ISSN 2278-2540 | DOI: 10.51583EMAS | Volume XV, Issue X, October2026

Medinsight AI: Explainable Chest X-Ray Analysis and Automated Radiology Report Generation with Knowledge Verification

M. Naga Mounika[1], S. Induja[2], G. Naga Sritha[3], C. Karthik Kumar[4], Khasim Syed[5*]

School of Computer Science and Engineering (SCOPE), Vellore Institute of Technology, Andhra Pradesh (VIT-AP University), Amaravati, Andhra Pradesh, India-522241

[*]Corresponding Author

DOI: https://doi.org/10.51583/IJLTEMAS.2026.151000062

Received: 15 October 2026; Accepted: 21 October 2026; Published: 06 November 2026

ABSTRACT

Small vision-language models (VLMs) can draft chest X-ray reports, but after fine-tuning they often fall back on one generic "normal" report, and they give no hint about when a draft should not be trusted. We studied a two-model pipeline on the Indiana University (IU) X-ray collection: a calibrated EfficientNetV2-S classifier and a MedGemma-4B generator fine-tuned with QLoRA. We asked whether the classifier's calibrated image-level predictions can send unreliable drafts to a human reader, and what training-free visual contrastive decoding (VCD) does to finding-level quality and diversity. On 240 held-out studies, VCD raised NIH-14 macro-F1 from 0.262 to 0.399 (paired bootstrap $p = 0.002$; Holm-adjusted $p = 0.012$) and cut the share of the single most common report from 71.2% to 35.8%. BLEU-4, ROUGE-L and CheXbert micro-F1 stayed flat or dropped slightly, no other clinical gain survived multiplicity correction, and because the greedy baseline used classifier hints in its prompt while VCD did not, the gains cannot be attributed to VCD alone. For verification, the classifier was adapted to IU with five-fold cross-fitting and every threshold was frozen before testing. At its 90%-sensitivity operating point the gate auto-approved 29.2% of VCD reports (70 of 240). Every approved report was a normal-reading draft, so the gate works as a screen for omissions and missed abnormalities in such drafts. Among approved reports, 4.3% (3 of 70; exact 95% CI 0.9–12.0%) had a labeler-detected label mismatch against the reference report, and 71 of the 74 label-mismatched reports were routed to review. On 1,033 test studies with retrieval-generated reports, the same gate approved 23.8% with a 6.9% mismatch rate (4.1–10.8%). A pre-declared second referee trained on CheXpert and MIMIC-CXR transferred poorly to IU (abnormality AUC 0.74) and halved coverage without a significant drop in risk. On 3,000 images from an external hospital system (VinDr-CXR, Vietnam) with radiologist labels, the frozen classifier reached a normal-versus-abnormal AUC of 0.891, IU-fitted calibration cut the mean expected calibration error from 0.30 to 0.037, and at the same operating point 28.7% of images would pass the gate, of which 3.6% (exact CI 2.5–5.1%) were abnormal by radiologist consensus. The report-level results are label-mismatch rates from a single institution with report-derived labels and no radiologist review of the generated reports; they are not radiologist-confirmed error rates and say nothing about clinical safety.

Index Terms—chest X-ray, radiology report generation, vision-language models, visual contrastive decoding, calibration, selective prediction

INTRODUCTION

Vision-language models are now being adapted to draft radi-ology reports, and domain models such as MedGemma report results on chest X-ray report generation [1]. Two kinds of mistake matter most. A draft can state a finding the image does not show, or it can leave out an abnormality that is there. Class imbalance makes both worse. Most studies are normal, so a model trained with cross-entropy can lower its loss just by writing a generic normal report, and earlier work found that report generators may do no better than unconditioned baselines [2].

We built a pipeline in which an image classifier acts as an independent referee for a report generator. Our first

ISSN 2278-2540 | DOI: 10.51583EMAS | Volume XV, Issue X, October2026

de-sign asked the two models to agree exactly on a set of find-ings. That did not work, and Section V-C explains why. We replaced it with a gate that uses classifier probabilities calibrated on the target domain and judged it as selective pre-diction, by coverage and risk [4]. Separately, we tried visual contrastive decoding [3], an inference-time method, to see whether it could reduce template collapse without retraining.

Research question. Can calibrated image-level abnor-mality predictions from an independent classifier pick out generated reports that should go to human review, and how does training-free VCD change finding-level quality and re-port diversity? The data answer the second part directly. The first part is supported only in a narrower form (Section III-D):the gate screens normal-reading reports for omissions and missed abnormalities.

Contributions. (1) A calibrated verification gate with a cross-fitted classifier and thresholds frozen before testing, evaluated by coverage and risk on 240 held-out studies and on 1,033 studies that include 793 never seen by the report generator, checked externally on 3,000 VinDr-CXR images with radiologist labels, and accompanied by an account of why exact agreement failed. (2) A paired-bootstrap evalua-tion of a VCD variant that contrasts against a no-image pass, applied to MedGemma on IU, with mixed results. (3) A pre-declared test of an independent second referee trained on other hospitals' data, which did not lower risk, and an analysis of the failure modes behind it: miscalibration under domain shift, template collapse that validation loss did not reveal, and errors that both models share. We do not claim state-of-the-art results, because our split, targets and metric code differ from the published IU protocols.

RELATED WORK

Classification and generation. ChestX-ray14, the 14-class extension of the ChestX-ray8 release [5], gives findings mined from reports, so its labels are weak, and Efficient-NetV2 [6] is a common convolutional backbone for this kind of multi-label task. Report generators range from memory-driven transformers such as R2Gen [7] to LLM-based systems. PromptMRG feeds classifier-derived diagno-sis prompts to the generator [8]. We fine-tune MedGemma-4B [1] with LoRA [9] on a 4-bit base (QLoRA) [10]. Babar et al. showed that unconditioned models can match encoder-decoder report generators on IU [2], which is why we pay attention to template collapse and diversity. TorchXRayVi-sion provides classifiers trained on large public datasets such as CheXpert and MIMIC-CXR [18–20], which we use as an outside referee.

Decoding-time hallucination control. VCD compares next-token scores computed from the original image and from a distorted copy, so that tokens driven by the language prior are suppressed [3]. Our variant contrasts against the same prompt with no image at all and adds a plausibility cutoff (Section III-C). It is not a reproduction of [3].

Report verification. Swin-Qwen3 uses LLM agents to draft, verify and refine a report; its verifier is a prompt that checks consistency, omissions and unsupported claims [11]. A phrase-grounded fact-checking method checks generated findings against the image and reports error detection on several generators [12]. Selective classification supplies the risk–coverage framework [4], and Platt scaling [14] and cali-bration analysis [15] supply the probability mapping.

Gap. In the papers we looked at, we did not find an evaluation that uses a separately calibrated, in-domain multi-label classifier as an outside referee for a small fine-tuned VLM, reports coverage and risk against report-derived labels, and tests a decoding intervention on the same reports. That is a narrow statement based on a limited search, not a novelty claim; related ideas exist [4, 11, 12].

Relation to our work. Reference [2] is a caution, not a method. It says generators should be judged by more than n-gram overlap. Reference [3] targets object hallucination in general-domain VLMs, so whether it helps radiology report content is an empirical question. PromptMRG [8] uses clas-**IU adaptation with cross-fitting.** Positive weighting and the change of hospital both inflate the raw probabilities (Sec-tion V-E), so we adapted Model 1 to IU with five-fold cross-fitting on the 2,160 development studies (Section IV-A). In each fold we started from the NIH weights and fine-tuned on the report-derived labels of the other four folds at 384 px for

a fixed six epochs (AdamW, learning rate 10^{-4} with a one-cycle schedule, batch size 16, positive weight

ISSN 2278-2540 | DOI: 10.51583EMAS | Volume XV, Issue X, October2026

capped at 10,

light geometric and intensity augmentation), then predicted the held-out fold with horizontal-flip test-time augmentation (TTA). Every development study therefore has an out-of-fold prediction, and test predictions are the average of the five fold models. Because the number of epochs was fixed in advance, no selection touches the out-of-fold predictions.

For each finding we keep the NIH head or the IU-tuned head, whichever has the higher out-of-fold AUC, and Platt scaling [14] fitted on the out-of-fold logits maps it to a cal-ibrated probability q. Calibration is a monotone map and cannot change AUC, so any AUC change comes from fine-tuning, resolution, TTA and head selection. A logistic re-gression over the 14 calibrated logits gives an abnormality score. Its five-fold cross-validated scores on the development studies set thresholds at 90%, 95% and 97.5% sensitivity.

Report Generation and VCD

Model 2 is MedGemma-4B-it [1] with a 4-bit NF4 base and LoRA adapters (rank 16, 29.8 M trainable parameters, 0.69%) on the attention and MLP projections of the language model;

the vision tower is frozen. It is trained to write "Findings: . . . Impression: . . . " (learning rate 2×10^{-4}, effective batch size 8, fp16, one epoch of about 4.8 hours on a T4 GPU). The adapter we evaluated was trained with an optional prompt line listing Model 1 predictions above 0.30.

At each decoding step, VCD replaces the token scores withsifier output as an input signal; we tested classifier hints in the prompt (no benefit in our ablation) and use the classifier for verification instead. The LLM verifier in [11] may shares(y) = (1 + α) log p(y | I, x)— α log p(y | x),(1)failure modes with the generator and has no image-grounded probability to calibrate. The checker in [12] is evaluated under its own protocol. We therefore make no numerical comparison with either.

PROPOSED METHOD

System Overview

Fig. 1 shows the pipeline. The same frontal image goes to Model 1 (the classifier) and Model 2 (the report generator). A labeler turns the draft report into 14 binary findings. A rule-based gate then compares those findings with Model 1's calibrated probabilities and outputs either AUTO-APPROVE or HUMAN REVIEW, with reason codes.

Chest X-ray Classification

Model 1 is tf_efficientnetv2_s.in21k_ft_in1k (20.2 M parameters). We fine-tuned it on NIH ChestX-ray14 at 224 × 224 for 14 labels, using binary cross-entropy with positive weights up to 653 (Hernia), AdamW (learning rate 3×10^{-4}, one-cycle schedule) and mixed precision for 15 epochs. We kept the checkpoint with the best validation mean AUC (0.8345, epoch 5).

where I is the image, x the prompt, and the second term is computed with the image removed. We keep only tokens whose image-conditioned probability is at least β times the maximum. We use α = 1, β = 0.1 and greedy search, with no classifier hints, and chose this setting from five configura-tions on 60 validation images. The cached contrast pass was checked against an uncached recomputation and agreed to

10^{-6}. VCD needs no retraining but roughly doubles decoding time (8.6 to 16.5 s per report on a T4). Among four training runs, we kept the adapter with the best overlap scores, notthe one with the lowest validation loss, as the baseline.

Calibrated Report Verification and Selective Review **Labeler.** Findings are extracted from text with CheXbert [13] for seven classes (atelectasis, cardiomegaly, consolidation,edema, effusion, pneumonia, pneumothorax)

ISSN 2278-2540 | DOI: 10.51583EMAS | Volume XV, Issue X, October2026

and with a

negation-aware, clause-scoped rule labeler for the other seven. Uncertain mentions count as positive. The same labeler pro-duces the reference labels from the radiologists' reports.

Rules. A class is a trusted referee if its out-of-fold AUC is at least 0.80 with at least 8 positives among the develop-ment studies; 11 of the 14 classes qualify (all except nodule, pneumonia and pneumothorax). For a trusted class c:

- UNSUPPORTED fires if the report states c and q(c) is below the 5th percentile of q among true positives;
- POSSIBLEOMISSION fires if the report is silent on c and

q(c) exceeds the F1-optimal threshold;

- POSSIBLEMISSEDABNORMALITY fires if the report states no finding but the abnormality score exceeds the threshold of the chosen operating point;
- ABNORMALREPORTSIGNOFF fires for any report that states a finding.

A report is auto-approved only if no rule fires. All thresholds are fitted on development studies only.

Consequence. Because any stated finding triggers sign-off, every auto-approved report states no finding. So in practice the gate only screens normal-reading reports for omissions and missed abnormalities. The UNSUPPORTED rule changes reason codes but not the routing decision in this configura-tion, and we did not ablate the rules. Coverage is the share of reports that are auto-approved. Risk is the share of approved reports whose 14-label vector differs from the reference la-bels, that is, the labeler-detected mismatch rate.

Independent Second Referee

Model 1 and Model 2 learned from the same IU pool and labels, so they can share blind spots. To test whether a model trained elsewhere could catch such errors, we added two TorchXRayVision DenseNet-121 classifiers [18], trained on CheXpert [19] and on MIMIC-CXR [20] and never on NIH or IU. Their 22 outputs (with flip TTA) feed a second abnor-mality regression, fitted with the same cross-validation on the development studies. In the dual gate a normal-reading report is approved only if both abnormality scores are below their thresholds at the same sensitivity; the fused gate uses one regression over both feature sets.

Pre-declared Evaluation

Before any test study was scored, all thresholds, head choices and calibration parameters were written to a file and hashed (SHA-256 prefix 6f7150d278e77a8a). The three operating points (90%, 95%, 97.5% sensitivity) and the primary com-parison, the dual gate at 90% sensitivity against the single-referee gate at the same point, were fixed in advance. A second, independent run reproduced the same hash and the same gate results.

Explainability

Grad-CAM [16] heatmaps from Model 1's head, for a se-lected finding, are shown in a local Gradio demonstration interface. We use them only as a qualitative visual aid. They mark regions that raised a class score; they are not verified diagnostic evidence, they do not show how the model reasons, and we did not measure whether they overlap annotated ab-normalities, so we make no localization claim. "Explainable" in this paper therefore means that every routing decision comes with explicit reason codes and calibrated probabilities (Section III-D), not that the heatmaps were validated. In one emphysema example the strongest activation lay partly outside the lung fields, which is why a heatmap still needs an expert to read it.EXPERIMENTAL SETUP

ISSN 2278-2540 | DOI: 10.51583EMAS | Volume XV, Issue X, October2026

Datasets and Splits

NIH ChestX-ray14 [5] (112,120 frontal images, 30,805 pa-tients, 14 text-mined labels) trains Model 1 with the official split: 73,916 / 12,608 / 25,596 images (23,806 / 4,202 / 2,797

patients), and the code asserts that no patient appears in two parts. IU X-ray [17] has 7,470 images and 3,955 reports, and 3,193 studies have a frontal image plus non-empty Find-ings and Impression. Model 2 used 2,400 randomly sampled studies, split at study level (seed 42) into 1,920 train, 240 validation and 240 test; 793 studies were never sampled.

Model 1's IU adaptation and all gate thresholds use the 2,160 Model 2 training and validation studies (the develop-ment pool). Model 1 is tested on the remaining 1,033 studies: the 240 Model 2 test studies plus the 793 never-sampled ones. The 240 test studies were never used for training, model se-lection, calibration or thresholds, and the Stage 2 split was recomputed and asserted equal to the stored test file. IU files carry no patient identifier, so patient-level separation cannot be confirmed. The 60-image VCD selection used validation images only.

External dataset. VinDr-CXR [21] contains frontal chest X-rays from two hospitals in Vietnam; in its public training set each image was labelled independently by three radiol-ogists. We used a random sample of 3,000 of these images (seed 42), counted a finding as present when a majority of the image's readers marked it, and mapped nine VinDr find-ings to the NIH classes (atelectasis, cardiomegaly, pleural effusion, infiltration, nodule/mass, pneumothorax, consolida-tion, pulmonary fibrosis, pleural thickening). VinDr has no reports, and nothing was fitted on it: the classifier, calibration and thresholds are the frozen IU ones, confirmed by the same hash. In the sample, 30.5% of images had at least one finding of any kind and 25.7% had one of the nine shared findings.

All IU labels are mined from report text: 34.4% of the 3,193 studies have at least one of the 14 findings, and so do 73 of the 240 test studies (30.4%). They are not radiologist annotations of the images. Model 1 and Model 2 both draw on the same development pool and the same report-derived labels.

Implementation and Metrics

Everything ran on a single NVIDIA T4. Clinical metrics compare labeler outputs for generated and reference reports: NIH-14 micro- and macro-F1; CheXbert micro-F1-14 (in-cluding "No Finding"), macro-F1-14 and micro-F1-5 [13]; and abnormal recall, the share of reference-abnormal studies whose report states any finding (not necessarily the right one). Macro-F1 averages over classes with at least one reference positive. We also report BLEU, ROUGE-L and METEOR on Findings+Impression using Hugging Face evaluate im-plementations; our METEOR cannot be compared with ME-TEOR 1.5 values in the literature. RadGraph F1 failed be-cause of a package incompatibility and is not reported. Di-versity is the share of unique reports and the share held by the most common one. For Table II, confidence intervals use 1,000 bootstrap resamples of test studies, and p-values are two-sided paired bootstrap tests against greedy decoding; we report Holm-adjusted values across the six clinical metrics. For the gates we report 2,000-resample bootstrap intervals

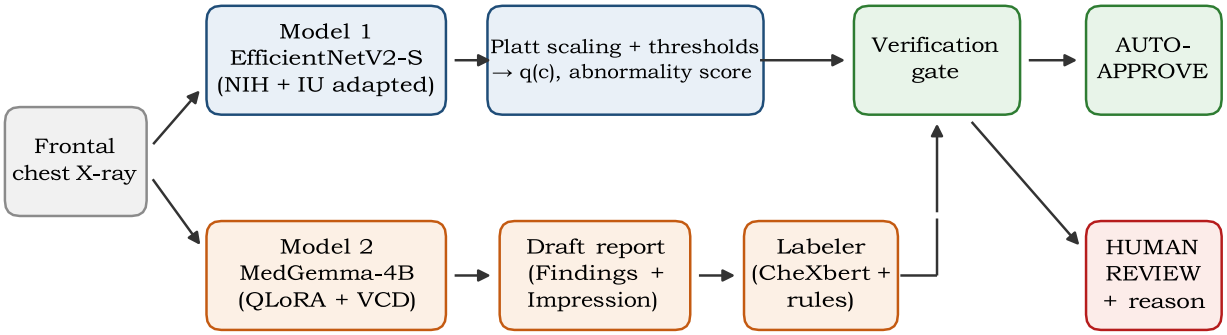


Figure 1. Pipeline. Model 1 yields calibrated probabilities q; Model 2 yields a draft report that the labeler converts to 14 binary findings; the gate routes each case.

ISSN 2278-2540 | DOI: 10.51583EMAS | Volume XV, Issue X, October2026

Table I. CONFIGURATIONS COMPARED IN THIS PAPER.

Configuration	What differs
Greedy	Model 2, greedy decoding, prompt includes Model 1 hints (as in training)
VCD	Model 2, contrastive decoding ($\alpha = 1$, $\beta = 0.1$), no hints
Retrieval	Report of the nearest development study (no generation)
Plan A, raw 0.30	Exact agreement on 8 NIH-trusted findings, raw NIH probabilities
Plan A, test-fit	As above, thresholds fitted on test images (leaky, for illustration)
Plan C	Calibrated gate, Model 1 only; 90/95/97.5% sensitivity points
Plan C dual	Plan C plus CheXpert/MIMIC referee; both must pass
Plan C fused	Plan C with one abnormality score from both models
External screen	Plan C on VinDr images, assuming a normal-reading draft table

and exact (Clopper–Pearson) intervals for risk. As a reference point we also ran a retrieval baseline that returns the report of the nearest development study in NIH Model 1 feature space. Calibration is measured by the expected calibration error (ECE, 15 equal-width bins) and the Brier score, per finding, for raw NIH probabilities and for the IU-calibrated ones.

External screen. On VinDr there are no draft reports, so we evaluate the part of the gate that does not depend on the draft. Because the gate only approves normal-reading drafts, an image would be approved exactly when no omission flag fires and the abnormality score is below threshold; we call this passing the screen. The share of passing images that radiologists call abnormal is the risk the gate would carry for a generator that always writes "normal", a worst case for the report side. Table I summarizes all configurations compared in this paper.

Comparability. Published IU results usually follow the R2Gen protocol [7]: a 7:1:2 split with about 590 test studies, Findings-only targets and Java-based scorers. We use a ran-dom study-level split, Findings+Impression targets and Hug-ging Face scorers. Our BLEU-4 (0.162–0.176) and ROUGE-L (0.366–0.381) fall in a similar range to common values, but they cannot be ranked against them.

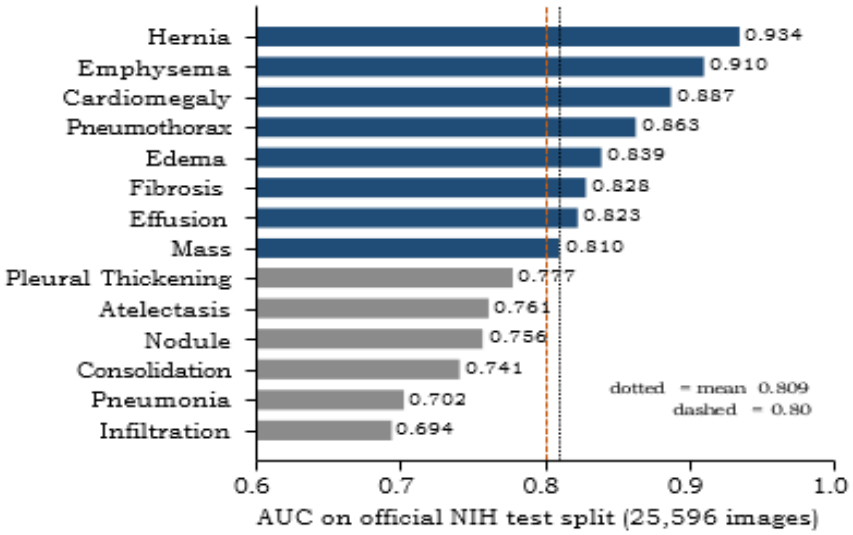


Figure 2. Model 1 AUC per finding on the official NIH test split. Blue bars reach 0.80 (dashed line); the dotted line is the mean.

Reproducibility. All seeds are 42, cuDNN runs in deter-ministic mode, the Stage 2 split is regenerated and

ISSN 2278-2540 | DOI: 10.51583EMAS | Volume XV, Issue X, October2026

asserted in the evaluation notebook, and every gate threshold is frozen from development studies before testing (Section III-F).

RESULTS AND DISCUSSION

Classification

On the official NIH test split, Model 1 reaches a mean AUC of 0.809, from 0.694 (infiltration) to 0.934 (hernia) (Fig. 2). On the 1,033 IU test studies, where all 14 findings have at least 3 positives, mean AUC rises from 0.837 for the NIH heads to 0.864 for the selected cross-fitted heads, and the normal-versus-abnormal AUC is 0.838. On the 240 Model 2 test studies, among the 10 findings with at least 3 positives, the corresponding values are 0.852 and 0.879, and the normal-versus-abnormal AUC is 0.874. Five of these ten findings have only 3–7 positives, so per-finding AUCs on the 240 studies are indicative only. All IU AUCs are measured against report-derived labels and should not be pooled with the NIH values.

VCD Comparison

Because the greedy arm used classifier hints and the VCD arm did not, every difference in this subsection reflects VCD and the removal of hints together. With that in mind, VCD raised NIH-14 macro-F1 and abnormal recall (Table II,**Table II.** REPORT QUALITY ON THE 240 IU TEST STUDIES. GREEDY VS. VCD WITH 95% BOOTSTRAP CIS, PAIRED BOOTSTRAP p AND

HOLM-ADJUSTED p ACROSS THE SIX CLINICAL METRICS. RETRIEVAL: NEAREST DEVELOPMENT REPORT, SAME 240 STUDIES. *HF evaluate

METEOR.

Metric	Greedy (95% CI)	VCD (95% CI)	p	Holm p	Retrieval
NIH-14 micro-F1	0.373 (0.281–0.453)	0.447 (0.358–0.525)	0.062	0.208	0.249
NIH-14 macro-F1	0.262 (0.177–0.349)	0.399 (0.262–0.505)	0.002	0.012	0.180
Abnormal recall	0.438 (0.324–0.548)	0.562 (0.446–0.667)	0.020	0.100	0.575
Abnormal precision	0.889 (0.771–0.974)	0.746 (0.625–0.864)	–	–	0.512
CheXbert micro-F1-14	0.579 (0.518–0.638)	0.576 (0.523–0.630)	0.912	0.912	0.399
CheXbert macro-F1-14	0.324 (0.242–0.388)	0.386 (0.286–0.455)	0.052	0.208	0.229
CheXbert micro-F1-5	0.388 (0.253–0.504)	0.480 (0.374–0.574)	0.074	0.208	0.281
BLEU-1 / BLEU-4	0.440 / 0.176	0.433 / 0.162	–	–	0.429 / 0.139
ROUGE-L	0.381	0.366	–	–	0.298
METEOR*	0.453	0.451	–	–	0.400
Unique reports (%)	27.5	40.4	–	–	81.7
Most common report (%)	71.2	35.8	–	–	4.2

The first design (Plan A) required the two models to produce identical sets of eight “high-trust” findings. It failed for four reasons. (i) Its keyword matcher found 262 positives in the reference reports, against 97 once negation was handled properly. (ii) The classifier’s raw-0.30 flags matched the radiologist reference exactly in only 4 of 240 cases. (iii) Trust tiers were taken from NIH AUC, which left out classes that transferred well to IU. (iv) Its recalibrated thresholds were fitted on the test images. On greedy reports it auto-approved

ISSN 2278-2540 | DOI: 10.51583EMAS | Volume XV, Issue X, October2026

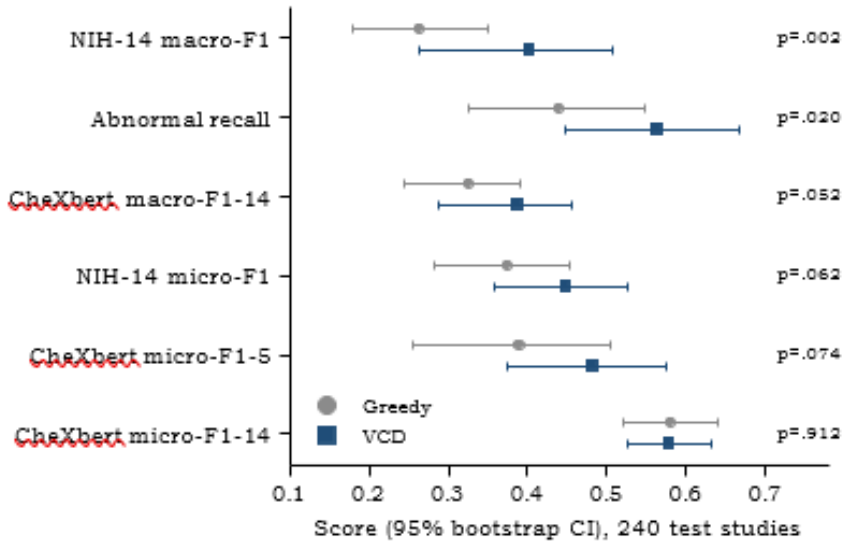


Figure 3. Greedy vs. VCD on the six clinical metrics, with 95% bootstrap intervals and paired bootstrap p-values. Metrics are ordered by p-value.

Fig. 3), and lifted the share of unique reports from 27.5% to 40.4% (66 to 97 distinct texts, Fig. 4b). NIH-14 micro-F1, CheXbert macro-F1-14 and CheXbert micro-F1-5 moved with p of 0.062, 0.052 and 0.074. CheXbert micro-F1-14 did not move (0.579 vs. 0.576), probably because that metric is dominated by "No Finding". Overlap metrics fell slightly (BLEU-4 0.176 to 0.162; ROUGE-L 0.381 to 0.366), as expected once boilerplate text is no longer rewarded. The confidence intervals overlap for every metric, and after Holm correction only the macro-F1 gain remains significant (ad-justed $p = 0.012$).

There are three caveats. First, greedy decoding used the training-time hints while VCD used none. The 60-image ablation suggests the hints were neutral for greedy decoding (micro-F1 0.348 with, 0.318 without), but no test arm isolates this. Second, the VCD setting was chosen on only 60 images. Third, abnormal precision fell from 0.889 to 0.746, so the recall gain came with more false abnormal statements. The retrieval baseline shows the limits of mention-based recall: on the same 240 studies its abnormal recall is 0.575, about the same as VCD, yet its CheXbert micro-F1-14 is only 0.399. A 90-minute abnormal-enriched fine-tune with focal weighting raised abnormal recall (0.667 to 0.905) but lowered validation micro-F1 (0.500 to 0.418), so we rejected it.21.2% at raw thresholds, with no genuine agreement (all 51 were cases where both sides were silent), and 55.4% after test-set recalibration, which is not a valid estimate (Fig. 5). So we moved to Plan C. (The Plan A rows of Table III re-evaluate Plan A with the final labeler, so their coverage differs from the 21.2% and 55.4% above, which came from the original keyword matcher.)

Single-referee gate. At the 90% operating point Plan C auto-approved 29.2% of the VCD reports (70 of 240), and all 70 were normal-reading drafts, as the sign-off rule guarantees (Section III-D). Three of them (4.3%; exact CI 0.9–12.0%) had a labeler-detected mismatch with the reference on at least one label, and 71 of the 74 label-mismatched reports (96%) were routed to review (Table III, Fig. 6a). Results on the greedy reports are almost identical (4.2%, 66 of 69 routed). The stricter operating points trade coverage for risk: 20.0% coverage with 1 mismatch at 95% sensitivity, and 9.6% with none at 97.5%. Plan C missed fewer abnormalities than the ungated pipeline: 32 VCD reports said "normal" for a reference-abnormal study, and 3 of them were approved.

Larger test set. On all 1,033 test studies, with retrieval-generated reports (46.0% mismatched without a gate), Plan C approved 23.8% with a 6.9% mismatch rate (17 of 246; exact

CI 4.1–10.8%) and routed 458 of 475 mismatched reports to review (Fig. 6b). Here the comparison with raw Plan A is clear: at a similar coverage (18.4%) Plan A had a 16.3% mismatch rate (11.4–22.4%), and the intervals do not overlap. On the 240 VCD reports the same ordering holds for the point estimates (4.3% vs. 7.4%), but the intervals overlap.

ISSN 2278-2540 | DOI: 10.51583EMAS | Volume XV, Issue X, October2026

Second referee (pre-declared primary). The dual gate approved 14.6% of VCD reports with 1 mismatch (2.9%; exact CI 0.1–14.9%). Compared with the single-referee gate

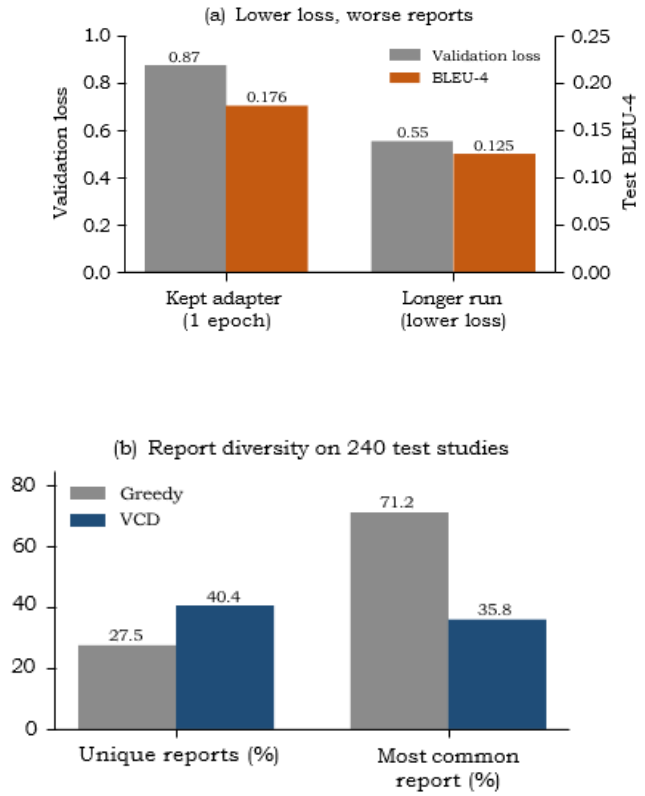


Figure 4. (a) A longer Stage 2 run lowered validation loss from 0.87 to 0.55 while test BLEU-4 fell from 0.176 to 0.125. (b) VCD roughly halves the share of the single most common report.

Table III. VERIFICATION GATES. COV. = SHARE AUTO-APPROVED; RISK = APPROVED REPORTS WITH A LABELER-DETECTED MISMATCH ON ANY OF THE 14 LABELS (95% BOOTSTRAP AND EXACT CIS); CAUGHT = LABEL-MISMATCHED REPORTS ROUTED TO REVIEW. PLAN C ROWS USE THE PRE-DECLARED SENSITIVITY TARGETS. † THRESHOLDS FITTED ON TEST IMAGES (LEAKY). ‡ PRE-DECLARED PRIMARY COMPARISON.

Reports	Gate	Cov. %	Mism. appr.	Risk % (bootstrap CI)	Exact CI	Caught
VCD, 240	No gate	100	74	30.8 (25.0–36.7)	25.1–37.1	0/74
	Plan A, raw 0.30	22.5	4	7.4 (1.7–15.1)	2.1–17.9	70/74
	Plan A, test-fit thr.†	61.7	19	12.8 (7.5–18.5)	7.9–19.3	55/74
	Plan C, 90% sens.	29.2	3	4.3 (0.0–9.7)	0.9–12.0	71/74
	Plan C, 95% sens.	20.0	1	2.1 (0.0–7.1)	0.1–11.1	73/74
	Plan C, 97.5% sens.	9.6	0	0.0	0.0–14.8	74/74
	Plan C dual, 90%‡	14.6	1	2.9 (0.0–9.7)	0.1–14.9	73/74
	Plan C	30.8	5	6.8 (1.4–13.1)	2.2–15.1	69/74

ISSN 2278-2540 | DOI: 10.51583EMAS | Volume XV, Issue X, October2026

	fused, 90%					
Greedy, 240	No gate	100	69	28.7 (23.3–35.0)	23.1–34.9	0/69
	Plan C, 90% sens.	29.6	3	4.2 (0.0–9.7)	0.9–11.9	66/69
Retrieval, 1,033	No gate	100	475	46.0 (42.8–49.0)	42.9–49.1	0/475
	Plan A, raw 0.30	18.4	31	16.3 (11.1–21.8)	11.4–22.4	444/475
	Plan C, 90% sens.	23.8	17	6.9 (3.9–10.2)	4.1–10.8	458/475
	Plan C, 95% sens.	14.2	8	5.4 (2.1–9.4)	2.4–10.4	467/475
	Plan C dual, 90%‡	12.1	7	5.6 (1.8–10.0)	2.3–11.2	468/475

fused. A classifier trained on other hospitals did not transfer well enough to IU to catch the errors Model 1 misses, so the pre-declared hypothesis was not supported. These fig-ures describe routing on one dataset. The mismatch rates are labeler-detected, not radiologist-confirmed clinical error rates, and they are not a safety measure.

External Validation and Calibration

Discrimination transfers. On VinDr the mean AUC over the eight findings with at least 10 positives was 0.894 (95% CI 0.883–0.906) for the NIH heads and 0.901 (0.889–0.913)

for the selected IU-adapted heads; the difference was not sig-nificant (CI −0.002 to 0.014). The normal-versus-abnormal

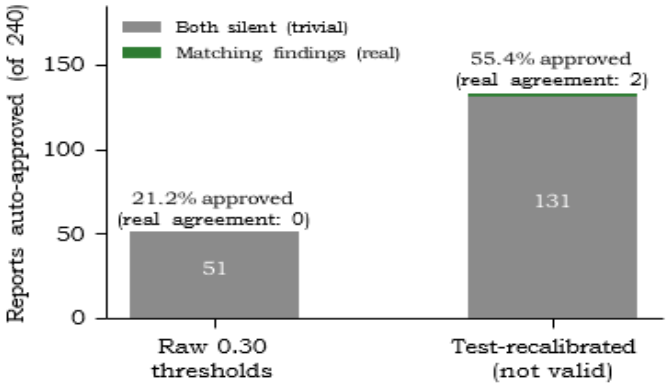


Figure 5. Plan A on greedy reports. Almost every "agreement" was both models saying nothing; only 2 of 240 were matching findings, even after fitting thresholds on the test images.

on the same reports, risk changed by −1.4 percentage points (95% CI −6.7 to +3.3) while coverage fell by 14.6 points (−19.2 to −10.0). On the 1,033 retrieval reports the change was −1.3 points (−4.6 to +1.8) for a coverage loss of 11.7 points, and at the stricter operating points the dual gate's risk

was not lower at all. The reason is visible in the abnormal-ity AUCs on the 1,033 test studies: 0.838 for Model 1 but only 0.740 for the outside referee, and 0.835 when both are AUC of the abnormality score was 0.891 (0.878–0.904) for any radiologist finding and 0.900 for the nine shared find-ings (Table IV). The CheXpert/MIMIC referee again lagged (0.775), and fusing it changed little (0.894). These AUCs are higher than on IU, plausibly because radiologist labels are cleaner than report-mined ones, so the two datasets should not be ranked against each other.

ISSN 2278-2540 | DOI: 10.51583EMAS | Volume XV, Issue X, October2026

Calibration partly transfers. Averaged over findings, calibration reduced the ECE from 0.206 to 0.011 on the IU test studies and from 0.30 to 0.037 on VinDr, and the Brier score from 0.109 to 0.027 and from 0.156 to 0.032 (Fig. 7a). Calibration fitted on IU is therefore much better

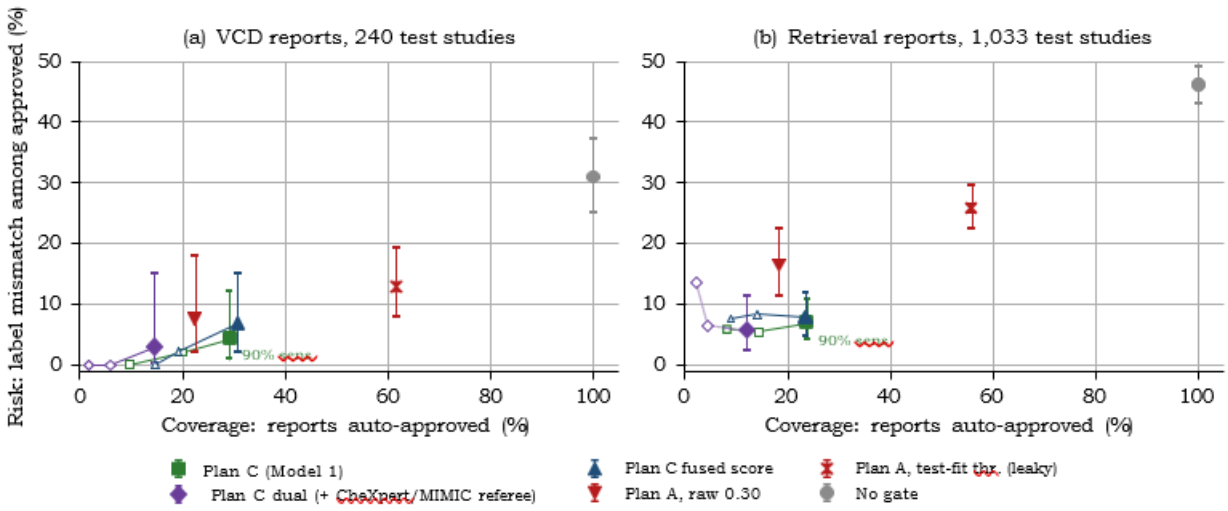


Figure 6. Coverage and risk of each gate. Filled markers with exact 95% CIs are the 90%-sensitivity operating points; open markers are the 95% and 97.5% points of the same gate. (a) 240 VCD reports. (b) 1,033 retrieval reports.

Table IV. EXTERNAL VALIDATION ON 3,000 VINDR-CXR IMAGES (RADIOLOGIST MAJORITY LABELS), WITH THE FROZEN IU PIPELINE. AUC OF THE

NIH HEAD VS. THE SELECTED IU-ADAPTED HEAD; ECE OF RAW NIH VS. IU-CALIBRATED PROBABILITIES; SHARE OF IMAGES FLAGGED AT A RAW 0.30 THRESHOLD VS. THE FROZEN CALIBRATED THRESHOLDS.

Finding	Pos.	Prev. %	AUC NIH	AUC IU-adapted	ECE raw	ECE cal.	Flag raw 0.30 %	Flag cal. %
Atelectasis*	8	0.3	0.598	0.775	0.274	0.076	41.0	11.7
Cardiomegaly	407	13.6	0.878	0.935	0.176	0.069	47.5	9.5
Pleural effusion	141	4.7	0.971	0.971	0.182	0.020	28.8	5.0
Infiltration	51	1.7	0.849	0.854	0.409	0.014	87.3	6.4
Nodule/mass	78	2.6	0.883	0.889	0.375	0.026	63.4	3.3
Pneumothorax	13	0.4	0.946	0.946	0.218	0.003	28.0	2.2
Consolidation	19	0.6	0.963	0.963	0.296	0.014	42.0	1.8
Pulmonary fibrosis	210	7.0	0.817	0.779	0.449	0.062	62.5	0.6
Pleural thickening	201	6.7	0.850	0.870	0.344	0.052	52.3	4.7
Mean (8 findings, ≥10 pos.)			0.894	0.901				

*Fewer than 10 positives; excluded from the mean AUC. ECE means over all nine findings: raw 0.30, calibrated 0.037.

ISSN 2278-2540 | DOI: 10.51583EMAS | Volume XV, Issue X, October2026

than raw probabilities at the new hospital, but about three times less accurate than at home. It does not transfer evenly: infiltration went from 87.3% flagged at raw 0.30 to 6.4% against a radiologist prevalence of 1.7%, while pulmonary fibrosis went from 62.5% to only 0.6% against a prevalence of 7.0%, so the IU threshold under-flags fibrosis at VinDr. The abnormality score was also somewhat overconfident on VinDr (mean score 0.347 vs. 0.257 prevalence of shared findings; ECE 0.102 vs. 0.040 on IU), an error in the cautious direction.

The gate's screen. At the 90% operating point, 28.7% of VinDr images (861 of 3,000) would pass the screen, against 26.7% of IU test studies, and 96.6% of radiologist-abnormal images were flagged. Of the passing images, 3.6% (31 of 861; exact CI 2.5–5.1%) had a radiologist finding of any kind and 2.4% (1.5–3.7%) one of the nine shared findings, compared with 6.9% abnormal among passing IU test studies under report-derived labels. The stricter points behaved as on IU: 14.4% passing with 2.5% abnormal at 95%, and 8.6% with 1.9% at 97.5% (Fig. 7b). The dual gate passed only 8.8% of images at 90%, with 2.3% abnormal, again giving up coverage for no clear gain over the single gate at a stricter point.

Interpreting the residual risk. The abnormal images that passed the screen at 90% were dominated by findings the system was never trained to report: aortic enlargement (13 images), calcification, "other lesion" and interstitial lung disease, together with pulmonary fibrosis (11), which the IU calibration under-flags, and a few nodules or masses (5). On IU, every approved mismatch was likewise a missed abnormality. In both datasets the residual risk is therefore concentrated in subtle findings and in findings outside the 14-class label set, which is exactly where an automatic sign-off is least acceptable. A deployment would need the stricter operating points, an explicit list of findings the gate cannot vouch for, and human reading of every approved report until a radiologist study shows otherwise.

A. Failure Modes

Calibration and domain shift. At a raw 0.30 threshold the NIH model flagged infiltration on 70.8% and fibrosis on 59.2% of the 240 IU test images (73.3% and 59.8% of all 1,033; Fig. 8), even though its ranking on IU was acceptable. After calibration and per-class thresholds the flag rates fell to 6.7% and 0.8%; the same pattern, with fibrosis now under-flagged, appeared at VinDr (Section V-D). AUC measures ranking. It cannot tell us which probability should count as "positive" in a new hospital.

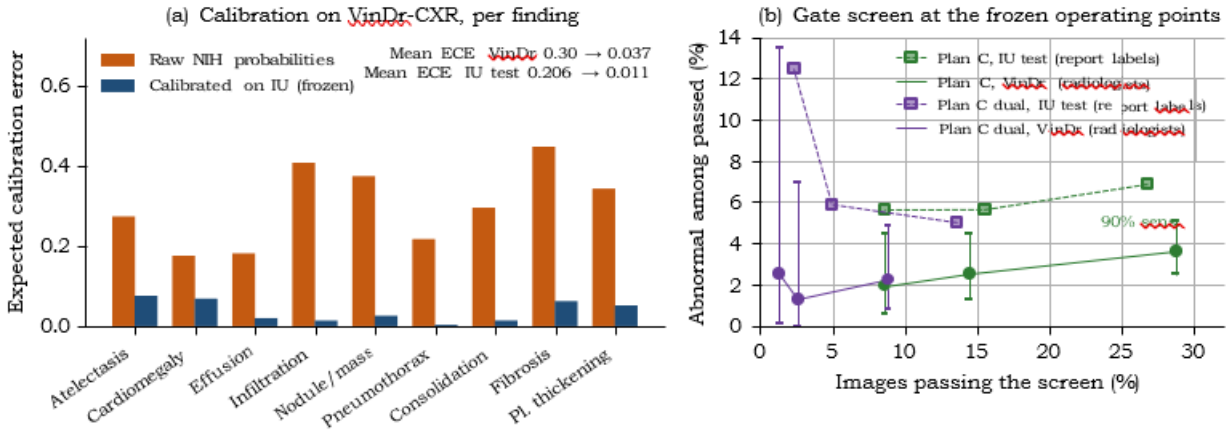


Figure 7. External validation on VinDr-CXR. (a) Per-finding expected calibration error of raw NIH probabilities and of the frozen IU-calibrated probabilities.

(b) Share of images passing the gate's screen and share of those that are abnormal, at the three frozen operating points; VinDr (radiologist labels, exact 95% CIs) vs. IU test studies (report-derived labels). The VinDr risk assumes a normal-reading draft for every image.

B. What the Results Do and Do Not Show

Our results support four points. First, a classifier calibrated in-domain can flag normal-reading drafts that conflict

ISSN 2278-2540 | DOI: 10.51583EMAS | Volume XV, Issue X, October2026

with the image evidence (29 of 32 such VCD reports went to review), and on 1,033 studies it clearly beats raw classifier agreement. Second, the same gate cannot catch mistakes that both mod-els share, and adding an outside referee trained elsewhere did not change that. Third, VCD reduces template collapse but does not remove it, and its gains show up in macro-F1 (rarer findings) rather than in micro-F1-14, which common and normal labels dominate. Fourth, the classifier side of the gate held up at an external hospital with radiologist la-bels: discrimination was maintained, calibration improved on raw probabilities though less than at home, and the screen's worst-case risk at the same operating point was 3.6%.

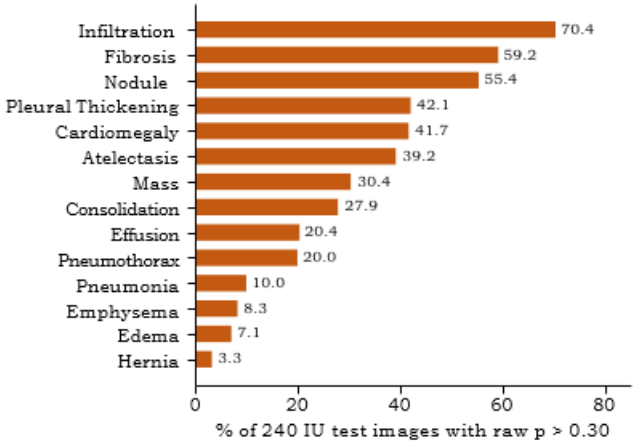


Figure 8. Share of the 240 IU test images that the NIH-trained Model 1

flags at a raw probability of 0.30, from the earlier Plan A evaluation. The final pipeline, which resizes images slightly differently, gives 70.8% for infiltration and the same 59.2% for fibrosis.

Template collapse. Validation loss did not track report quality. Across four Stage 2 runs, one lowered validation loss from 0.87 to 0.55 while BLEU-4 fell from 0.176 to 0.125, and in a 30-sample check the single most common output made up 73% of the samples (Fig. 4a). The adapter we evaluated wrote one identical normal report for 71.2% of test studies. VCD roughly halves that to 35.8% but does not remove it. Raw and calibrated classifier hints did not help in the 60-image ablation, although the adapter had been trained on noisy raw hints, so this does not show that conditioning cannot help.

Correlated errors. Every approved mismatch was a missed abnormality: the report stated no finding and Model 1's abnormality score was below threshold (3 of 3 for VCD reports, 17 of 17 for retrieval reports). Both models learned from the same IU pool and the same report-derived labels, so shared blind spots and label noise are invisible to the gate. The second-referee experiment shows that an off-the-shelf classifier from other hospitals does not close this gap. Estimating the residual risk would need indepen-dent evidence, such as a domain-adapted outside model or radiologist review.

explanation, which we did not test, is that penalizing tokens the language prior already favors moves probability mass from boilerplate to finding-specific tokens.

They do not show that VCD's image contrast, rather than dropping the hints or our selection procedure, produced the gain, or that there is any clinical benefit. Lower risk always came with lower coverage, and at about 29% coverage most normal-reading drafts still go to a reader. A model that always wrote a normal report would probably score a high micro-F1-14; we did not run that baseline, so we report macro-F1, diversity and abnormal recall next to it.

LIMITATIONS AND ETHICAL CONSIDERATIONS

The VCD and greedy comparisons rest on 240 studies, 73 of them abnormal, and many findings have only a handful of positives; with 3 mismatches among 70 approved reports the interval for risk is wide. Labels come from a rule/CheXbert labeler applied to reports, and a "mismatch" means a labeler-detected disagreement with

ISSN 2278-2540 | DOI: 10.51583EMAS | Volume XV, Issue X, October2026

the reference labels, not an error judged by a clinician. There was no radiologist review of the generated reports, and IU is a single institution with patient-level separation unconfirmed. The external validation covers only the classifier and the gate's screen: VinDr has no reports, so MedGemma was not tested outside IU, the screen assumes a normal-reading draft, only nine findings are shared between the two label sets, and the sample comes

from VinDr's public training split. The VCD comparison is confounded by the hints and rests on a small selection set, and our metric implementations differ from the published IU protocols, so we do not compare against published numbers. The demonstration app's gate uses the rule labeler only.

Threats to validity. The reference labels, Model 1's train-ing and calibration labels, and the mismatch criterion all come from the same labeler, so labeler quirks can make agree-ment look better than agreement with the real image findings. A MeSH cross-check showed 95.5–99.8% per-finding agree-ment, but that figure includes true negatives and does not validate precision. Model 1's development pool is the same set of studies Model 2 was trained and validated on, which is one source of the shared blind spots. The VCD setting was chosen on 60 validation images.

AUTO-APPROVE is an experimental routing label, not a clinical authorization. A missed pneumothorax or effusion in an approved report could delay care, false statements could mislead readers, and fluent text may be trusted too easily. Model outputs are predictions, not confirmed findings, and the system must not be used for diagnosis. Grad-CAM maps are explanatory aids only.

CONCLUSION AND FUTURE WORK

Calibrated classifier probabilities worked as a usable screen for normal-reading reports. Every auto-approved report stated no finding; at about 29% coverage, 4.3% of approved VCD reports had a labeler-detected label mismatch and 96% of label-mismatched reports were sent to review, and on 1,033 studies the mismatch rate among approved reports was 6.9%. These are label-mismatch rates, not radiologist-confirmed error rates. On an external dataset with radiologist labels, the frozen classifier kept its discrimination, its calibration remained far better than raw probabilities, and at the same operating point 3.6% of images that would pass the gate were abnormal. A second referee trained on CheXpert and MIMIC-CXR did not lower risk, because it transferred poorly to IU. The image-free-contrast VCD variant improved macro-F1 and diversity without retraining, but not the overlap metrics, and only the macro-F1 gain holds up under multiplicity cor-rection. Since the greedy baseline used classifier hints and VCD did not, part of this change may come from removing the hints. The pipeline is still a research prototype. Its errors are correlated and its evaluation is built on report-derived labels.

Future work (not completed): recompute BLEU, ME-TEOR and ROUGE-L with the standard scorers and fix RadGraph; add a greedy-without-hints arm and an image-distortion VCD baseline; ablate the gate rules; try a second referee that is adapted to IU on data Model 2 never saw; recalibrate per site and add findings outside the 14 classes (such as aortic enlargement); test report generation on an external dataset with reports, such as ReXGradient-160K or MIMIC-CXR; and have radiologists review approved and rejected cases.

ACKNOWLEDGMENT

This work was carried out as a final-year B. Tech project at the School of Computer Science and Engineering (SCOPE), VIT-AP University, under the guidance of Dr. Khasim Syed.

We thank him for his supervision and guidance throughout the project. Model training and evaluation used the free GPU quota provided by Kaggle. We also thank the NIH Clinical Center and Indiana University for making their chest X-ray datasets publicly available.

DATA AND CODE AVAILABILITY

NIH ChestX-ray14 [5] and the IU chest X-ray collection [17] are publicly available from their providers. The external evaluation used VinDr-CXR [21] training images and labels as distributed through the VinBigData

ISSN 2278-2540 | DOI: 10.51583EMAS | Volume XV, Issue X, October2026

Kaggle competition, under its non-commercial research terms. The second referee uses publicly released TorchXRayVision weights [18]; we did not access CheXpert or MIMIC-CXR images. The code, trained model weights and generated reports from this study are kept private and are not publicly released.

FUNDING

This work received no external funding.

Ethics Statement

This study used only existing, publicly available, de-identified datasets (NIH ChestX-ray14, the IU chest X-ray collection and VinDr-CXR) and publicly released pretrained model weights. No new images or data were collected, and no patients, radiologists or other human participants took part. The system is a research prototype and is not intended for clinical use.

REFERENCES

1. A. Sellergren et al., "MedGemma technical report," arXiv:2507.05201, 2025.
2. Z. Babar, T. van Laarhoven, and E. Marchiori, "Encoder-decoder models for chest X-ray report generation perform no better than un-conditioned baselines," PLoS ONE, vol. 16, no. 11, Art. no. e0259639, 2021, doi: 10.1371/journal.pone.0259639.
3. S. Leng, H. Zhang, G. Chen, X. Li, S. Lu, C. Miao, and L. Bing, "Mit-igating object hallucinations in large vision-language models through visual contrastive decoding," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2024, pp. 13872–13882.
4. Y. Geifman and R. El-Yaniv, "Selective classification for deep neural networks," in Adv. Neural Inf. Process. Syst. (NeurIPS), 2017.
5. X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 2097–2106.
6. M. Tan and Q. V. Le, "EfficientNetV2: Smaller models and faster training," in Proc. Int. Conf. Mach. Learn. (ICML), 2021, pp. 10096–10106.
7. Z. Chen, Y. Song, T.-H. Chang, and X. Wan, "Generating radiology reports via memory-driven transformer," in Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP), 2020, pp. 1439–1449.
8. H. Jin, H. Che, Y. Lin, and H. Chen, "PromptMRG: Diagnosis-driven prompts for medical report generation," in Proc. AAAI Conf. Artif. Intell., vol. 38, no. 3, 2024, pp. 2607–2615, doi: 10.1609/aaai.v38i3.28038.
9. E. J. Hu et al., "LoRA: Low-rank adaptation of large language mod-els," in Proc. Int. Conf. Learn. Represent. (ICLR), 2022.
10. T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient finetuning of quantized LLMs," in Adv. Neural Inf. Process. Syst. (NeurIPS), 2023.
11. H. Abdaoui, S. Barbaria, A. Al Kuwaiti, N. Raouafi, R. Ben Salah, and H. Boussi Rahmouni, "Swin-Qwen3: A three-stage vision–language framework for automated radiology report generation with multi-agent verification," Front. Radiol., vol. 6, Art. no. 1875905, 2026, doi: 10.3389/fradi.2026.1875905.
12. R. Mahmood et al., "Phrase-grounded fact-checking for automatically generated chest X-ray reports," arXiv:2509.21356, 2025.
13. A. Smit, S. Jain, P. Rajpurkar, A. Pareek, A. Y. Ng, and M. P. Lungren, "CheXbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT," in Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP), 2020, pp. 1500–1519.
14. J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in Advances in Large Margin Classifiers, A. J. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, Eds. Cambridge, MA, USA: MIT Press, 1999, pp. 61–74.
15. C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in Proc.

ISSN 2278-2540 | DOI: 10.51583EMAS | Volume XV, Issue X, October2026

Int. Conf. Mach. Learn. (ICML), 2017, pp. 1321–1330.
16. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2017, pp. 618–626.
17. D. Demner-Fushman et al., “Preparing a collection of radiology examinations for distribution and retrieval,” J. Am. Med. Inform. Assoc., vol. 23, no. 2, pp. 304–310, 2016, doi: 10.1093/jamia/ocv080.
18. J. P. Cohen et al., “TorchXRayVision: A library of chest X-ray datasets and models,” in Proc. Int. Conf. Medical Imaging with Deep Learning (MIDL), PMLR, vol. 172, 2022, pp. 231–249.
19. J. Irvin et al., “CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison,” in Proc. AAAI Conf. Artif. Intell., vol. 33, 2019, pp. 590–597, doi: 10.1609/aaai.v33i01.3301590.
20. A. E. W. Johnson et al., “MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports,” Sci. Data, vol. 6, Art. no. 317, 2019, doi: 10.1038/s41597-019-0322-0.
21. H. Q. Nguyen et al., “VinDr-CXR: An open dataset of chest X-rays with radiologist’s annotations,” Sci. Data, vol. 9, Art. no. 429, 2022, doi: 10.1038/s41597-022-01498-w.