

# Process Reward Models for Sentence-Level Verification of LVLM Radiology Reports

Alois Thomas   Maya Varma   Jean-Benoit Delbrouck   Curtis P. Langlotz

AIMI Center, Stanford University, CA, USA

{aloistho, mayavarma, jbde1, langlotz}@stanford.edu

## Abstract

Automating radiology report generation with Large Vision-Language Models (LVLMs) holds great potential, yet these models often produce clinically critical hallucinations, posing serious risks. Existing hallucination detection methods frequently lack the necessary sentence-level granularity or robust generalization across different LVLM generators. We introduce a novel approach: a sentence-level Process Reward Model (PRM) adapted for this vision-language task. Our PRM predicts the factual correctness of each generated sentence, conditioned on clinical context and preceding text. When fine-tuned on MIMIC-CXR with weakly-supervised labels, a lightweight 0.5B-parameter PRM outperforms existing verification techniques, demonstrating, for instance, relative improvements of 7.5% in Matthews Correlation Coefficient and 1.8% in AUROC over strong white-box baselines on outputs from one LVLM. Unlike methods reliant on internal model states, our PRM demonstrates strong generalization to an unseen LVLM. We further show its practical utility: PRM scores effectively filter low-quality reports, improving F1-CheXbert scores by 4.5% (when discarding the worst 10% of reports). Moreover, when guiding a novel weighted best-of-N selection process on the MIMIC-CXR test set, our PRM show relative improvements in clinical metrics of 7.4% for F1-CheXbert and 0.6% for BERTScore. These results demonstrate that a lightweight, context-aware PRM provides a model-agnostic safety layer for clinical LVLMs without access to internal activations

## 1 Introduction

Large Language Models (LLMs) and Vision-Language Models (VLMs or LVLMs) have achieved remarkable success in generating free-text content, including in healthcare (Chen et al., 2024a; Bannur et al., 2024). Radiology report generation involves translating medical images (like X-rays or CT scans) and patient-specific clinical information into a structured textual summary of findings. Automating this complex task can help alleviate the increasing workload on radiologists and improve patient care. In radiology, these models offer potential for automating report generation from medical images and patient context, possibly alleviating radiologist workload (Paschali et al., 2025). However, a significant challenge remains: the propensity of these models to generate "hallucinations", factually incorrect statements, which pose severe risks in clinical settings where diagnostic accuracy is important (Kim et al., 2025).

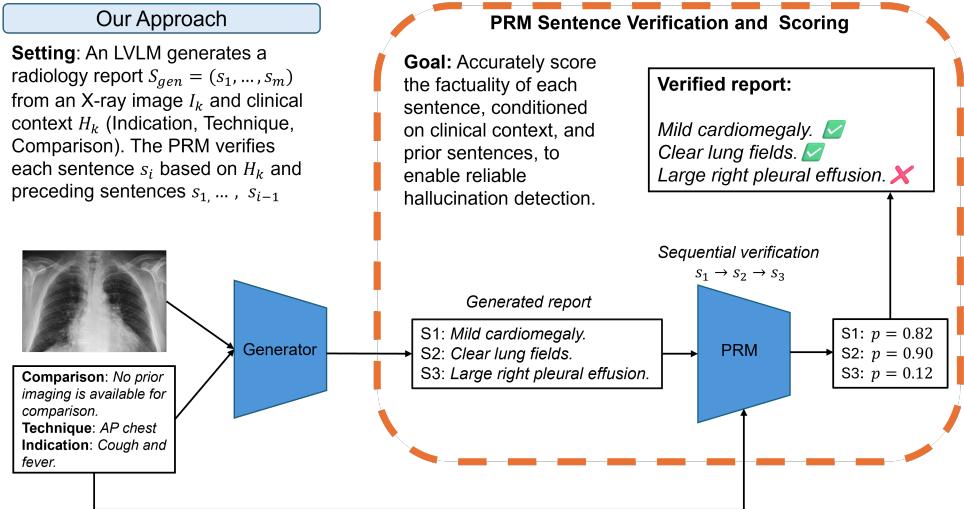


Figure 1: Overview of our proposed sentence-level Process Reward Model (PRM) for verifying radiology reports. The PRM takes clinical context and previously generated/verified sentences as input to predict the correctness of the current sentence, enabling fine-grained hallucination detection.

Existing hallucination detection methods often operate at the report level or lack robustness across different generator models (Chen et al., 2024a). Black-box methods, relying on sampling consistency or entailment checks, can be computationally expensive (Zhang et al., 2024; Manakul and Gales, 2023). White-box methods, using internal model states like hidden activations or logits, may offer efficiency but often exhibit poor generalization when applied to different LVLMs (Hardy et al., 2024; Azaria and Mitchell, 2023). Neither approach typically provides the sentence-level granularity needed in radiology, where a single incorrect sentence (e.g., misstating the presence of a critical finding like pneumothorax) can have serious clinical consequences, potentially leading to misdiagnosis, delayed or incorrect treatment, and ultimately, adverse patient outcomes. Therefore, identifying and mitigating errors at the sentence level is important for safe clinical deployment.

To address these limitations, we propose adapting Process Reward Models (PRMs), initially developed for evaluating step-by-step reasoning in LLMs (Lightman et al., 2023), to the task of sentence-level factuality verification in radiology reports. Instead of assessing mathematical reasoning steps, our modified PRM is a novel adaptation for this multimodal clinical domain, and learns to sequentially evaluate the factual grounding of radiology report sentences. It assigns a correctness probability to each sentence, conditioned on the full clinical context (patient history, imaging technique, comparison studies) and all previously generated and verified sentences in the report. This fine-grained, context-aware verification operates in a black-box manner, requiring only the generated text and clinical prompt, making it broadly applicable. Our contributions are:

1. We propose a novel sentence-level Process Reward Model (PRM) specifically designed for verifying LVLG-generated radiology findings. This PRM operates sequentially, conditioned on clinical context and prior generated text, and functions as a black-box verifier, marking a new application of PRMs to fine-grained, multimodal verification in this critical healthcare domain.
2. We introduce a scalable training methodology for this PRM using a large corpus ( $\sim 200k$  instances) of weakly-supervised sentence correctness labels derived from an existing domain-specific entailment model (RadNLI).
3. Through extensive experiments on the MIMIC-CXR dataset (Johnson et al., 2019), we demonstrate that our PRM verifier, leveraging the proposed approach, outperforms strong

baselines (including ReXTrust, [Hardy et al. \(2024\)](#)) in sentence-level hallucination detection and exhibits superior transferability to an unseen LVLM.

4. We further showcase its practical utility by showing how PRM scores can effectively filter low-quality reports via rejection sampling and improve overall report quality via Best-of-N selection, especially using a novel weighted strategy leveraging CheXbert labels.

This work offers a path towards more reliable and clinically safer deployment of LVLMs in radiology.

## 2 Approach

We formalise sentence-level hallucination detection for radiology reports, and introduce a process reward model (PRM) that performs this verification in-context. We position two strong baselines against it, and describe how the resulting sentence probabilities can lead to practical downstream improvements of radiology LVLMs.

### 2.1 Problem statement and notation

**Task.** For each study  $k$  we observe a chest X-ray image  $I_k$  and structured clinical context  $H_k$  comprising the indication, imaging technique and any comparison studies. A report generator produces a findings section  $S_k^{\text{gen}} = (s_{k,1}^{\text{gen}}, \dots, s_{k,m_k}^{\text{gen}})$ . Our goal is to estimate, for every sentence  $s_{k,i}^{\text{gen}}$ , the probability that it is factually correct with respect to  $\{I_k, H_k\}$ .

**Supervision.** We label sentences automatically with RadNLI ([Yuan et al., 2021](#)) (which achieves 80.0% accuracy on its test set, see Appendix D Fig. 4): if any ground-truth sentence semantically entails  $s_{k,i}^{\text{gen}}$  we assign  $y_{k,i} = 1$ , otherwise  $y_{k,i} = 0$ . This yields a noisy but large training corpus that we balance to an approximate 1:1 ratio (details in Appendix D).

### 2.2 Sentence-level process reward model

**Core idea.** We instantiate the PRM as a sequential binary decoder: given the clinical prompt and the running prefix of previously generated sentences with their predicted labels, the model outputs a token  $y_{k,i} \in \{0, 1\}$  that denotes the factual status of the current sentence.

**Input construction.** During training we provide the model with the interleaved sequence

$$[H_k, s_{k,1}^{\text{gen}}, \text{\textbackslash n}, y_{k,1}, \dots, s_{k,m_k}^{\text{gen}}, \text{\textbackslash n}, y_{k,m_k}],$$

where  $\text{\textbackslash n}$  is a hard separator. At inference the labels are omitted; the verifier must supply them.

**Model backbone and scaling.** We fine-tune Qwen2.5-0.5B-base and Qwen2.5-3B-base ([Qwen et al., 2024](#)) with the TRL PRM-trainer, thereby evaluating the effect of parameter count on verification quality (training protocol in Section 3.2).

**Objective.** Let  $\phi$  denote the PRM parameters and  $\text{prefix}_{k,i} = [H_k, s_{k,1}^{\text{gen}}, \text{\textbackslash n}, y_{k,1}, \dots, s_{k,i}^{\text{gen}}, \text{\textbackslash n}]$ . Constraining the logits to  $\{‘0’, ‘1’\}$ , we compute

$$p_{\phi,k,i} = p_{\phi}(‘1’ | \text{prefix}_{k,i}),$$

and minimise the binary cross-entropy

$$\mathcal{L}_{\text{PRM}}(\phi) = - \sum_k \sum_i \left[ y_{k,i} \log p_{\phi,k,i} + (1 - y_{k,i}) \log(1 - p_{\phi,k,i}) \right]. \quad (1)$$

At test time the verifier outputs  $\{p_{\phi,k,i}\}_{i=1}^{m_k}$ , a sentence-level plausibility map for the generated report.

### 2.3 Baseline verifiers

**ReXTrust (white-box).** Following Hardy et al. (Hardy et al., 2024), we feed MAIRA-2’s 16-th layer token embeddings for each sentence into a multi-head self-attention block, mean-pool the result, and apply a linear classifier.

**Grey-box MLP.** A two-layer perceptron inspired by Liu et al. (Liu et al., 2024) uses a 13-dimensional feature vector of sentence-level token statistics (length, logit and entropy derived features) and outputs a correctness logit.

Hyper-parameters for both baselines are listed in Appendix F.2.

### 2.4 Downstream applications

**Report rejection.** Aggregating the sentence-level correctness probabilities produces a scalar report score  $\text{Score}_k = f(\{p_{k,i}\})$ , which can be used to filter low-quality reports. Let  $p_{\text{PRM},k,i}$ ,  $p_{\text{Rex},k,i}$ , and  $p_{\text{MLP},k,i}$  denote the predicted correctness probability for sentence  $i$  of report  $k$  from the PRM, ReXTrust, and MLP verifiers, respectively, and let  $m_k$  be the number of sentences in report  $k$ . For the rejection experiments (Section 4.3), we study several aggregation functions  $f$  (matching labels like ‘PRM avg’, ‘PRM prod’ in Fig 2):

- **MinProb (min):** The minimum sentence probability across the report ( $\text{Score}_k = \min_i p_{\text{PRM},k,i}$ ). Applied to PRM, ReXTrust, MLP.
- **AvgProb (avg):** The arithmetic mean of sentence probabilities ( $\text{Score}_k = \frac{1}{m_k} \sum_i p_{\text{PRM},k,i}$ ). Applied to PRM, ReXTrust, MLP.
- **ProdProb (prod):** The geometric mean of sentence probabilities ( $\text{Score}_k = (\prod_{i=1}^{m_k} p_{\text{PRM},k,i})^{1/m_k}$ ). Applied to PRM, ReXTrust, MLP.
- **Entropy (baseline):** The average sentence-level token entropy calculated from the generator model’s output probabilities (MAIRA-2). Lower scores indicate higher generator confidence.

Reports are ranked by these scores. To evaluate rejection, we discard reports below a tunable percentile threshold (e.g., the worst 10% based on the chosen score).

**Best-of-N selection.** For each study we sample  $N=128$  candidate reports from MAIRA-2 (temperature 1.0, top- $p$  0.9, top- $k$  50) and rank them by: (i) generator log-probability; (ii) PRM-derived scores; or (iii) a weighted PRM strategy that first groups candidates by CheXbert label vector (Smit et al., 2020), sums PRM scores within each group, and then selects the top-scoring candidate in the best group (full algorithm in Appendix G). Evaluation relies on BLEU, ROUGE, BERTScore, F1-RadGraph and F1-CheXbert.

Together these components define a pipeline that not only flags hallucinated sentences but also demonstrably improves report quality.

## 3 Experimental Setup

This section details the data, training protocol, and evaluation methodology used to quantify the effectiveness of our sentence-level PRM verifier relative to strong grey- and white-box baselines.

### 3.1 Dataset and preprocessing

**Dataset.** We adopt MIMIC-CXR v2.0.0 (Johnson et al., 2019) and adhere to its official train, validation, and test splits.

**Report generation and sentence labelling.** For each study we prompt MAIRA-2 to generate a findings section, segment the output into sentences, and determine for every sentence

whether it is entailed by at least one ground-truth sentence using the domain-specific RadNLI model (Yuan et al., 2021). Sentences predicted as entailment receive the label  $y=1$  (correct); all others are marked  $y=0$  (hallucinated). Initial labeling revealed class imbalance. We down-sampled the majority class in the training, validation, and test sets to achieve an approximate 1:1 ratio, mitigating potential bias. Full details on the balancing procedure and the final sentence counts are provided in Appendix D.

**Input representations.** ReXTrust uses hidden states from MAIRA-2’s 16-th layer, while the grey-box baseline relies on a 13-dimensional vector of sentence-level token statistics (length, logit and entropy moments). The PRM receives a single sequence comprising the clinical prompt followed by each generated sentence, with labels interleaved during training as described in Section 2.2. Appendix E illustrates the exact formatting.

**Transferability corpus.** To assess out-of-distribution generalisation we repeat the above procedure with CheXagent-8B (Chen et al., 2024b), yielding an independent test bed of automatically labelled reports.

### 3.2 Model training

All models are trained on the same sentence-balanced splits. We fine-tune Qwen2.5-0.5B and Qwen2.5-3B with the TRL PRM-trainer for three epochs, using AdamW ( $lr = 10^{-5}$ ) with linear warm-up and decay, an effective batch size of 16 (2 samples per GPU, 8 gradient-accumulation steps), gradient checkpointing, and a maximum sequence length of 1,024 tokens.

For ReXTrust we follow Hardy et al. (Hardy et al., 2024): AdamW ( $lr = 10^{-4}$ ), cosine annealing, weight decay 0.01, batch size 128, dropout 0.1, and early stopping after three stagnant validation epochs. The grey-box MLP (two hidden layers) is optimised with Adam at  $lr = 10^{-3}$  for 20 epochs and the same batch size. Each model minimises binary cross-entropy. We run the training of PRMs on a single A100 GPU. Best-of- $N$  experiments run on an 8×H100 GPU cluster. Full hyper-parameters appear in Appendix F.

### 3.3 Evaluation metrics

Sentence-level discrimination is reported via accuracy, macro-averaged F1, Matthews correlation coefficient (MCC), AUROC, and AUPRC. Report-level quality is assessed with two clinical metrics—F1-CheXbert (Smit et al., 2020) and F1-RadGraph (Delbrouck et al., 2024)—and three textual metrics (BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and BERTScore (Zhang et al., 2019)).

### 3.4 Scope of experiments

We design seven complementary studies: (i) baseline comparison on sentence-level metrics; (ii) generator transfer to CheXagent-8B; (iii) report rejection, where low-scoring reports are discarded; (iv) Best-of-128 sampling, including our weighted selection strategy; (v) context ablation to isolate which prompt fields are indispensable; and (vi) model-scaling analysis contrasting 0.5B and 3B PRMs.

## 4 Results

We present results evaluating the PRM verifier against baselines on verification, transferability, and downstream tasks.

## 4.1 Sentence-level verification performance

We compared our PRM verifiers (Qwen2.5-0.5B, Qwen2.5-3B) against the grey-box MLP (Liu et al., 2024) and the white-box ReXTrust (Hardy et al., 2024) on the MIMIC-CXR test set. Table 1 shows performance with 95% CIs (1000 bootstrap resamples).

Table 1: Sentence-level verification performance on the MIMIC-CXR test set. Metrics computed using 1,000 bootstrap resamples with 95% confidence intervals.

Model	Accuracy	F1 Macro	MCC	AUROC	AUPRC
Grey-box	0.652 (0.628, 0.677)	0.646 (0.622, 0.669)	0.300 (0.253, 0.345)	0.701 (0.674, 0.727)	0.684 (0.647, 0.721)
	0.735	0.733	0.467	0.819	0.798
ReXTrust	(0.712, 0.755)	(0.710, 0.754)	(0.422, 0.510)	(0.798, 0.839)	(0.768, 0.828)
<b>Qwen2.5-0.5B-PRM</b>	<b>0.752</b> (0.728, 0.773)	<b>0.751</b> (0.727, 0.771)	<b>0.502</b> (0.455, 0.544)	0.834 (0.815, 0.853)	0.832 (0.806, 0.856)
Qwen2.5-3B-PRM	0.746 (0.723, 0.771)	0.744 (0.722, 0.769)	0.490 (0.444, 0.539)	<b>0.841</b> (0.821, 0.861)	<b>0.835</b> (0.811, 0.861)

The PRM models consistently outperform both baselines. Qwen2.5-0.5B-PRM achieves the highest Accuracy, F1 Macro, and MCC (+3.5 MCC points over ReXTrust). Qwen2.5-3B-PRM yields the best AUROC (+2.2 points over ReXTrust) and AUPRC, indicating superior discrimination despite slightly lower accuracy/F1 than the 0.5B PRM. This shows the effectiveness of the PRM approach, potentially leveraging contextual understanding better than hidden-state or grey-box derived methods. Qualitative examples are in Appendix H.

**Keyword-specific performance.** We assessed F1-micro scores for sentences containing specific keywords (Table 2). PRM variants generally outperform ReXTrust on most keywords (e.g., "pleural effusion", "consolidation", "pneumothorax"). Qwen2.5-0.5B-PRM leads on "edema", "atelectasis", "left", while Qwen2.5-3B-PRM excels on "tube", "right". This shows PRM superiority extends to clinically relevant terms.

Table 2: F1-micro scores on sentences containing selected keywords (test set), 95% CIs.

Keyword	Count	Qwen2.5-0.5B-PRM F1	Qwen2.5-3B-PRM F1	ReXTrust F1
"pleural effusion"	218	<b>0.789</b> (0.734, 0.839)	0.780 (0.725, 0.830)	0.775 (0.716, 0.830)
"pneumothorax"	215	0.865 (0.819, 0.907)	<b>0.874</b> (0.833, 0.916)	0.870 (0.823, 0.907)
"consolidation"	125	<b>0.664</b> (0.576, 0.752)	0.632 (0.544, 0.720)	0.648 (0.568, 0.728)
"pneumonia"	5	<b>0.800</b> (0.400, 0.800)	<b>0.800</b> (0.400, 0.800)	<b>0.800</b> (0.400, 0.800)
"edema"	51	<b>0.686</b> (0.549, 0.804)	0.667 (0.529, 0.785)	0.608 (0.471, 0.745)
"atelectasis"	26	<b>0.962</b> (0.846, 1.000)	0.923 (0.808, 0.962)	0.923 (0.808, 0.962)
"tube"	42	0.690 (0.571, 0.833)	<b>0.786</b> (0.667, 0.905)	0.738 (0.595, 0.857)
"right"	134	0.672 (0.597, 0.746)	<b>0.739</b> (0.664, 0.806)	<b>0.739</b> (0.664, 0.813)
"left"	93	<b>0.774</b> (0.688, 0.860)	0.699 (0.613, 0.785)	0.710 (0.624, 0.796)

## 4.2 Transferability and generalization

We tested models trained on MAIRA-2 reports on reports from CheXagent-8B (Chen et al., 2024b) (Table 3).

<sup>†</sup> An MCC of 0.000 occurs when a binary classifier predicts only a single class for all samples. Here, ReXTrust failed to identify any hallucinated sentences (predicted all as non-hallucinated), resulting in this degenerate MCC value.

ReXTrust achieves high overall accuracy but has weaker performance on balanced metrics (MCC=0, AUROC<0.5), likely due to majority class prediction, indicating poor generalizability.

Table 3: Transferability performance on CheXagent-8B reports (95% CIs).

Metric	Qwen2.5-0.5B-PRM	Qwen2.5-3B-PRM	ReXTrust
Accuracy	0.700 (0.673, 0.724)	0.699 (0.672, 0.724)	<b>0.805</b> (0.783, 0.825)
F1 Macro	0.599 (0.567, 0.629)	<b>0.627</b> (0.597, 0.654)	0.446 (0.439, 0.452)
MCC	0.220 (0.156, 0.279)	<b>0.306</b> (0.250, 0.358)	0.000 (0.000, 0.000) <sup>†</sup>
AUROC	0.729 (0.697, 0.760)	<b>0.754</b> (0.722, 0.781)	0.411 (0.372, 0.450)
AUPRC	0.368 (0.307, 0.425)	<b>0.379</b> (0.322, 0.437)	0.169 (0.144, 0.196)

In contrast, both PRM models maintain strong performance on balanced metrics (e.g., Qwen2.5-3B-PRM MCC=0.306, AUROC=0.754), demonstrating superior transferability. This suggests PRMs learn more generalizable factuality patterns than white-box methods reliant on specific generator internals.

### 4.3 Rejecting low-quality reports

We used aggregated sentence probabilities (Min Prob, Avg Prob, Prod Prob) from PRM and baselines to score and reject low-quality reports. Figure 2 shows the impact on F1-CheXbert and F1-RadGraph as rejection rate increases.

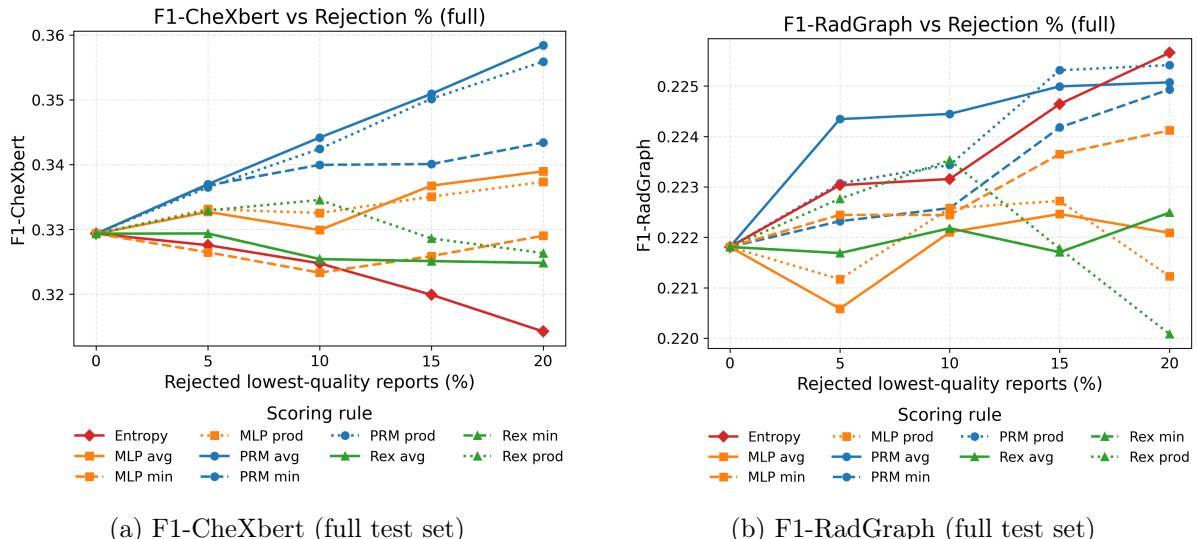


Figure 2: Impact of rejecting lowest-scoring reports (x-axis) on clinical factuality metrics (y-axis) of remaining reports. PRM-*avg* and PRM-*prod* show steepest improvements, indicating effective quality filtering. Entropy also shows good filtering quality for F1-RadGraph scores, sometimes outperforming PRM methods at high rejection thresholds.

PRM-based scoring, particularly Avg Prob and Prod Prob, generally improves report quality as more low-scoring reports are rejected (steeper curves in Fig 2). For instance, discarding the worst-scoring 10 % of reports using PRM-Avg raises F1-CheXbert from 0.329 (baseline F1-CheXbert 0.32934) to 0.344 (0.34417), a relative increase of +4.5 %. Full results across all rejection thresholds (0/5/10/15/20 %) for both F1-CheXbert and F1-RadGraph, along with scores for all aggregation methods, are detailed in App. I, Tab. 6. This confirms PRM scores can effectively identify hallucination-prone reports. The baseline Entropy scoring method also demonstrates strong performance for F1-RadGraph, sometimes exceeding PRM-based aggregation at higher rejection thresholds (Fig. 2b and Tab. 6).

#### 4.4 Best-of-N sampling

We used PRM scores for Best-of-N (BoN) selection from  $N = 128$  candidates (MAIRA-2, temp=1.0). Figure 3 shows key results for weighted strategies.

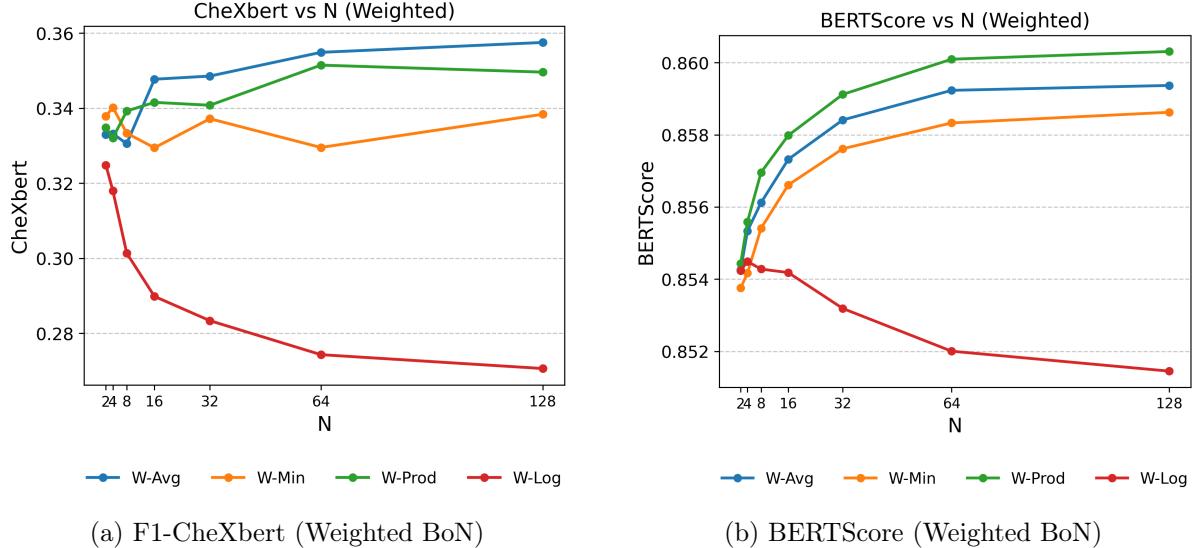


Figure 3: Performance of Weighted BoN strategies vs.  $N$  (temp=1.0). W-Avg and W-Min show strong improvement. Scoring strategies W-Avg, W-Min, W-Prod, and W-Log refer to weighted BoN using base scores from AvgProb, MinProb, ProdProb, and LogProb, respectively.

Increasing  $N$  generally improves quality. Weighted BoN strategies consistently outperform non-weighted ones (details in Appendix J). W-Avg (Weighted AvgProb) and W-Min (Weighted MinProb) yield strong results across clinical and textual metrics (Fig. 3 and App. J, Fig. 13). At  $N=128$ , W-Avg achieves an F1-CheXbert of 0.357 and a BERTScore of 0.859. This improves over non-weighted selection (e.g., the non-weighted AvgProb strategy yields F1-CheXbert of 0.308 and BERTScore of 0.858 at  $N=128$ , see App. J, Fig. 14). Compared to non-weighted AvgProb, W-Avg provides a substantial +5.0 point improvement in F1-CheXbert (0.357 vs 0.308) and a +0.001 point improvement in BERTScore (0.8594 vs 0.8584) at  $N=128$ . This demonstrates PRM scores, especially in a weighted BoN framework considering clinical finding profiles (CheXbert labels), effectively select higher-quality, factually accurate reports. Further details about this experiment is available in Appendix J.

#### 4.5 Ablation studies

We removed individual context sections (“INDICATION:”, “TECHNIQUE:”, “COMPARISON:”) from the PRM input at inference time (Table 4).

Removing TECHNIQUE produces the largest degradation for both verifiers (e.g. Qwen2.5-0.5B: AUROC decreases of 0.007; Qwen2.5-3B: AUROC decreases of 0.018). Eliminating INDICATION has a comparable negative impact. The effect of COMPARISON is size-dependent: for the 0.5 B model its removal is mildly harmful ( $\Delta MCC = -0.012$ ), whereas the 3B model actually improves across every metric (e.g. Accuracy + 0.004, MCC + 0.007, AUPRC + 0.001). A plausible explanation is that few training reports contain a well-formed comparison section, and many of those refer to prior studies that are unavailable at inference. For the larger network, whose capacity lets it model subtler correlations, this inconsistent information may act as noise, so discarding it sharpens the verification signal.

Removing TECHNIQUE causes the largest performance drop across models and metrics (e.g., Qwen2.5-3B AUROC drops from 0.841 to 0.823). Removing INDICATION also noticeably degrades

Table 4: Ablation results (Balanced test set, 95 % CIs). Bold indicates the highest value within each model for every metric.

Model	Variant	Accuracy	F1 Macro	MCC	AUROC	AUPRC
Qwen2.5-0.5B	Original	<b>0.752</b> (0.728, 0.773)	<b>0.751</b> (0.727, 0.771)	<b>0.502</b> (0.455, 0.544)	<b>0.834</b> (0.815, 0.853)	<b>0.832</b> (0.806, 0.856)
	Ablate INDICATION	0.743 (0.720, 0.767)	0.741 (0.718, 0.765)	0.483 (0.438, 0.531)	0.823 (0.802, 0.844)	0.813 (0.783, 0.841)
	Ablate TECHNIQUE	0.737 (0.715, 0.760)	0.734 (0.712, 0.757)	0.473 (0.428, 0.517)	0.827 (0.806, 0.846)	0.814 (0.783, 0.842)
	Ablate COMPARISON	0.746 (0.724, 0.768)	0.745 (0.722, 0.767)	0.490 (0.446, 0.535)	0.833 (0.812, 0.852)	0.827 (0.801, 0.851)
Qwen2.5-3B	Original	0.746 (0.723, 0.771)	0.744 (0.722, 0.769)	0.490 (0.444, 0.539)	<b>0.841</b> (0.821, 0.861)	0.835 (0.811, 0.861)
	Ablate INDICATION	0.743 (0.721, 0.765)	0.739 (0.717, 0.762)	0.484 (0.440, 0.530)	0.825 (0.805, 0.846)	0.810 (0.779, 0.841)
	Ablate TECHNIQUE	0.739 (0.716, 0.760)	0.737 (0.714, 0.759)	0.476 (0.429, 0.520)	0.823 (0.802, 0.843)	0.806 (0.777, 0.835)
	Ablate COMPARISON	<b>0.750</b> (0.728, 0.771)	<b>0.748</b> (0.726, 0.770)	<b>0.497</b> (0.454, 0.541)	<b>0.841</b> (0.821, 0.861)	<b>0.836</b> (0.809, 0.861)

performance. Ablating COMPARISON has minimal impact, suggesting the PRM primarily verifies consistency with the current study context and generator text, and less so with prior studies.

## 5 Discussion

**Verification accuracy.** On the MIMIC–CXR test set our smallest verifier (Qwen2.5-0.5B) achieved MCC 0.502 (95% CI: 0.455–0.544) and AUROC 0.834 (95% CI: 0.815–0.853), exceeding the strongest baseline (ReXTrust: MCC 0.467, AUROC 0.819) by +0.035 and +0.015, respectively (Table 1). Interestingly, while the 0.5B model led on threshold-dependent metrics like MCC, the larger 3B model attained the best AUROC/AUPRC. This could indicate that the smaller model better fits the specific classification threshold implied by the binary RadNLI labels, while the larger model learns a more robust underlying ranking of sentence correctness, reflected in the threshold-agnostic metrics.

**Generalisation to unseen generators.** When applied to CheXagent-8B with reports out-of-distribution relative to training, the PRM retained balanced discrimination (MCC 0.306, AUROC 0.754). In contrast, ReXTrust degraded to chance level (MCC 0.000). The drop is consistent with the dependency of white-box methods on generator-specific latent representations, which may not preserve class-separable structure after a distribution shift.

**Impact on downstream report quality.** Rejecting the lowest-scoring 10% of MAIRA-2 outputs according to the PRM’s mean probability increases F1-CheXbert from 0.329 to 0.344 (+4.5%, relative). Within a Best-of-128 sampling protocol, a weighted selection guided by PRM scores yields F1-CheXbert 0.357, a substantial relative improvement over standard log-probability ranking and non-weighted PRM selection, indicating tangible clinical benefit (Section 4.4, Appendix I, Appendix J).

**Context ablations.** TECHNIQUE and INDICATION are indispensable as removing either lowers AUROC by up to 0.018 and MCC by up to 0.029 (Table 4). In contrast, COMPARISON has asymmetric effects: for the 0.5B verifier its removal mildly degrades AUROC ( $\Delta$ AUROC = −0.001), whereas the 3B model actually improves across metrics (e.g.  $\Delta$ MCC = +0.007). We hypothesise that the comparison section contributes noise because (i) only about 50.6% of training reports include a substantive comparison (with similar rates in validation (50.0%) and test (49.1%)), and (ii) these sentences often reference external priors non related to the current

study. The larger model’s greater capacity thus allows it to exploit the cleaner context once this noisy section is removed.

**Limitations.** Our work has several limitations. First, the sentence correctness labels used for training were generated automatically by RadNLI and are therefore noisy. Using cleaner supervision, such as human annotations, could improve verifier performance. Second, our evaluation focused on the MIMIC-CXR dataset and reports from two specific generators. Testing on additional datasets like CheXpert or PadChest and with other generators is necessary to establish broader external validity. Lastly, to annotate the dataset we considered only entailment as a measure of correctness, meaning that correct statements generated by the LVLM that are not present in the ground truth are ultimately labeled as hallucinations. Leveraging LLMs to annotate or human evaluators could mitigate this issue.

**Future work.** Future research could investigate how to calibrate the PRM’s output probabilities, which could lead to more reliable thresholds for decision-making. Another avenue is integrating the PRM as a reward signal within a reinforcement learning framework to directly improve the factuality of generated reports during decoding. Furthermore, the verification approach could be extended to handle the impression section of radiology reports, which often involves synthesizing findings and clinical reasoning. Finally, incorporating the images of the studies directly as context for the PRM is another lead to further improve performance.

## 6 Conclusion

Our experiments demonstrate that a lightweight, context-aware PRM provides a robust detector of sentence-level hallucinations in LVLM-generated radiology reports, outperforming grey- and white-box baselines and maintaining discrimination under generator shift. The verifier not only improves clinical metrics through rejection and sampling strategies but also offers a model-agnostic safety layer that can be attached to any report-generation pipeline.

## Code availability

Code and model weights can be found at [alothomas/radiology\\_prm\\_verifier](https://alothomas/radiology_prm_verifier)

## References

- Amos Azaria and Tom M Mitchell. Internal state analysis for hallucination detection in large language models. *arXiv preprint arXiv:2310.12345*, 2023. URL <https://arxiv.org/abs/2310.12345>.
- Shruthi Bannur, Kenza Bouzid, Daniel Coelho de Castro, Anton Schwaighofer, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, Felix Meissen, Mercy Ranjit, Shaury Srivastav, Julia Gong, Fabian Flack, Ozan Oktay, Anja Thieme, Matthew P Lungren, Maria Teodora Wetscherek, Javier Alvarez-Valle, and Stephanie Hyland. Maira-2: Grounded radiology report generation. Technical Report MSR-TR-2024-18, Microsoft, June 2024. URL <https://www.microsoft.com/en-us/research/publication/maira-2-grounded-radiology-report-generation/>.
- Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. Inside: Llms’ internal states retain the power of hallucination detection. In *International Conference on Learning Representations (ICLR)*. ICLR, OpenReview, 2024a. URL <https://arxiv.org/abs/2402.03744>.

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

Jean-Benoit Delbrouck, Pierre Chambon, Zhihong Chen, Maya Varma, Andrew Johnston, Louis Blankemeier, Dave Van Veen, Tan Bui, Steven Truong, and Curtis Langlotz. Radgraph-xl: A large-scale expert-annotated dataset for entity and relation extraction from radiology reports. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12902–12915, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.765. URL <https://aclanthology.org/2024.findings-acl.765>.

Sebastian Farquhar, Lukas Schott, Jannik Kossen, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 618:492–497, 2024. doi: 10.1038/s41586-024-07421-0. URL <https://www.nature.com/articles/s41586-024-07421-0>.

Romain Hardy, Sung Eun Kim, and Pranav Rajpurkar. Rextrust: A model for fine-grained hallucination detection in ai-generated radiology reports. *arXiv preprint*, 2024. URL <https://arxiv.org/pdf/2412.15264v2.pdf>.

Ziwei Ji, Delong Chen, Etsuko Ishii, Samuel Cahyawijaya, Yejin Bang, Bryan Wilie, and Pascale Fung. Llm internal states reveal hallucination risk faced with a query. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 88–104. Association for Computational Linguistics, November 2024. URL <https://aclanthology.org/2024.blackboxnlp-1.6.pdf>.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-Ying Deng, Roger G. Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317, 2019. doi: 10.1038/s41597-019-0322-0. URL <https://doi.org/10.1038/s41597-019-0322-0>.

Yubin Kim, Hyewon Jeong, Shan Chen, Shuyue Stella Li, Mingyu Lu, Kumail Alhamoud, Jimin Mun, Cristina Grau, Minseok Jung, Rodrigo Gameiro, Lizhou Fan, Eugene Park, Tristan Lin, Joonsik Yoon, Wonjin Yoon, Maarten Sap, Yulia Tsvetkov, Paul Liang, Xuhai Xu, Xin Liu, Daniel McDuff, Hyeonhoon Lee, Hae Won Park, Samir Tulebaev, and Cynthia Breazeal. Medical hallucination in foundation models and their impact on healthcare. *medRxiv*, 2025. doi: 10.1101/2025.02.28.25323115. URL <https://www.medrxiv.org/content/early/2025/03/03/2025.02.28.25323115>.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Nayeon Lee, Wei Ping, Peng Xu, Mostafa Patwary, Pascale Fung, Mohammad Shoeybi, and Bryan Catanzaro. Factuality enhanced language models for open-ended text generation, 2023. URL <https://arxiv.org/abs/2206.04624>.

Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. The dawn after the dark: An empirical study on factuality hallucination in large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10879–10899, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.586. URL <https://aclanthology.org/2024.acl-long.586>.

Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. Making large language models better reasoners with step-aware verifier. *arXiv preprint arXiv:2206.02336*, 2022. doi: 10.48550/arXiv.2206.02336. URL <https://arxiv.org/abs/2206.02336>.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let’s verify step by step. *OpenAI*, 2023. URL [https://cdn.openai.com/improving-mathematical-reasoning-with-process-supervision/Lets\\_Verify\\_Step\\_by\\_Step.pdf](https://cdn.openai.com/improving-mathematical-reasoning-with-process-supervision/Lets_Verify_Step_by_Step.pdf).

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL <https://aclanthology.org/W04-1013/>.

Litian Liu, Reza Pourreza, Sunny Panchal, Apratim Bhattacharyya, and Roland Memisevic. Enhancing hallucination detection through noise injection. *arXiv preprint arXiv:2402.12345*, 2024. URL <https://arxiv.org/abs/2402.12345>.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2017. doi: 10.48550/arXiv.1608.03983. URL <https://arxiv.org/abs/1608.03983>. ICLR 2017 conference paper.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2019. doi: 10.48550/arXiv.1711.05101. URL <https://arxiv.org/abs/1711.05101>. Published as a conference paper at ICLR 2019.

Potsawee Manakul and Mark Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2304.05633*, 2023. URL <https://arxiv.org/abs/2304.05633>.

Nhat Minh Nguyen, Andrew Baker, Clement Neo, Allen Roush, Andreas Kirsch, and Ravid Shwartz-Ziv. Turning up the heat: Min-p sampling for creative and coherent llm outputs. 2024. URL <https://arxiv.org/abs/2407.01082>.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics, 2002. doi: 10.3115/1073083.1073135. URL <https://doi.org/10.3115/1073083.1073135>.

Magdalini Paschali, Zhihong Chen, Louis Blankemeier, Maya Varma, Alaa Youssef, Christian Bluethgen, Curtis Langlotz, Sergios Gatidis, and Akshay Chaudhari. Foundation models in radiology: What, how, why, and why not. *Radiology*, Feb 2025. doi: 10.1148/radiol.240597. Deputy Editor: Linda Moy; Scientific Editor: Sarah Atzen.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL <http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf>.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL <https://arxiv.org/abs/2412.15115>.

Amrith Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, and Aviral Kumar. Rewarding progress: Scaling automated process verifiers for llm reasoning. In *International Conference on Learning Representations (ICLR)*, 2024. URL <https://openreview.net/forum?id=A6Y7AqlzLW>.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y. Ng, and Matthew P. Lungren. Chexbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using bert. *arXiv preprint arXiv:2004.09167*, 2020. doi: 10.48550/arXiv.2004.09167. URL <https://arxiv.org/abs/2004.09167>.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process- and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022. URL <https://arxiv.org/abs/2211.14275>.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. TRL: Transformer Reinforcement Learning. <https://github.com/huggingface/trl>, 2020.

Hanyin Wang, Chufan Gao, Qiping Xu, Bolun Liu, Guleid Hussein, Hariprasad Korsapati, Mohamad El Labban, Kingsley Iheasirim, Mohamed Hassan, Gokhan Anil, Brian Bartlett, and Jimeng Sun. Process-supervised reward models for verifying clinical note generation: A scalable approach guided by domain expertise. *arXiv preprint arXiv:2412.12583v2*, 2025. URL <https://arxiv.org/abs/2412.12583v2>.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierrick Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface’s transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019. URL <https://arxiv.org/abs/1910.03771>.

Hongyuan Yuan, Yifan Peng, Zhiyong Lu, and Matthew P. Lungren. Radnli: A natural language inference dataset for the radiology domain. *PhysioNet*, 2021. URL <https://physionet.org/content/radnli-report-inference/>.

Serena Zhang, Sraavya Sambara, Oishi Banerjee, Julian Acosta, L. John Fahrner, and Pranav Rajpurkar. Radflag: A black-box hallucination detection method for medical vision language models. *arXiv preprint arXiv:2411.00299*, 2024. URL <https://doi.org/10.48550/arXiv.2411.00299>.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019. doi: 10.48550/arXiv.1904.09675. URL <https://arxiv.org/abs/1904.09675>. To appear in ICLR 2020.

## A Related Work

Hallucination detection in large language models (LLMs) and Large Vision-language models (LVLMs) has attracted considerable attention given its importance in safety-critical applications. Owing to the architectural similarities between LLMs and LVLMs, many techniques are transferable across these domains ([Li et al. \(2024\)](#)). In the following, we review approaches from two complementary perspectives: (i) **black-box methods**, which require only model inputs and outputs, and (ii) **white and grey box methods**, which leverage internal model representations and sampling information.

### A.1 Black-Box Approaches

Black-box methods are appealing due to their model-agnostic nature but typically incur higher computational costs.

**Sampling-Based Methods.** Sampling-based techniques generate multiple outputs for a given input to assess consistency. For example, [Zhang et al. \(2024\)](#) introduced RadFlag in the radiology domain, where multiple sampled reports are compared via an entailment model to derive a consistency score. Similarly, entropy-based methods [Farquhar et al. \(2024\)](#) compute semantic entropy over sampled outputs to quantify uncertainty, though at the expense of increased computational demand.

**LLMs as Self-Judges.** An alternative strategy uses LLMs to evaluate their own outputs. Self-critique methods such as SelfCheckGPT [Manakul and Gales \(2023\)](#) prompt the LLM to assess consistency between an initial response and subsequent samples, with metrics like BERTScore serving as proxies for factual correctness.

**Reward Models.** Process Reward Models (PRMs) and Outcome Reward Models (ORMs) provide a distinct verification approach by leveraging pre-trained LLMs. Originally proposed for mathematical reasoning [Uesato et al. \(2022\)](#), PRMs assess correctness at each intermediate step and have demonstrated superior performance over ORMs in certain tasks [Lightman et al. \(2023\)](#). Recent work has further explored PRMs for online policy improvement [Setlur et al. \(2024\)](#) and clinical note verification [Wang et al. \(2025\)](#).

While PRMs have seen applications in general multimodal reasoning and other clinical text scenarios ([Wang et al., 2025](#)), their adaptation for fine-grained, sentence-by-sentence factuality assessment of LVLM-generated radiology reports, conditioned on structured clinical context and evolving report text, represents a novel application area which we explore.

### A.2 White-Box and Grey-Box Approaches

White-box and grey-box methods tap into internal model dynamics to detect hallucinations more directly.

**Manipulating Token Probabilities.** Grey-box approaches modify token log probabilities during generation to steer outputs towards factuality. For instance, [Lee et al. \(2023\)](#) showed that greedy decoding can reduce hallucination risk, while [Nguyen et al. \(2024\)](#) proposed min-p sampling, which adjusts the token sampling threshold based on model confidence.

**Hidden State Methods.** White-box techniques leverage latent information in hidden states or logits. Classifiers trained on hidden layer activations have been shown to predict factuality effectively [Azaria and Mitchell \(2023\); Ji et al. \(2024\)](#). In radiology, [Hardy et al. \(2024\)](#) utilized token embeddings from intermediate layers (ReXTrust) to detect hallucinated findings, and

Chen et al. (2024a) proposed EigenScore, a geometrically motivated measure that computes eigenvalues of the covariance matrix of sentence embeddings. These methods often lack cross-model generalizability, a limitation our PRM approach aims to overcome.

## B Reproducibility Statement

All datasets, pretrained weights, training scripts, and configuration files necessary to reproduce every table and figure will be released after submission. Detailed instructions cover data acquisition (MIMIC-CXR v2.0.0 via PhysioNet), label generation, hyper-parameters, and hardware requirements ( $8 \times$  NVIDIA H100 and A100).

## C Ethical Considerations Statement

MIMIC-CXR consists of de-identified images and reports collected under IRB waiver; we adhered to its data-use agreement and did not attempt re-identification. The proposed verifier reduces, but does not eliminate, the risk of hallucination. Model- and data-induced biases (e.g., demographic performance gaps) remain unmeasured and should be carefully assessed before any clinical integration.

## D Dataset creation and labeling details

**Dataset source** We used MIMIC-CXR v2.0.0 (Johnson et al., 2019) from PhysioNet, adhering to official train/validation/test splits.

**Report generation** Findings sections were generated using MAIRA-2 (Bannur et al., 2024) (default settings) and CheXagent-8B (Chen et al., 2024b) for transferability tests, using the image and clinical context (indication, technique, comparison) from MIMIC-CXR.

**Sentence labeling procedure** Generated findings were segmented into sentences. Each generated sentence  $s_{k,i}^{\text{gen}}$  was compared against all ground truth sentences  $s_{k,j}^{\text{gt}}$  using RadNLI (Yuan et al., 2021). If RadNLI predicted ‘entailment’ for any pair  $(s_{k,i}^{\text{gen}}, s_{k,j}^{\text{gt}})$ , the label was  $y_{k,i} = 1$  (correct); otherwise,  $y_{k,i} = 0$  (hallucinated). We found RadNLI effective for this domain-specific task compared to general models like GPT-4o (see Figure 4).

**Dataset balancing** Initial labeling of sentences generated by MAIRA-2 using RadNLI gives an unbalanced result across the official MIMIC-CXR splits. To mitigate potential bias during training and evaluation, we balanced the training, validation, and test sets. This was achieved by downsampling the majority class (sentences labeled as correct,  $y = 1$ ) to achieve roughly a 1:1 ratio with the minority class (hallucinated sentences,  $y = 0$ ). The final sentence counts after balancing for each split are presented in Table 5.

## E Input data preparation details

Metric	RadNLI model	GPT-4o
Accuracy	<b>0.8000</b>	0.7750
F1 macro	0.7582	<b>0.7801</b>

(a) Performance on RadNLI test dataset.

You are a highly accurate Natural Language Inference (NLI) classifier and an expert in radiology.  
 Given two sentences, determine their relationship as ‘entailment’, ‘neutral’, or ‘contradiction’.  
 Respond with only one of these three labels based on the relationship between the sentences.

Sentence 1: {sentence1}  
 Sentence 2: {sentence2}

Label:

(b) Prompt used for GPT-4o on RadNLI dataset.

Figure 4: Comparison of RadNLI (Yuan et al., 2021) and GPT-4o for entailment labeling, used for generating weak labels. (a) Performance metrics evaluated on the RadNLI test dataset. (b) Prompt used for GPT-4o evaluation on the RadNLI test dataset.

Table 5: Final sentence counts per split (train, validation, test) for MAIRA-2 (Bannur et al., 2024) generated reports on MIMIC-CXR (Johnson et al., 2019) after downsampling the majority class to achieve an approximate 1:1 ratio of correct ( $y = 1$ ) vs. hallucinated ( $y = 0$ ) sentences, based on RadNLI labels.

Split	Correct ( $y = 1$ )	Hallucinated ( $y = 0$ )
Train	96,255	106,950
Validation	780	865
Test	694	771

**ReXTrust baseline input** For sentence  $s_{k,i}^{\text{gen}}$ , final hidden states for each token were extracted from MAIRA-2’s 16th transformer layer and used as input. The hidden states size is a 4096 dimensional vector.

**Grey-box baseline input** A 13-dimensional feature vector per sentence  $s_{k,i}^{\text{gen}}$  derived from MAIRA-2: token count, and mean/std/min/max of token-level logits, probabilities, and entropies.

**PRM verifier input** A single sequence per report  $k$ : clinical context  $H_k$ , followed by generated sentences  $s_{k,i}^{\text{gen}}$  interleaved with newline separators (`\n`) and (during training) ground truth labels  $y_{k,i}$  ('1' or '0'). See Figure 5.

**Prompt:**

Provide a description of the findings in the radiology study in comparison to the prior frontal image. INDICATION: Middle-aged man with possible pneumonia. TECHNIQUE: Anteroposterior (AP) and lateral chest radiographs. COMPARISON: Not applicable.

**Completions:**

"There are patchy opacities in the right upper lung, right lower lung, and left lower lung."

"No pleural effusion or pneumothorax."

"Cardiac size is normal."

**Groundtruth Labels:**

False, False, True

Figure 5: Example PRM input structure for training, using report mimic-54422699 from the MIMIC-CXR dataset. The input includes clinical context (Indication, Technique, Comparison), MAIRA-2 generated sentences, and ground-truth labels (interleaved during training only). For inference, labels are omitted and predicted by the PRM.

## F Model training details

Trained using PyTorch ([Paszke et al., 2019](#)), Hugging Face Transformers ([Wolf et al., 2019](#)) and TRL ([von Werra et al., 2020](#)).

### F.1 PRM verifier

- **Base Models:** Qwen2.5-0.5B, Qwen2.5-3B ([Qwen et al., 2024](#)).
- **Framework:** TRL PRM Trainer.
- **Epochs:** 3.
- **Optimizer:** AdamW ([Kingma and Ba, 2014](#); [Loshchilov and Hutter, 2019](#)).
- **Learning Rate:**  $1 \times 10^{-5}$ , linear warmup/decay.
- **Batching:** Effective batch size 16 (2 per device, 8 accum steps).
- **Efficiency:** Gradient checkpointing.

- **Sequence Length:** Max 1024 tokens (prompt max 512).
- **Evaluation:** Every 50 steps on validation set.

## F.2 Baseline models

**ReXTrust** (Hardy et al., 2024)

- **Architecture:** Multi-head self-attention (proj to 1024), mean pooling, linear classifier.
- **Optimizer:** AdamW (wd=0.01).
- **Learning Rate:**  $1 \times 10^{-4}$ , Cosine Annealing ( $T_{max} = 10$ ) (Loshchilov and Hutter, 2017).
- **Batch Size:** 128.
- **Loss:** BCE.
- **Regularization:** Dropout (0.1), Early Stopping (val loss, patience=3).

**Grey-box MLP** (Inspired by Liu et al. (2024))

- **Architecture:** 2-layer MLP (13 -> 50 ReLU -> 1 logit).
- **Optimizer:** Adam.
- **Learning Rate:** Fixed  $1 \times 10^{-3}$ .
- **Epochs:** 20.
- **Batch Size:** 128.
- **Loss:** BCE.

## G Best-of-N implementation details

**Candidate generation**  $N = 128$  candidates per study using MAIRA-2 (temp=1, top-p=0.9, top-k=50)

**Scoring methods** Used for non-weighted and as base for weighted BoN (matching labels like ‘Avg Prob’, ‘Min Prob’, ‘Prod Prob’ or ‘Log Prob’ in Appendix J):

- **MinProb:** The minimum sentence probability using PRM.  $\text{Score}_k = \min_i p_{k,i}$ .
- **AvgProb:** The arithmetic mean of sentence probabilities using PRM.  $\text{Score}_k = \frac{1}{m_k} \sum_i p_{k,i}$ .
- **ProdProb:** The geometric mean of sentence probabilities using PRM.  $\text{Score}_k = (\prod_{i=1}^{m_k} p_{k,i})^{1/m_k}$ .
- **LogProb:** The sum of log-probabilities (equivalent to maximizing the product  $\prod p_{k,i}$ ).  $\text{Score}_k = \sum_i \log p_{k,i}$ .

**Weighted Best-of-N (PRM-based)** Adapted from Li et al. (2022):

1. **Grouping:** Assign candidates to groups  $G_j$  based on identical CheXbert (Smit et al., 2020) label vectors.
2. **Intra-Group Scoring:** Calculate base score  $f(S_{k,l}^{\text{gen}})$  (using MinProb, AvgProb, ProdProb, or LogProb as defined above) for each candidate  $l$ .
3. **Inter-Group Aggregation:** Sum scores within each group:  $\text{TotalScore}(G_j) = \sum_{l \in G_j} f(S_{k,l}^{\text{gen}})$ .
4. **Group Selection:** Select group  $G_{j^*}$  with max TotalScore.
5. **Final Selection:** Select candidate  $l^*$  within  $G_{j^*}$  maximizing individual score  $f(S_{k,l^*}^{\text{gen}})$ . (Note: For weighted plots labeled W-Avg, W-Min, W-Prod, W-Log, the base score  $f$  corresponds to AvgProb, MinProb, ProdProb, LogProb respectively).

## H Qualitative examples of PRM verification

Examples of Qwen2.5-0.5B-PRM predictions on MAIRA-2 generated findings. Background: **Verifier Correct**, **Verifier Incorrect**. Border: **GT Correct**, **GT Incorrect**. **Underlined**: Verifier prediction differs from GT.  $p_{\phi,k,i}$  shown.



**Probabilities:** 0.480, 0.786, 0.103, 0.082

**Report:**

The lungs are clear.

Negative for pleural effusion or pneumothorax.

Cardiomediastinal silhouette is within normal limits.

There are atherosclerotic aortic calcifications.

Figure 6: Qualitative Example 1 (mimic-56486000). Mostly correct, one false positive.



**Probs:** 0.229, 0.913, 0.941, 0.881, 0.896, 0.525, 0.058, 0.094

**Report:**

The lungs are clear.

No focal consolidation.

No effusion.

No pneumothorax.

The cardiomediastinal silhouette is normal.

No displaced rib fractures are noted.

There is no evidence of a compression deformity...

The thoracic spine demonstrates mild degenerative changes.

Figure 7: Qualitative Example 2 (mimic-58865157). Mostly correct, one false positive.



**Probs:** 0.020, 0.955, 0.922, 0.210, 0.916

**Report:**

There is elevation of the left hemidiaphragm...

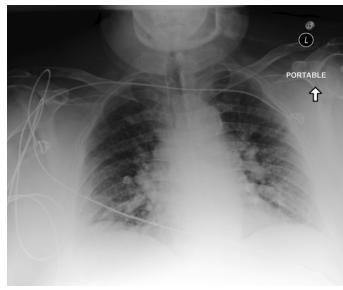
No pleural effusion.

No pneumothorax.

Normal cardiomedastinal silhouette.

There are healed left-sided rib fractures.

Figure 8: Qualitative Example 3 (mimic-51153135). All sentences correctly verified.



**Probs:** 0.146, 0.049, 0.670, 0.901, 0.616

**Report:**

Mild to moderate cardiomegaly is present.

There is mild perihilar vascular congest...

There is no focal lung consolidation.

There is no pneumothorax.

There is no large pleural effusion.

Figure 9: Qualitative Example 4 (mimic-54259835). All sentences correctly verified.



**Probs:** 0.220, 0.582, 0.810, 0.794, 0.422, 0.140, 0.206

**Report:**

The lungs are hyperinflated but clear.

No focal consolidations.

No pneumothoraces.

No pleural effusions.

Cardiomedastinal silhouette is within normal limits.

Mild degenerative changes of the spine.

Surgical clips in the left upper quadrant.

Figure 10: Qualitative Example 5 (mimic-59688743). Mixed predictions with false negatives and a false positive.

## I Additional rejection curves

This appendix section provides the full set of report rejection curves, complementing the [subsection 4.3](#) of the main paper. These plots illustrate how various report-level quality metrics change as an increasing percentage of the lowest-scoring reports are discarded. Figure 11 shows these results on the full (unbalanced) MIMIC-CXR test set across all evaluated textual and clinical metrics. Figure 12 presents results for key metrics on the balanced MIMIC-CXR test set, demonstrating the robustness of PRM-based filtering to dataset balance. Table 6 provides the precise numerical values for F1-CheXbert and F1-RadGraph scores at various rejection thresholds (0%, 5%, 10%, 15%, and 20%) on the full unbalanced test set, complementing the visual trends shown in Figure 11 and Figure 2. The table details confirm that PRM-based scoring methods lead to notable improvements in report quality as more low-scoring reports are rejected. For instance, using PRM-*avg* to reject the worst 20% of reports increases F1-CheXbert by approximately 8.8% (from a baseline of 0.32934 to 0.35838). For F1-RadGraph, improvements are generally more modest; PRM-*prod* achieves a 1.6% increase (from 0.22181 to 0.22541) at 20% rejection, while Entropy scoring shows a slightly higher 1.7% improvement (to 0.22566) at the same threshold. These results confirm the effectiveness of PRM-based methods, particularly PRM-*avg* and PRM-*prod*, in enhancing the clinical factuality of the reports with the verifier.

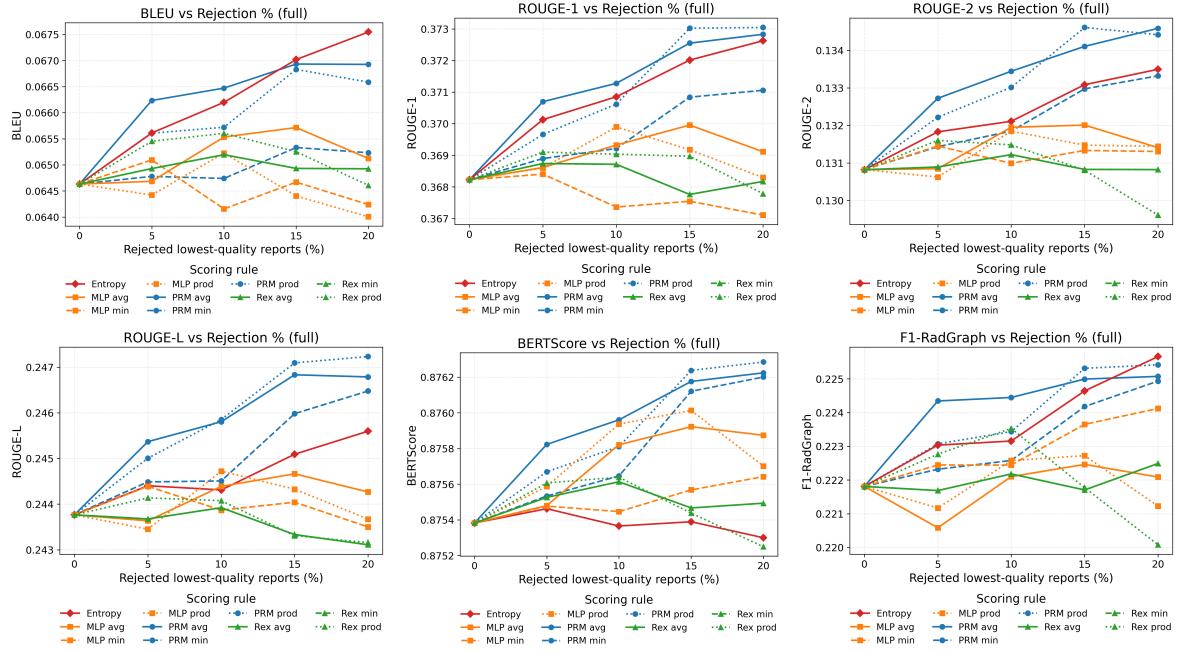


Figure 11: Full metric results for report rejection on the MIMIC-CXR full unbalanced test set (MAIRA-2 generated reports). The y-axis shows the metric score for the remaining reports after rejecting the x-axis percentage of lowest-scoring reports, based on various verifier aggregation rules. PRM-*avg* and PRM-*prod* show best improvement slopes for most metrics. Entropy as a rejection criterion also shows good improvements for BLEU and F1-RadGraph scores. (F1-CheXbert shown in main paper Fig. 2)

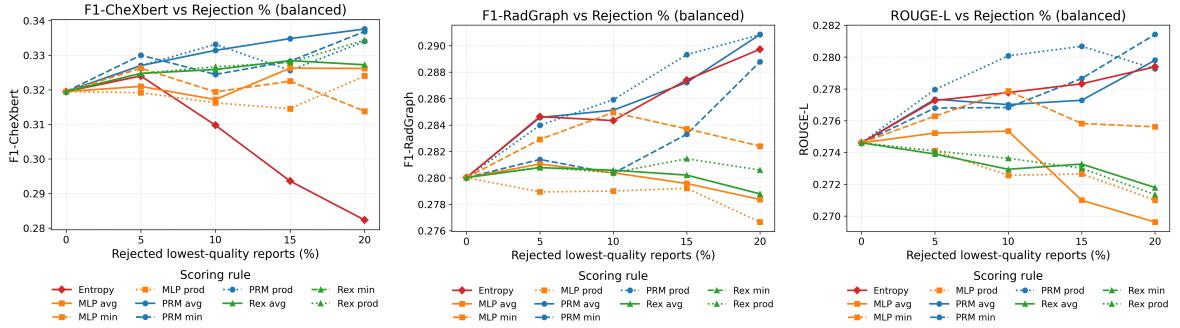


Figure 12: Selected rejection curves on the MIMIC-CXR balanced test set (MAIRA-2 generated reports). Trends mirror the unbalanced set (Figure 11), confirming PRM effectiveness is robust to dataset balance.

Table 6: Absolute F1-CheXbert and F1-RadGraph scores on the MIMIC-CXR full unbalanced test set (MAIRA-2 generated reports) after rejecting lowest-scoring reports at 0%, 5%, 10%, 15%, and 20% percentile thresholds. Scores are shown for various aggregation methods (PRM-based, MLP-based, ReXTrust-based, and Entropy). Baseline (0% rejection) F1-CheXbert is 0.32934 and F1-RadGraph is 0.22181.

Metric	Rej. %	PRM avg	PRM min	PRM prod	MLP avg	MLP prod	MLP min	Rex avg	Rex min	Rex prod	Entropy
F1-CheXbert	0%	0.32934	0.32934	0.32934	0.32934	0.32934	0.32934	0.32934	0.32934	0.32934	0.32934
	5%	0.33698	0.33679	0.33648	0.33270	0.33313	0.32648	0.32937	0.32937	0.33293	0.32757
	10%	0.34417	0.33996	0.34242	0.32990	0.33254	0.32332	0.32543	0.32543	0.33455	0.32477
	15%	0.35091	0.34008	0.35016	0.33676	0.33505	0.32588	0.32511	0.32511	0.32861	0.31992
	20%	0.35838	0.34341	0.35588	0.33896	0.33732	0.32903	0.32483	0.32483	0.32635	0.31426
F1-RadGraph	0%	0.22181	0.22181	0.22181	0.22181	0.22181	0.22181	0.22181	0.22181	0.22181	0.22181
	5%	0.22434	0.22232	0.22307	0.22059	0.22117	0.22245	0.22169	0.22169	0.22277	0.22303
	10%	0.22445	0.22258	0.22344	0.22210	0.22258	0.22244	0.22218	0.22218	0.22353	0.22316
	15%	0.22499	0.22418	0.22531	0.22246	0.22272	0.22365	0.22171	0.22171	0.22178	0.22464
	20%	0.22507	0.22493	0.22541	0.22209	0.22123	0.22412	0.22249	0.22249	0.22008	0.22566

## J Additional Best-of-N results plots

This appendix section provides a more detailed exploration of our Best-of-N (BoN) sampling experiments, complementing Section 4.4. We investigated BoN selection to assess if PRM scores could guide the selection of higher-quality reports from multiple candidates generated by MAIRA-2 (temperature 1.0,  $N_{max} = 128$  candidates per study from the MIMIC-CXR full unbalanced test set). Beyond standard BoN ranking using PRM scores directly (non-weighted), we explored weighted strategies (e.g., W-Avg). As detailed in Appendix G, 'W-Avg' refers to a weighted BoN approach where candidates are first grouped by their CheXbert label vectors; PRM-Avg scores are then aggregated within groups to select the best group, and finally, the top-scoring individual report from that group is chosen. This aims to leverage broader clinical finding profiles for more robust selection.

Our results (Figures 13 and 14) show that weighted strategies generally outperform non-weighted ones. For instance, Weighted Avg (W-Avg) using PRM scores improves F1-CheXbert by +5.0 points (0.3575 vs 0.3077 for non-weighted AvgProb) at  $N = 128$ . Similarly, W-Avg boosts BERTScore by +0.001 points (0.8594 vs 0.8584) under the same conditions. The choice of  $N$  also impacts performance, with larger  $N$  generally leads to better results.

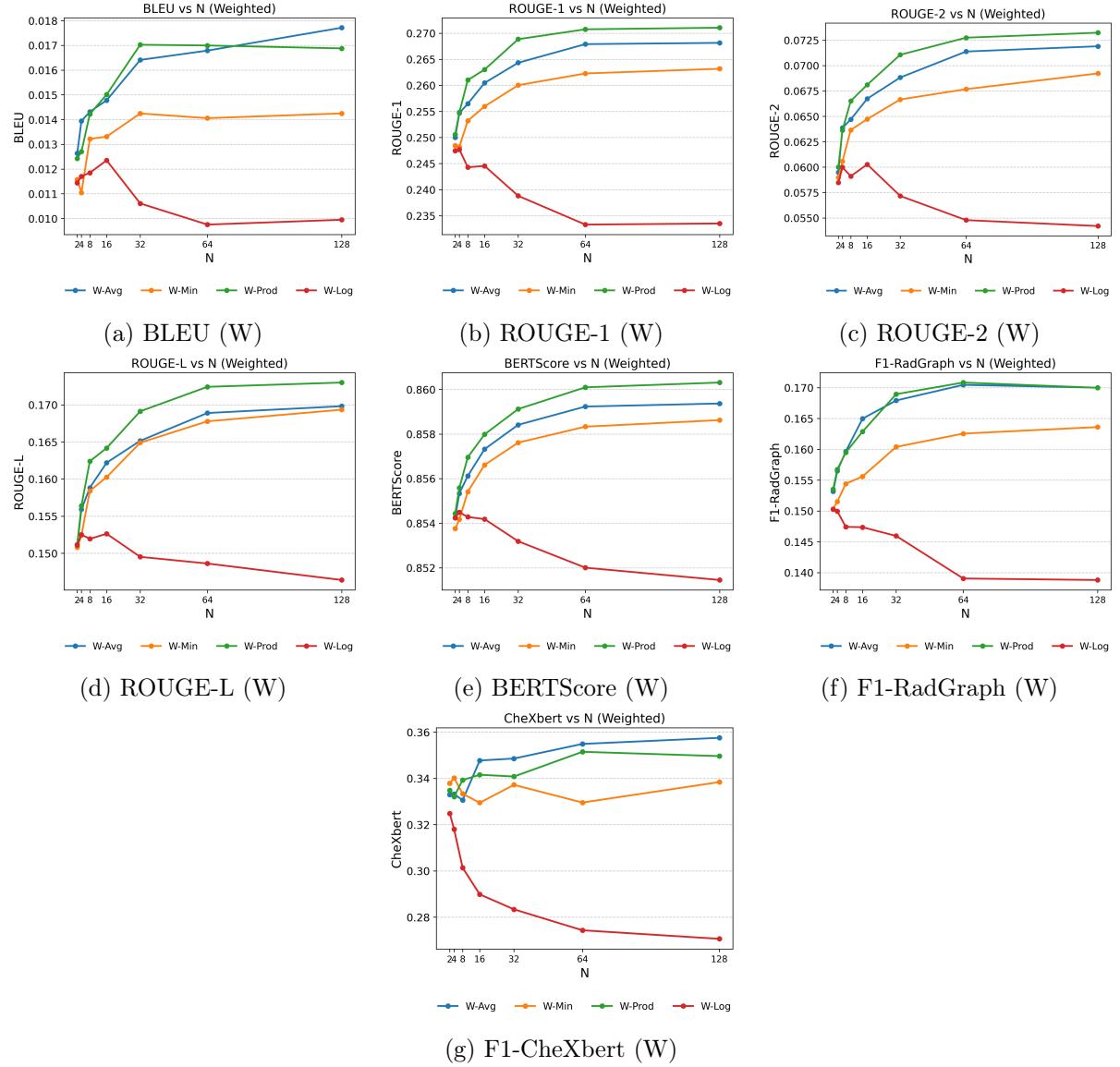


Figure 13: Performance of Weighted Best-of- $N$  (BoN) selection strategies across all metrics versus  $N$ . Candidates were generated by MAIRA-2 (temperature 1.0) for each study in the MIMIC-CXR full unbalanced test set. Scoring strategies W-Avg, W-Min, W-Prod, and W-Log refer to weighted BoN approaches using base PRM scores from AvgProb, MinProb, ProdProb, and LogProb, respectively. (F1-CheXbert and BERTScore also shown in main paper Fig. 3)

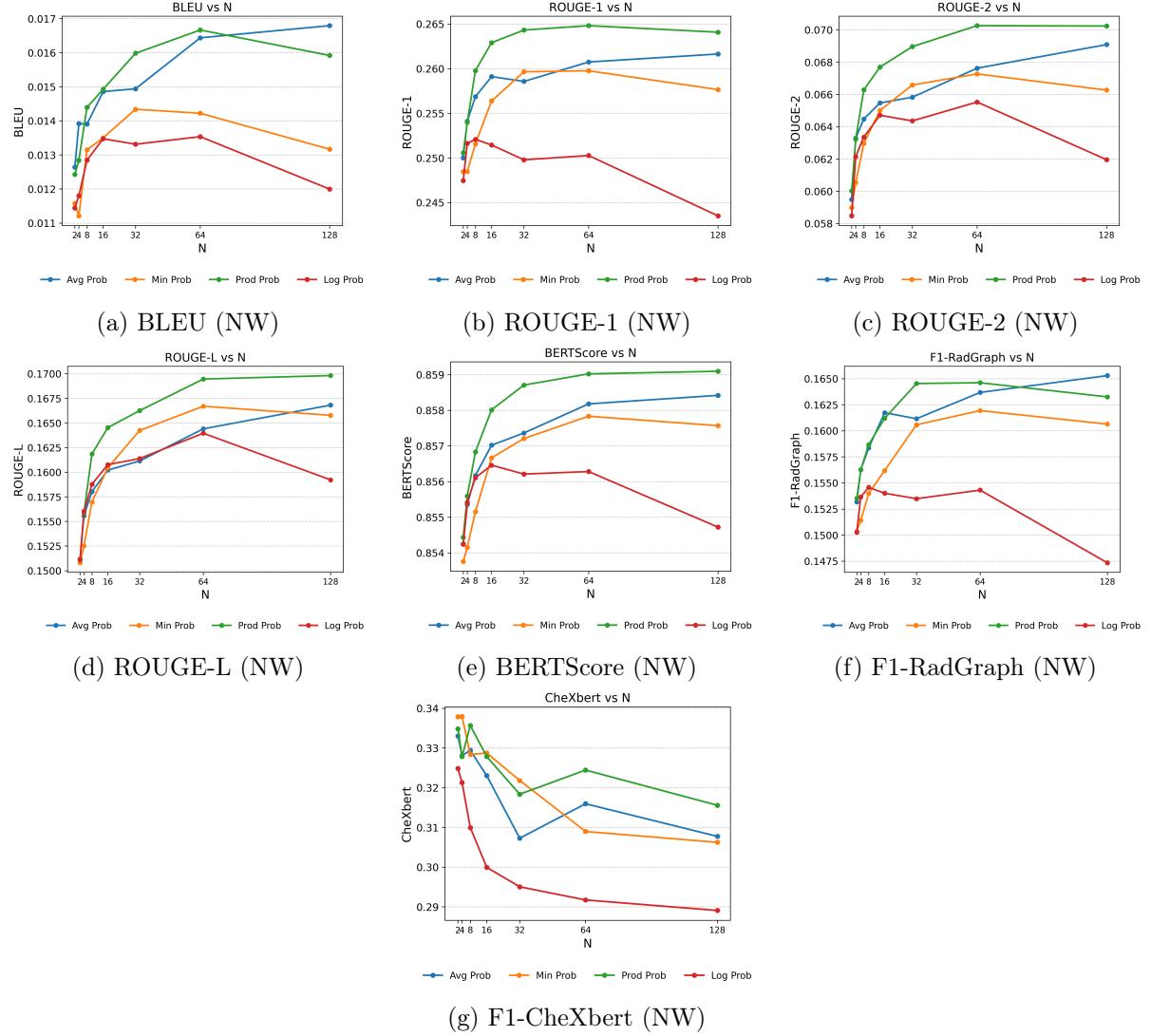


Figure 14: Performance of Non-Weighted (NW) Best-of- $N$  (BoN) selection strategies across all metrics versus  $N$ . Candidates were generated by MAIRA-2 (temperature 1.0) for each study in the MIMIC-CXR full unbalanced test set. Reports are ranked directly by PRM-derived scores (AvgProb, MinProb, ProdProb) or generator LogProb.