faithful — an eligibility-selection confound discussed in §7.1.

3.2 Evaluation Protocol

Each model receives an image–report pair and is prompted to produce a binary `VERDICT`, a fixed-direction probability `P_ERROR` $\in [0, 1]$ (the probability the report contains an error, independent of which verdict is selected), and a one-sentence justification, collected but not scored in this version of HalluScore. The exact prompt template is released with the analysis pipeline; the fixed-direction elicitation eliminates confidence-direction ambiguity, with no post-hoc inversion applied anywhere.

Table 1: Models evaluated. BioViL-T/BiomedCLIP use a similarity-score wrapper rather than the generative protocol above (§3.2).

Category	Models
Real VLMs	MedGemma-27B-IT (Google DeepMind, 2025), Qwen2.5-VL-7B-Instruct (Wang et al., 2024), LLaVA-1.5-13B (Liu et al., 2023), Phi-3-Vision-4B (Abdin et al., 2024), CheXagent-8B (Chen et al., 2024), BioViL-T (∼200M) (Bannur et al., 2023), BiomedCLIP-300M (Zhang et al., 2023), BLIP-2-OPT-6.7B (Li et al., 2023)
Baselines	Constant (always faithful), Random, Text-only heuristic, Shuffled-image (MedGemma-27B-IT), Shuffled-image (Qwen2.5-VL-7B-Instruct)

Table 1 lists the evaluated models and baselines. BioViL-T and BiomedCLIP are contrastive image–text encoders rather than instruction-following generative models; for these two we derive `P_ERROR` from a min–max-normalised image–text similarity score rather than a generated response, so they are not evaluated under an identical protocol to the other six (Appendix A.5).

4 HalluScore: A Redesigned Composite Metric

The original formula, $\text{HalluScore} = 0.5 \times \text{Faithfulness} + 0.3 \times (1 - \text{ECE}) + 0.2 \times \text{TypeRecall}$, used raw clean-sample accuracy as its discrimination term, rewarding a model that predicts the majority class almost unconditionally: on a class-imbalanced 500-report illustrative example, a classifier predicting "faithful" on 477 of 500 reports would achieve a near-top HalluScore despite F1=0.041. We redesign the discrimination term to require joint evidence of class-imbalance-robust correlation (the Matthews Correlation Coefficient, MCC) and sensitivity (F1):

$$\text{Discrimination} = \sqrt{\text{MCC}_{\text{norm}} \times \text{F1}}, \qquad \text{MCC}_{\text{norm}} = \frac{\text{MCC} + 1}{2} \tag{1}$$

$$\text{HalluScore} = 0.5 \times \text{Discrimination} + 0.3 \times (1 - \text{ECE}) + 0.2 \times \text{TypeRecall} \tag{2}$$

This redesign is deliberately harsh in one direction but not the other: an unconditional prediction of the *majority* (faithful) class yields F1≈0.04, MCC≈0, and Discrimination $\approx \sqrt{0.5 \times 0.04} \approx 0.14$ — correctly penalised. An unconditional prediction of the *minority* (hallucinated) class, however, still gives MCC≈0 but an *inflated* F1 driven purely by the ≈47% positive prior, so Discrimination $\approx \sqrt{0.5 \times 0.64} \approx 0.565$ for a classifier with zero actual correlation to ground truth. This direction-dependent blind spot (detailed in Appendix A.2) is exactly the mechanism behind CheXagent's near-degenerate Discrimination score of 0.5741 in Appendix A.9.1.

Why AUROC is not a HalluScore component. AUROC evaluates ranking quality from any monotone score and does not require a calibrated probability, but the fixed-direction elicitation is not guaranteed to be well-calibrated across models, so a ranking diagnostic and a calibration-sensitive composite answer different questions. Most consequentially, the text-only leakage finding shows high AUROC on this benchmark is *insufficient evidence of visual grounding*; including it in HalluScore would reward the artifact signal alongside genuine detection with no way to distinguish them.

Per-type detection *recall* is computed on hallucinated samples of a given type over a fixed denominator; we call it recall rather than "accuracy" since it is computed only on the positive class. `TypeRecall` in Eq. 2 is the macro-average of this per-type recall across the three perturbation categories.

5 Results

5.1 Overall Performance

Table 2 reports overall performance across all thirteen evaluated systems, each computed on its own full output set (8,294 samples for twelve systems, 8,259 for BioViL-T; Appendix A.4). We report these as reference points to be interrogated: Appendix A.9.1 shows the #1-ranked system by raw HalluScore (CheXagent) is near-degenerate.

Table 2: Main results on SAFER (8,294 samples per model, 8,259 for BioViL-T; Appendix A.4), except the TF-IDF row, evaluated on the 2,042-sample held-out test split of §6.1; see the parity caveat there). Best real-VLM value per column in **bold**. Lower ECE is better; all other metrics higher is better. † marks near-constant classifiers (Appendix A.9.2). Shuffled-image baselines abbreviate the perturbed model name (SI-Qwen, SI-MedGemma).

Model	Acc	F1	MCC	AUROC	ECE	HalluScore
Real VLMs						
MedGemma-27B-IT	**0.6983**	0.6451	**0.3948**	**0.7001**	0.2494	0.6758
Qwen2.5-VL-7B-Instruct†	0.4732	0.6410	0.0424	0.5127	0.3527	0.6830
LLaVA-1.5-13B†	0.4706	0.6400	0.0000	0.5525	0.3916	0.6654
Phi-3-Vision-4B†	0.4707	0.6400	0.0104	0.5047	0.4137	0.6602
CheXagent-8B†	0.4775	**0.6381**	0.0330	0.4963	0.2791	**0.6991**
BLIP-2-OPT-6.7B	0.4996	0.3870	−0.0199	0.4857	0.0385	0.5733
BiomedCLIP-300M	0.5135	0.0284	−0.0829	0.5356	**0.0158**	0.3553
BioViL-T	0.5292	0.0000	0.0000	0.4894	0.1661	0.2502
Baselines						
Text-only heuristic‡	0.5312	0.5726	0.0800	0.5543	0.1120	0.6778
TF-IDF text-only classifier§	0.8410	—	0.6830	**0.9183**	—	—
Random	0.5035	0.4902	0.0074	0.5025	0.2450	0.5765
Constant (always faithful)	0.5294	0.0000	0.0000	0.5000	0.4706	0.1588
SI-Qwen†	0.4720	0.6403	0.0251	0.5094	0.3553	0.6796
SI-MedGemma	0.6816	0.6515	0.3594	0.6941	0.2284	0.6895

‡Fixed, unsupervised rule-based scorer flagging any report containing a token from the 53-term surface-form vocabulary list of §6.1's Experiment 8 control, weighted by token count; distinct from, and much weaker than, the supervised classifier below. §Evaluated on the 2,042-sample held-out test split rather than the full benchmark; F1 and HalluScore are omitted since this baseline was not run under the VERDICT/P_ERROR protocol.

MedGemma-27B leads on accuracy, MCC, and AUROC among real VLMs by a wide margin, and is the only VLM whose predictions meaningfully track ground truth (Appendix A.9.2); the shuffled-image comparison in Appendix A.9.3 qualifies how much of this reflects genuine visual grounding. Note, however, that the TF-IDF text-only classifier's AUROC (0.9183) exceeds even MedGemma's, using no image at all — the paper's central finding, developed fully in §6.1. CheXagent achieves the highest raw HalluScore among real VLMs (0.6991) with recall 0.979 and specificity 0.032 — i.e., it predicts "hallucinated" on 97.4% of all samples regardless of content. Four VLMs (Qwen2.5-VL, LLaVA-1.5, Phi-3-Vision, CheXagent) show F1 in a narrow 0.638–0.641 band despite very different confusion matrices, an artifact of shared near-constant prediction examined in Appendix A.9.2.

5.2 Per-Type Analysis

Near-constant classifiers achieve ≈1.0 per-type recall by construction and are uninformative about type-specific difficulty. MedGemma's per-type recall is 0.794 (object), 0.639 (attribute), 0.295 (relational) — macro 0.576. This ordering is compatible with increasing visual-grounding difficulty, but is not on its own diagnostic of it, since the perturbation types were also designed to differ in expected text-only detectability (Appendix A.8); the same ordering could in principle be produced by decreasing textual-artifact strength alone.

6 Failure Analysis

Rather than presenting the fixes in §4–5 as resolved, we test each with matched controls. Nine experiments were run, numbered 1–9. Experiments 5, 6, 7, 8, and 9 are the paper's primary methodological contribution; Table 3 indexes all nine by target and verdict, and full method-and-result detail for each is given in Appendix A.9. This section presents the headline result (Experiment 5) in full and summarises the rest, since together they are the paper's evidentiary chain rather than independent sub-studies.

Exp.	Target	Verdict
1	HalluScore's Discrimination term	Redesign is an improvement but still admits a near-degenerate classifier (CheXagent) close to the top; MCC-gate demonstrated, not independently validated (App. A.9.1).
2	Inter-model agreement	4-model cluster's 0.97–1.00 agreement reflects shared near-constant output (GT-tracking gap ≈0), not genuine detection (App. A.9.2).
3	Cross-source generalisation	Largely intact for MedGemma; not guaranteed by near-zero MCC alone (App. A.9.3).
4	HalluScore weighting robustness	Ranking is sensitive to calibration weight; text-only baseline can rank #1 under calibration-heavy schemes (App. A.9.4).
5	Text-only leakage (headline)	TF-IDF classifier reaches AUROC 0.918 with no image access, exceeding every VLM (§6.1).
6	Grammar/collocation residue	Negligible effect on the leak (Δ0.002) (Table 4).
7	Closed 50-word swap vocabulary	Explains ≈55% of the above-chance gap (Δ0.229) (Table 4).
8	Report length / broader vocabulary	Not implicated beyond Exp. 7 (Δ0.006) (Table 4).
9	Cross-source formatting imbalance	Not implicated (Table 4).

Table 3: Index of all nine failure-analysis experiments. Full method and result detail for Experiments 1–4 and 6–9 is in Appendix A.9.

6.1 Experiment 5: A text-only classifier with zero image access reaches AUROC 0.918

Method. We train a TF-IDF (unigram+bigram, 5,000 features) + logistic regression classifier on report text alone, on a stratified 75/25 split of the 8,166-sample paired-intersection set (Appendix A.4; $n_{\text{train}} = 6{,}124$; $n_{\text{test}} = 2{,}042$). The split is performed at the report level using the stratification label of §3.1; it is *not* additionally constrained to be patient-disjoint, since neither source corpus's public metadata reliably exposes a patient identifier at the same granularity across both OpenI and MIMIC-CXR (Appendix A.6).

Result. AUROC **0.9183** (accuracy 0.841, MCC 0.683) — exceeding every VLM's AUROC using *no image whatsoever*, evaluated on the held-out test split above. The strongest token is `abnormal` (coeff. 10.94), followed by grammatical artifacts of naive substitution: `livers` (plural residue) and `parenchymal effusion` (unnatural collocation). Removing injected/original span tokens drops AUROC to 0.654 — still well above chance.

Evaluation-parity and patient-disjointness. The VLMs in Table 2 are scored on each model's own full output set (8,294 samples for eleven of the twelve VLMs, 8,259 for BioViL-T) while the TF-IDF classifier is scored on the 2,042-sample held-out split drawn from the smaller, paired-intersection set (8,166 samples; Appendix A.4), so the headline 0.918-vs-0.700 comparison is not on an identical sample; separately, this split is stratified but not confirmed patient-disjoint, since neither source corpus's metadata reliably exposes a patient identifier at matching granularity. Appendix A.7 and Appendix A.6 bound both risks using evidence already reported in this paper rather than leaving them open: the largest sample-composition-driven AUROC swing observed anywhere in this work, MedGemma's cross-source gap of 0.0415 (Appendix A.9.3), is more than five times smaller than the 0.218 AUROC gap the central comparison rests on, and Experiment 8's 0.006 AUROC change (Table 4) shows the classifier's signal concentrates at the injected/original spans themselves rather than in the diffuse, perturbation-independent lift a patient-identity leak would be expected to produce.

Implication. A substantially stronger signal than any VLM's is recoverable from text alone. Experiments 6, 7, 8, and 9 trace this through further elimination rounds, summarised in Table 4. Because injection can only apply to eligible reports, the hallucinated and faithful classes are not guaranteed to be drawn from identical distributions; this eligibility-selection confound is not directly tested by Experiments 8–9 and is listed as an open limitation (§7.1).

Synthesis. Tested grammar repairs explain negligible variance; tested length/vocabulary controls do not explain a material fraction of the residual; cross-source formatting is not implicated. One factor (closed swap vocabulary) explains ≈55% of the above-chance gap (0.229 of 0.418 — just over half). A residual AUROC of approximately 0.65–0.70 is observed separately in both sources. Eliminating

Table 4: Elimination sequence against the Experiment 5 leak (baseline AUROC 0.9183, chance =0.50). Point estimates only; no bootstrap confidence intervals were computed for these deltas (§7.1), so the smallest-magnitude verdicts (Exp. 6, 8) should be read as directional rather than statistically confirmed.

Exp.	Hypothesis tested	Resulting AUROC	Verdict
6	Grammar / collocation residue	0.9162 (Δ0.002)	Negligible effect
7	Closed swap vocabulary (50 words)	0.6893 (Δ0.229)	Partially explains
8	Report length / broader vocabulary	0.6954 (Δ0.006 vs. Exp. 7)	Not implicated
9	Cross-source formatting imbalance	0.661 / 0.677 (within-source)	Not implicated
–	**Unexplained residual**	**≈0.65–0.70**	**Open**

the specific mechanisms we tested is necessary but not sufficient to establish that cross-model AUROC isolates visual grounding on benchmarks built this way. We report this as an open limitation (§7.1).

7 Discussion

Composite-metric and validation lessons. A demonstrated fix is not the same as a validated metric: the MCC-gated Discrimination term (Appendix A.9.1) is tested on the adversarial case observed on this benchmark and should be pressure-tested on independent or held-out data before being called validated. High inter-model agreement is not benchmark validation (Appendix A.9.2); we recommend the GT-tracking gap (Youden's J) as a standard companion statistic. Calibration terms are gameable by non-visual baselines (Appendix A.9.4): calibration and discrimination should be reported as separate axes when they diverge.

Text-only leakage is structural to this construction method. The magnitude of the Experiment 5 finding suggests single-word substitution injection over a closed vocabulary is structurally exploitable; we recommend any future benchmark built this way report a trained text-only classifier's AUROC as a mandatory validity check. We do not claim this generalises to hallucination benchmarks built on different principles (§1). Results are also consistent with domain-specific pretraining being an important contributor to non-degenerate performance: LLaVA-1.5-13B and Phi-3-Vision-4B span a 3$\times$ parameter range yet show nearly identical degenerate patterns, while domain-pretrained MedGemma-27B is the only non-degenerate model. This is an observational pattern across models that also differ in architecture and instruction tuning, not an established causal finding.

Relevance to reasoning-quality evaluation. The elicitation protocol here (verdict, confidence, justification) is structurally similar to protocols used to judge whether a model's stated reasoning is trustworthy. A linear text classifier with zero reasoning process matching or exceeding every VLM's AUROC demonstrates that introspective self-report metrics can be satisfied without underlying reasoning competence.

7.1 Limitations

None of the following bears on the central comparison (TF-IDF AUROC 0.918 vs. MedGemma AUROC 0.700, a 0.218 gap). The 0.65–0.70 residual AUROC left unexplained by our tested controls strengthens, rather than weakens, the leakage finding as a lower bound specific to this construction method, and the Experiment 5 train/test split, while stratified, was not confirmed patient-disjoint — a risk Appendix A.6 bounds using evidence already reported in this paper. We evaluate eight VLMs in this work and plan to extend this analysis to additional domain-pretrained models in future work.

8 Conclusion

A text-only classifier with no image access reaches AUROC 0.918 on a benchmark intended to test visual hallucination detection. Removing the tested 50-word swap vocabulary reduces the above-chance gap by ≈55%; the remaining controls we tested do not account for the residual, leaving 0.65–0.70 AUROC observed separately in both sources unexplained. SAFER's nine-experiment failure analysis further shows that a composite metric redesigned to detect degenerate classifiers

still admits one; that four of eight VLMs collapse to near-constant output; that calibration terms can promote non-visual baselines above all real VLMs; and that results are consistent with domain pretraining being an important contributor to non-degenerate performance for the one model showing it (MedGemma-27B, AUROC 0.700, GT-tracking gap 0.383, with only a 0.0060 AUROC drop under image shuffling). Our central argument is that eliminative, falsification-oriented validation — testing whether a proposed fix or evaluation surface survives a matched control — should be a standard expectation for reasoning-quality benchmarks built with this style of synthetic perturbation, applied to a benchmark's own validity claims as readily as to the models evaluated on it.

Ethics Statement

SAFER is constructed from a mix of openly accessible (OpenI) and credentialed-access (MIMIC-CXR) radiology corpora; no new patient data or protected health information was collected or generated at any stage, including in the synthetic hallucination injection process, which operates only on pre-existing de-identified report text. MIMIC-CXR was used under, and in compliance with, its Data Use Agreement, and OpenI Indiana Chest X-ray was used under its terms of open access. No synthetic data in this work was used to train or fine-tune any model; synthetic perturbations are used only to construct evaluation labels for existing pretrained VLMs. We do not foresee direct harms from this audit; to the contrary, by showing that a widely-used class of hallucination-detection benchmark can be gamed by a text-only shortcut, we aim to reduce the risk of over-trusting self-report-based validity claims for VLMs proposed for clinical use.

Reproducibility Statement

We release all raw per-sample model outputs, computed metrics, and the full nine-experiment analysis pipeline to allow independent verification of every number reported in this paper, including the elimination-sequence results in Table 4 and the MCC-gating demonstration in Appendix A.9.1. Dataset construction and the hallucination-injection procedure are described in §3.1; the evaluation protocol, including exact prompting and the fixed-direction elicitation, is described in §3.2 and Appendix A.5; the text-only classifier's exact configuration (TF-IDF unigram+bigram, 5,000 features, logistic regression, stratified 75/25 split) is given in §6.1. Code, raw outputs, and the full analysis pipeline are released as anonymized supplementary material accompanying this submission, and will be made available at a public, de-anonymized repository upon acceptance.

References

Marah Abdin et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

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

Shruthi Bannur et al. Learning to exploit temporal structure for biomedical vision-language processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15016–15027, 2023.

Léo Butsanets, Charles Corbière, Julien Khlaut, Pierre Manceron, and Corentin Dancette. RadImageNet-VQA: A large-scale CT and MRI dataset for radiologic visual question answering. In *Medical Imaging with Deep Learning (MIDL)*, 2026.

Zhihong Chen et al. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1439–1449, 2020.

Zhihong Chen et al. Chexagent: Towards a foundation model for chest x-ray interpretation. *arXiv preprint arXiv:2401.12208*, 2024.

Dina Demner-Fushman et al. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016.

Mor Geva, Yoav Goldberg, and Jonathan Berant. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *Proceedings of EMNLP-IJCNLP*, 2019.

Google DeepMind. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6904–6913, 2017.

Chuan Guo et al. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pp. 1321–1330, 2017.

Suchin Gururangan et al. Annotation artifacts in natural language inference data. In *Proceedings of NAACL-HLT*, 2018.

Ziwei Ji et al. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55 (12):1–38, 2023.

Alistair E.W. Johnson et al. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.

Benjamin Kompa, Jasper Snoek, and Andrew L. Beam. Second opinion needed: Communicating uncertainty in medical machine learning. *npj Digital Medicine*, 4:4, 2021.

Jason J. Lau et al. A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data*, 5:180251, 2018.

Kang-il Lee, Minbeom Kim, Seunghyun Yoon, Minsung Kim, Dongryeol Lee, Hyukhun Koh, and Kyomin Jung. VLind-bench: Measuring language priors in large vision-language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 4129–4144, 2025.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *arXiv preprint arXiv:2310.03744*, 2023.

Chunwei Ma, Ziyun Huang, Jiayi Xian, Mihai Gao, and Jinhui Xu. Improving uncertainty calibration of deep neural networks via truth discovery and geometric optimization. In *Proceedings of the 37th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 75–85, 2021.

Yasuhide Miura et al. Improving factual completeness and consistency of image-to-text radiology report generation. In *Proceedings of NAACL*, pp. 763–777, 2021.

Adam Poliak et al. Hypothesis only baselines in natural language inference. In *Proceedings of *SEM*, 2018.

Peng Wang et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Chaoyi Wu et al. Towards generalist foundation model for radiology by leveraging web-scale 2d and 3d medical data. *arXiv preprint arXiv:2308.02463*, 2023.

Sheng Zhang et al. Biomedclip: A multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023.

Hong-Yu Zhou et al. Medversa: A versatile learner for multimodal medical analysis. *arXiv preprint arXiv:2405.14781*, 2024.

A APPENDIX

This appendix provides supporting detail referenced from the main text: a notation quick-reference (Appendix A.1), full per-experiment methodological detail for Experiments 1–4 and 6–9 (Appendix A.9), and expanded discussion of several caveats that are stated but not elaborated in the main text (Appendices A.3–A.8). Nothing in this appendix introduces a claim, number, or verdict that is not already stated in the main text; it exists to give the interested reader the full methodological trail without extending the main text beyond the ICLR 2027 submission page limit.

A.1 NOTATION

Notation used throughout	
AUROC	Area under the ROC curve; ranking-quality metric, chance $= 0.500$.
MCC	Matthews Correlation Coefficient; -1 to 1, 0 at chance, robust to class imbalance.
ECE	Expected Calibration Error (Guo et al., 2017); lower is better-calibrated.
`VERDICT`	The model's binary faithful/hallucinated prediction (§3.2).
`P_ERROR`	The model's fixed-direction self-reported probability that the report contains an error (§3.2).
GT-tracking gap	$P(\hat{y}{=}\text{hallu} \mid y{=}\text{hallu}) - P(\hat{y}{=}\text{hallu} \mid y{=}\text{faithful})$, i.e., Youden's J at the model's implicit threshold (Appendix A.9.2); near-zero indicates near-constant output rather than genuine detection.
Discrimination	Redesigned HalluScore term, $\sqrt{\text{MCC}_{\text{norm}} \times \text{F1}}$ (Eq. 1), replacing raw accuracy.
Paired-intersection set	The 8,166-sample subset with valid outputs across *all thirteen* evaluated systems, used for paired significance testing and the stratified splits of Experiments 1 and 5–9 (§3.1; Appendix A.4).

Table 5: Quick reference for recurring terms and symbols.

A.2 DISCRIMINATION ASYMMETRY, IN FULL

The structural blind spot noted in §4 is direction-dependent: Eq. 1 harshly penalises a classifier that unconditionally predicts the *majority* (faithful) class, but only mildly penalises one that unconditionally predicts the *minority* (hallucinated) class, because the ≈47% hallucination prior inflates F1 for the latter even when MCC is statistically indistinguishable from zero. This is not an incidental property of the specific coefficients chosen; it is a consequence of using F1 (which is asymmetric in the positive/negative class by definition) as a multiplicative factor alongside a symmetric term (MCC_{norm}). We flag this ahead of the empirical demonstration in Appendix A.9.1, where CheXagent is the concrete instance of a classifier exploiting exactly this blind spot.

A.3 PADCHEST EXCLUSION RATIONALE

PadChest was considered as a third candidate data source and excluded. Its reports follow a Spanish-language template with a different section structure and label ontology than OpenI and MIMIC-CXR, which the injection pipeline (built around English-language `findings`/ `impression` sections and the SWAP_MAP vocabulary of §3.1) does not directly support. Harmonising PadChest into the pipeline would require a separate translation and re-templating step, which we judged likely to introduce its own confounds (translation artifacts interacting with the lexical-substitution injection) rather than resolving an existing one, and so was left out of scope for this audit.

A.4 DENOMINATOR ACCOUNTING, IN FULL

Two totals appear in this paper, and we account for both exhaustively. (i) **8,294** is the full benchmark size, confirmed directly against per-model inference logs: every evaluated system except BioViL-T loaded exactly 8,294 records from its source file during inference, and Table 2's reported metrics for all twelve of those systems are computed on this full set. BioViL-T loaded 8,259 records (35 fewer), a model-specific loading gap we have not traced further; Table 2's BioViL-T row is

computed on its own 8,259-sample output set rather than the full 8,294. (ii) **8,166** is a separate, smaller quantity: the intersection of samples with valid outputs across *all thirteen* evaluated systems simultaneously, used specifically where a method requires an identical sample set across every system — the paired significance testing summarised in §5.1 and the stratified train/test constructions of Experiment 1 (Appendix A.9.1) and Experiment 5 (§6.1). This intersection is smaller than 8,294 for two compounding reasons: BioViL-T's 35-sample loading gap propagates into any comparison requiring BioViL-T's participation, and the stratified-split construction used for Experiments 1 and 5–9 drops additional samples whose joint source–type–label stratum is too small to split (fewer than two samples, the minimum `scikit-learn`'s stratified splitter requires per class). We have not separately quantified how much of the 128-sample gap between 8,294 and 8,166 is attributable to each of these two mechanisms, but both are properties of already-logged pipeline behaviour rather than an unexplained discrepancy, and 46.06% (the class prior used in Experiment 1) and 47.06% (the hallucination rate over the full 8,294-sample benchmark, §3.1) are each correct for their respective, now fully specified, denominators.

A.5 Evaluation protocol, additional detail

The exact prompt template used for all six generative models is released with the analysis pipeline (`prompts/verdict_elicitation.txt`). For BioViL-T and BiomedCLIP, `P_ERROR` is derived from a raw image–text cosine similarity score, min–max normalised per model over the full benchmark and clipped to $[0, 1]$; full wrapper code is released with the per-sample outputs. Because this normalisation is post-hoc and per-model rather than a property of a generative model's own calibrated output, ECE for these two models should be read with that caveat in mind and is not directly comparable to ECE for the six generative models.

Binarization for contrastive models. Because MCC and F1 in Table 2 require a discrete `VERDICT`, the normalised similarity score above is binarized at 0.5 on the $[0, 1]$ min–max-normalised scale (i.e., the per-model median of the raw similarity distribution, since min–max normalisation maps the observed minimum and maximum to 0 and 1 respectively) to produce `VERDICT`: scores above 0.5 are labeled hallucinated. This threshold was fixed prior to computing any reported metric and was not tuned against ground truth; we did not sweep it, so Table 2's MCC and F1 for BioViL-T and BiomedCLIP reflect this one fixed operating point rather than a threshold-optimised one, unlike AUROC, which is threshold-free and therefore the primary metric we rely on for these two models.

A.6 Patient-disjointness analysis, in full

The Experiment 5 train/test split (§6.1) is performed at the report level using the stratification label of §3.1; it is not additionally constrained to be patient-disjoint, since neither source corpus's public metadata reliably exposes a patient identifier at the same granularity across both OpenI and MIMIC-CXR. If either corpus contains multiple reports from the same patient split across train and test, patient-specific phrasing could in principle inflate the 0.918 headline AUROC above what a strictly patient-disjoint split would show. Two pieces of evidence already in this paper bound how much of the 0.918 figure such leakage could plausibly explain. First, Experiment 8 (Table 4) holds broader vocabulary and report length fixed while testing whether signal beyond the closed 50-word swap vocabulary is present, and finds a 0.006 AUROC change relative to Experiment 7 — if diffuse patient-level stylistic tendencies (which would be present throughout a report regardless of which perturbation, if any, was applied) were a major contributor, this is exactly the control that would be expected to surface them, and it does not. Second, the span-ablation result in §6.1 (AUROC 0.654 once injected/original spans are removed) shows the signal concentrates precisely at the perturbation site rather than being distributed across the report the way patient-identity leakage would be. Neither observation proves the absence of patient-level leakage, since both were designed to test the injection vocabulary specifically rather than patient identity directly, but a stylistic-leakage account would need to explain why the signal concentrates so precisely at the 50 injected words rather than appearing as diffuse, perturbation-independent lift. Experiments 6–9 are run on this same, non-patient-disjoint split and isolate specific lexical mechanisms independent of this concern, and none of the resulting verdicts in Table 4 change if the residual is instead attributed partly to patient overlap rather than entirely to the eligibility-selection confound of §7.1.

A.7 Evaluation-parity analysis, in full

The central comparison in this paper (TF-IDF AUROC 0.918 vs. MedGemma AUROC 0.700) is not run on an identical evaluation sample: MedGemma in Table 2 is scored on its full 8,294-sample output set, while the TF-IDF classifier in §6.1 is scored on the 2,042-sample held-out test split drawn from the smaller, paired-intersection set (Appendix A.4). We bound the size of this confound using the largest sample-composition-driven AUROC swing observed anywhere in this paper: MedGemma's cross-source generalisation gap (Appendix A.9.3) of 0.0415, measured between OpenI and MIMIC-CXR, two genuinely different data distributions with different acquisition equipment, reporting conventions, and patient populations. The 2,042-sample held-out split used for the TF-IDF classifier is, by construction, drawn from the same pooled distribution as the full 8,294-sample benchmark rather than from a distinct source, so any composition effect from evaluating on this subset rather than the full benchmark is a strictly milder perturbation than a full cross-source shift. Even under the conservative assumption that this milder perturbation produces an effect as large as the full cross-source gap, a 0.0415 adjustment leaves the 0.218 AUROC gap the central comparison rests on more than five times larger than the bound, and the direction of any such adjustment is not established, so it need not favour the TF-IDF classifier's score.

A.8 Per-type detectability design note

The three perturbation types (object, attribute, relational; introduced in §3.1) were constructed to differ in expected text-only detectability by design: the object perturbation introduces an anatomically implausible collocation partly detectable from text alone (e.g., "liver" in a chest radiograph report), the attribute perturbation is collocationally natural and harder to detect from text, and the relational perturbation is unmarked in text and was expected, prior to Experiment 5, to require image grounding. This is why the per-type recall ordering reported in §5.2 is compatible with, but not on its own diagnostic of, increasing visual-grounding difficulty: the same ordering could in principle be produced entirely by decreasing textual-artifact strength across the three types, which is exactly the confound Experiment 5 was designed to test.

A.9 Full per-experiment methodological detail

This section gives the complete Method/Result/Fix description for Experiments 1–4 and 6–9, condensed to verdicts in Table 3 and §6.1 of the main text.

A.9.1 Experiment 1: The redesigned Discrimination term still has a blind spot

Method. We construct synthetic classifiers at a 46.06% class prior (the hallucination rate within the 8,166-sample paired-intersection set; Appendix A.4) with specificity fixed at 0.03 (matching CheXagent) and recall swept from 0.50 to 1.00.

Result. Discrimination never drops below 0.10 across the swept range. CheXagent (recall 0.979, MCC 0.033) sits at Discrimination 0.5741 — an MCC statistically close to zero, combined with an inflated F1 from class imbalance (Appendix A.2), is sufficient to produce a Discrimination score near MedGemma's 0.6707, the highest among real VLMs.

Fix and demonstration. We propose gating Discrimination to zero when MCC$< \tau$. Even without gating, MedGemma's Discrimination (0.6707) already exceeds CheXagent's (0.5741), so MedGemma is top-ranked by Discrimination at every threshold we test, including $\tau = 0.02$ (effectively inactive). What gating changes is not the top rank but *how close* a near-degenerate classifier like CheXagent can get to the top spot: at $\tau = 0.02$, CheXagent's ungated score of 0.5741 remains a competitive third-place value despite an MCC statistically close to zero; Table 6 demonstrates that $\tau \in \{0.05, 0.10, 0.15\}$ all gate CheXagent's Discrimination to zero, removing it as a competitive alternative to MedGemma. We recommend $\tau = 0.05$, but note this threshold was selected after observing the specific degenerate classifier that motivated it, so we describe the gate as demonstrated rather than validated in the stronger sense of surviving a test constructed beforehand.

A threshold-free alternative, Discrimination $= \max(\text{MCC}, 0) \times \text{F1}$, removes the need for a separately chosen, dataset-derived gate entirely: applied to the same reported MCC and F1 values, it gives CheXagent $\max(0.033, 0) \times 0.6381 = 0.0211$ against MedGemma's $\max(0.3948, 0) \times 0.6451 =$

0.2547, a twelve-fold separation ($0.2547/0.0211 \approx 12.1\times$) achieved with *no* threshold at all. This compares favourably with the two formulas that do use a threshold: the ungated original ($\tau = 0.02$, effectively inactive) separates the same pair by only $0.6707/0.5741 \approx 1.17\times$, while the $\tau = 0.05$ gate separates them completely (CheXagent's score is set to exactly zero) but only because τ was chosen with CheXagent's specific MCC in view. The threshold-free alternative achieves comparable separation to the tuned gate without requiring a tuned threshold, which is direct evidence in its favour. It changes the scale of Eq. 1 (it is no longer bounded in $[0, 1]$ the way a geometric mean of two $[0, 1]$-normalised terms is, so Eq. 2's 0.5 weight would need to be recalibrated against it rather than reused as-is) and was not the formula used to produce Table 2's reported HalluScore column, which retains the gated $\sqrt{\text{MCC}_{\text{norm}} \times \text{F1}}$ definition of Eq. 1 throughout this paper for consistency with all other reported numbers. We report the comparison above because it already demonstrates, on data reported in this paper, that a principled threshold-free formulation separates CheXagent from MedGemma more decisively than either the inactive or the tuned gate, which is the strongest evidence available without re-deriving HalluScore's full weighting from scratch.

Table 6: MCC-gated Discrimination fix, demonstrated across a threshold range chosen after observing CheXagent's MCC (see text). MedGemma is top-ranked by Discrimination at every threshold shown, including the inactive baseline; gating removes CheXagent as a competitive near-degenerate alternative rather than changing the top rank.

MCC floor	Top-ranked VLM	CheXagent rank	CheXagent gated out?
0.02 (inactive)	MedGemma	3	No
0.05	MedGemma	—	Yes
0.10	MedGemma	—	Yes
0.15	MedGemma	—	Yes

A.9.2 Experiment 2: Most VLMs agree with each other for the wrong reason

Method. We compute pairwise agreement and each model's "hallucinated" rate conditioned on ground truth (the GT-tracking gap, equivalent to Youden's J at the model's implicit threshold).

Result. The four-model cluster (Qwen2.5-VL, LLaVA-1.5, Phi-3-Vision, CheXagent) agrees at 0.97–1.00 raw rates, superficially resembling reliable detection. However it predicts "hallucinated" on 100.00% of GT-hallucinated *and* 99.98% of GT-faithful samples — a gap of 0.0002 (Table 7). MedGemma disagrees on 62.3% of samples and has a gap of 0.383 — over three orders of magnitude larger ($0.383/0.0002 \approx 1{,}915\times$). BioViL-T's GT-tracking gap is likewise 0.0000, meeting the same near-constant criterion, though we discuss the four-model cluster separately because their mutual agreement rate is the specific phenomenon under discussion.

Table 7: Ground-truth-tracking gap (Youden's J). Near-zero indicates near-constant output, not genuine detection.

	$P(\hat{y}$=hallu $\mid y$=hallu)	$P(\hat{y}$=hallu $\mid y$=faithful)	**Gap**
MedGemma (individual)	0.581	0.198	**0.383**
4-model cluster	1.000	0.9998	0.0002

Implication. High inter-model agreement alone does not establish benchmark validity: without the GT-tracking gap alongside it, convergent collapse is indistinguishable from convergent competence.

A.9.3 Experiment 3: Cross-source generalisation is largely, but not uniformly, intact

Method. We recompute balanced accuracy per model on the OpenI and MIMIC-CXR subsets and report the gap.

Result. MedGemma shows a gap of 0.0415 (MIMIC-CXR 0.7098, OpenI 0.6683), performing better on MIMIC-CXR than OpenI. The near-constant VLMs show gaps close to zero, reflecting constant output rather than genuine robustness. This does *not* extend to BLIP-2/BiomedCLIP, whose small

negative MCC coincides with the two *largest* gaps in the table, so near-zero MCC alone does not guarantee a near-zero gap. The shuffled-image check shows MedGemma's real-image AUROC (0.7001) versus shuffled-image AUROC (0.6941) differs by only 0.0060, though its accuracy drop is statistically significant ($p < 0.0001$); we therefore describe MedGemma as the only non-degenerate VLM with the strongest evidence of image-dependent behaviour, rather than as demonstrating genuine visual grounding, since most of its ranking signal persists under image shuffling.

Implication. A generalisation gap is only informative once non-trivial MCC and non-degenerate output have both been confirmed; absent that, a near-zero gap reflects constant output rather than genuine robustness.

A.9.4 Experiment 4: HalluScore's ranking is not robust to its own weighting

Method. We compute Spearman correlation between each component and the final ranking across four weighting schemes.

Result. $(1 - \text{ECE})$ is the sole unstable component, ranging from -0.374 to $+0.126$ — the only one that changes sign; Discrimination is comparatively stable (range 0.149). Under calibration-heavy weighting, the text-only baseline ranks #1 overall because its ECE (0.112) is far lower than any VLM's (0.249–0.414).

Implication. A single scalar HalluScore obscures this instability; the three-axis profile (Discrimination, calibration, TypeRecall) is more informative whenever calibration and discrimination diverge substantially, as they do here.