

POLAR: Polarity-Aware and Anatomy-Conditioned Radiology Report Evaluation

Lukas Buess¹, Tri-Thien Nguyen¹, Xiaoliang Wang², Paula Andrea Perez-Toro^{1,3},
Tomas Arias-Vergara^{1,3}, Bernhard Kainz^{2,4}, Andreas Maier¹

¹Pattern Recognition Lab ²Department of Artificial Intelligence in Biomedical Engineering

Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

³GITA Lab, Facultad de Ingenieria, Universidad de Antioquia, Medellín, Colombia

⁴Department of Computing, Imperial College London, London, UK

Correspondence: Lukas.Buess@fau.de

Abstract

Automated radiology report generation is expanding across imaging modalities and anatomical regions, yet its evaluation remains misaligned with diagnostic priorities. Existing metrics either rely on surface-level text similarity or require slow, costly generative inference. This limits their ability to detect clinically significant errors, such as hallucinated or omitted abnormal findings, and prevents their use as efficient reward signals during training. We introduce POLAR (**P**olarity-Aware Long-context **A**natomy-conditioned **R**egression), an encoder-based evaluation metric built on ModernBERT. POLAR combines long-context inputs for multi-organ reports, anatomy conditioning for targeted body-region evaluation, and polarity-aware regression that explicitly distinguishes normal from abnormal findings, yielding a more clinically meaningful assessment than prior metrics. Across real-world, radiologist-annotated CXR benchmarks and expert preference datasets, POLAR achieves competitive alignment with radiologist judgments while running $11\times$ faster than generative evaluators, making it practical for large-scale benchmarking and as a training-time reward signal. We publicly release POLAR and the POLAR-Synth dataset, along with a fully open-source, privacy-preserving pipeline for data generation, training, and inference, enabling the metric to be extended to new modalities and anatomies.

1 Introduction

Recent advances in medical vision-language models have substantially improved automated radiology report generation across imaging modalities and anatomical regions (Chen et al., 2024; Zhou et al., 2026; Sellergren et al., 2026; Pellegrini et al., 2025; Buess et al., 2026), extending from single-view chest radiographs to long-form, multi-organ reporting for volumetric imaging (Blankemeier et al., 2026; Hamamci et al., 2026). As gener-

Metric	Clin.	Abn.	Ctx-lim.	Anat.	Speed	Lbl.	Open
NLG Metrics							
BLEU (Papineni et al., 2002)	X	X	X	X	+	X	✓
ROUGE (Lin, 2004)	X	X	X	X	+	X	✓
BERTScore (Zhang* et al., 2020)	X	X	✓	X	+	X	✓
Clinically Grounded Metrics							
CheXbert (CE) (Smit et al., 2020)	✓	✓	✓	X	+	✓	✓
CheXprompt (Zambrano Chaves et al., 2025)	✓	X	X	X	-	X	X
CRG (Hamamci et al., 2025)	✓	✓	X	X	+	✓	✓
FineRadScore (Huang et al., 2024)	✓	✓	X	X	-	X	X
MedVAL (Aali et al., 2026)	✓	✓	X	X	-	X	X
RadCliQ (Yu et al., 2023a)	✓	✓	✓	X	+	X	✓
RadEvalBERTScore (Xu et al., 2025)	✓	✓	X	X	+	X	✓
RadFact (Bannur et al., 2024)	✓	X	X	X	-	X	X
RadGraph (Yu et al., 2023a)	✓	X	X	X	+	X	✓
RaTEScore (Zhao et al., 2024)	✓	✓	X	X	+	X	✓
GREEN (Ostmeier et al., 2024)	✓	X	✓	X	-	X	(✓)
CRIMSON (Baharoon et al., 2026)	✓	X	X	X	-	X	(✓)
POLAR-LM (Ours)	✓	✓	X	✓	-	X	✓
Clinically Grounded Regression Metrics							
CREPE (Cho et al., 2025)	✓	X	✓	X	+	X	(✓)
POLAR (Ours)	✓	✓	X	✓	+	X	✓

Table 1: Radiology report evaluation metrics across seven dimensions. Clin.: clinically grounded (beyond lexical/semantic overlap). Abn.: differentiates normal vs. abnormal findings in scoring. Ctx-lim.: subject to input-length truncation. Anat.: supports region-specific evaluation. Speed: + fast, – slow. Lbl.: restricted to a closed label set. Open: open-source pipeline.

ation broadens across modalities and anatomies, evaluation has emerged as a critical bottleneck.

At its core, this bottleneck reflects a mismatch between natural language generation (NLG) metrics and clinical reasoning (Buess et al., 2025; Zhang et al., 2025d). Lexical and semantic similarity metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and BERTScore (Zhang* et al., 2020) reward surface-level overlap while remaining insensitive to clinical correctness, often assigning high scores to reports that omit critical abnormalities or misrepresent findings. Recent clinically grounded evaluators improve factual sensitivity and alignment with radiologist preferences (Ostmeier et al., 2024; Bannur et al., 2024; Cho et al., 2025), but still fall short on four fronts.

As summarized in Table 1, current evaluation methods exhibit four major shortcomings: they are (1) polarity-blind, scoring abnormal findings the same as normal ones despite their different clinical

cal risk profiles (Zhang et al., 2025b; Roschewitz et al., 2026); (2) anatomically coarse, operating at the full-report level and obscuring localized errors in clinically relevant regions; (3) context-limited, with short context windows that force truncation of long, multi-organ CT and MRI reports; and (4) slow, for generative evaluators, making them impractical to scale or use as reward signals in reinforcement learning (Delbrouck et al., 2022; Zhang et al., 2025a; Wang et al., 2025). Beyond these methodological gaps, several pipelines also depend on closed-source models for training data generation (e.g. GPT-4 (Ostmeier et al., 2024)) or inference, limiting reproducibility, extensibility, and privacy-preserving use in clinical settings.

To address these limitations, we propose POLAR (**P**olarity-Aware **L**ong-context **A**natomy-conditioned **R**egression), an encoder-based evaluation metric for radiology reports across CXR, CT, and MRI. POLAR combines polarity-aware multi-head regression with anatomy conditioning, enabling targeted assessment of clinically relevant findings within specific anatomical regions. Built on a long-context encoder backbone, POLAR supports full-report evaluation without truncation while remaining efficient for large-scale benchmarking and training-time feedback. POLAR is trained on POLAR-Synth, a large-scale multi-modal dataset of $\sim 80K$ reference-candidate report pairs spanning CXR, CT, and MRI with structured polarity- and anatomy-aware error annotations, which we release with this work.

Our contributions are summarized as follows:

- **Polarity-aware scoring:** We introduce a multi-head regression architecture that explicitly distinguishes normal from abnormal findings in both errors and matches, enabling more clinically meaningful scoring.
- **Anatomy-conditioned evaluation:** We introduce anatomy conditioning for targeted, region-specific assessment of report quality, revealing which anatomical regions a model handles well or poorly.
- **Long-context evaluation:** We introduce an encoder-based evaluator with up to 2,048 input tokens, enabling full-report assessment of long CT and MRI reports without truncation.
- **Open-source dataset and pipeline:** We release POLAR-Synth, a large-scale multi-modal corpus of $\sim 80K$ reference-candidate

report pairs across CXR, CT, and MRI with polarity- and anatomy-aware error annotations, alongside an open-source pipeline for dataset generation, training, and inference.

2 Related Work

Natural language generation metrics. NLG metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005) have long been used to evaluate radiology report generation due to their simplicity and efficiency. Semantic metrics such as BERTScore (Zhang* et al., 2020) improve robustness to paraphrasing by leveraging contextual embeddings. However, these metrics remain largely insensitive to clinical factuality and often fail to capture diagnostically important errors (Ostmeier et al., 2024).

Clinically grounded metrics. To improve clinical relevance, many radiology evaluation methods move beyond surface-level overlap by mapping reports to structured clinical representations such as finding labels or entity-relation graphs (Smit et al., 2020; Hamamci et al., 2025; Zhao et al., 2024), better capturing clinically meaningful content than lexical metrics. Some methods are tied to pre-defined label spaces and miss errors outside their scope (Smit et al., 2020; Hamamci et al., 2025); others rely on slow LLM-based extraction (Ostmeier et al., 2024; Aali et al., 2026; Huang et al., 2024). Many operate on short contexts, produce only global report-level scores, and do not account for the different clinical importance of normal and abnormal findings (Zhang et al., 2025b). RadEvalBERTScore (Xu et al., 2025) relaxes the context limit with a long-context encoder, but remains similarity-based without structured error, polarity, or anatomy modeling.

LLM-based evaluation. Recent work uses LLMs as evaluators that reason about discrepancies between reference and candidate reports. Methods such as GREEN (Ostmeier et al., 2024), CRIMSON (Baharoon et al., 2026), and RadFact (Bannur et al., 2024) align well with radiologist preferences and provide interpretable error explanations, but are substantially slower and more expensive, limiting their use for large-scale benchmarking or as reward signals during training. Several pipelines also rely on closed-source models for synthetic data generation (Ostmeier et al., 2024; Cho et al., 2025; Baharoon et al., 2026) or inference (Bannur et al., 2024), and the training corpora are typically not

released. This limits transparency, reproducibility, and privacy-preserving deployment in clinical settings, since reproducing or extending the metrics requires uploading patient data to external APIs.

Encoder-based regression metrics. Encoder-only regression approaches predict structured error counts directly from report pairs, offering a faster alternative to generative evaluation (Calamida et al., 2023). CREPE (Cho et al., 2025) shows this paradigm can achieve competitive clinical alignment while remaining efficient. However, these methods typically inherit the 512-token limits of legacy BERT-style encoders, often insufficient for long CT or MRI reports and forcing truncation of clinically relevant findings. POLAR extends this paradigm along three axes: long-context encoding for full multi-organ reports, polarity-aware regression separating normal from abnormal findings, and anatomy conditioning for localized assessment.

3 POLAR-Synth Dataset

POLAR is built in two stages: we first construct POLAR-Synth, a large-scale corpus of reference-candidate report pairs with structured polarity- and anatomy-aware error annotations (this section), and then train an encoder-based regressor on these annotations (Section 4). Each POLAR-Synth example consists of a radiologist-written reference, a candidate report, and clinical discrepancy annotations distilled from a locally hosted GPT-OSS-120B (Agarwal et al., 2025), yielding targets that can be learned by an encoder-based regressor.

3.1 Multi-Modal Data Sources

POLAR-Synth is constructed from report pairs, each consisting of a reference report R_{ref} (ground truth) and a candidate report R_{cand} (prediction), spanning multiple imaging modalities, anatomical regions, and report styles. We include four source datasets to capture variation in report length, anatomical coverage, and clinical complexity:

- **GREEN-CXR** (Ostmeier et al., 2024): Chest X-ray (CXR) reports collected from six public datasets with relatively short reports.
- **CT-RATE** (Hamamci et al., 2026): Chest CT reports with longer descriptions, richer clinical detail, and more anatomical complexity.
- **Merlin** (Blankemeier et al., 2026): Structured multi-organ abdominal CT reports that ex-

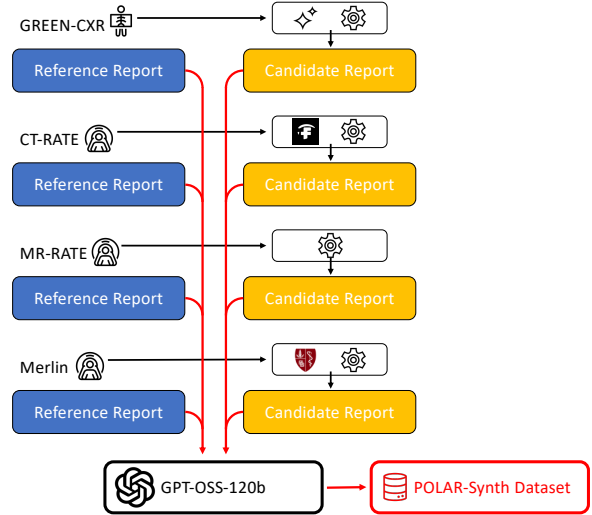


Figure 1: POLAR-Synth dataset pipeline: : chest x-ray, : CT/MRI, : chest x-ray report generation models, : synthetic generation of candidate reports, : CT-CHAT model, : Merlin model.

tend beyond the thoracic domain and provide broader anatomical coverage.

- **MR-RATE** (Forithmus, 2026): Brain and spine MRI reports for modality diversity.

3.2 Candidate Synthesis

For each source dataset, we pair R_{ref} with a candidate R_{cand} from five heuristics covering realistic predictions and controlled perturbations (Figure 1): (1) report generation with state-of-the-art report generation models, (2) LLM-based synthetic error insertion, (3) random report reassignment from a different patient, (4) meaning-preserving rewriting, and (5) sentence-level dropping or shuffling. This mixture follows prior synthetic-supervision pipelines (Ostmeier et al., 2024; Cho et al., 2025; Baharoon et al., 2026) and covers hallucination, paraphrase, and omission errors. Heuristics (1)–(2) each contribute 25% (not MR-RATE, which lacks a generation model) and (3)–(5) each contribute 16.7%; see Appendix A for details and prompts.

3.3 Polarity-Aware Error Analysis

Given a report pair $(R_{\text{ref}}, R_{\text{cand}})$, GPT-OSS produces a structured comparison with clinically significant errors, insignificant errors, and matched findings (Figure 2). The analysis covers eight error categories extending GREEN’s taxonomy (Ostmeier et al., 2024) with two attribute-level types for quantitative and descriptive errors common in CT

and MRI reports: (a) *false report*, (b) *missing finding*, (c) *incorrect location*, (d) *incorrect severity*, (e) *spurious comparison*, (f) *missing comparison*, (g) *incorrect quantitative detail*, and (h) *incorrect descriptive attribute*. Significant vs. insignificant errors are distinguished by whether the discrepancy involves an abnormal or normal finding (see Appendix C). Matched findings are likewise split into normal and abnormal findings. These counts form POLAR’s regression targets. We treat GPT-OSS-120B as a teacher rather than ground truth. Its annotations are validated against a human-annotated CXR benchmark (Section 6.1), Appendix C shows a representative output, and POLAR’s alignment with expert radiologists across seven independent benchmarks (Section 6) further supports the clinical informativeness of this supervision.

POLAR Summary	
[Explanation]:	<Explanation here>
[Clinically Significant Errors]:	
(a) False report of a finding in the candidate	
(b) Missing a finding present in the reference	
(c) Misidentification of a finding’s anatomic location or position...	
(d) Misassessment of the severity or extent of a finding	
(e) Mentioning a comparison that isn’t in the reference	
(f) Omitting a comparison detailing a change from a prior study	
(g) Misstatement of quantitative details such as size or number...	
(h) Misstatement of descriptive attributes (e.g., morphology, ...)	
[Clinically Insignificant Errors]:	
(a)–(h)	
[Matched Findings]:	
– Normal findings	
– Abnormal findings	

Figure 2: POLAR error analysis schema: eight error categories and two polarity-aware matched findings.

3.4 Anatomy Labeling

Beyond global report-level supervision, POLAR-Synth assigns each extracted error and matched finding to an anatomical region, enabling localized supervision and anatomy-conditioned evaluation. GPT-OSS assigns each error and matched finding exactly one label from a predefined anatomy vocabulary, with Other when no specific anatomy applies (Figure 3). We use dataset-specific anatomy vocabularies (region taxonomies for structuring reports per modality): (Moll et al., 2025) for CXR, RadGenome (Zhang et al., 2025c) for CT-RATE, a brain/spine split for MR-RATE, and the original work (Blankemeier et al., 2026) for Merlin. Full per-dataset vocabularies are in Appendix B.

Anatomy Labeling Output	
...	
[Errors]:	
(b) Missing a finding: 1. “Right pleural effusion”	<a>Pleura;
[Matched Findings]:	
– Normal findings: 1. “No hydronephrosis”	<a>Kidney
– Abnormal findings: 1. “Pancreatic duct dilatation”	<a>Pancreas

Figure 3: Example anatomy-labeled output.

Source dataset	Modality	Full Reports	Anatomy Reports	# Anatomies	Tokens (mean / P99)
GREEN (CXR)	CXR	23,204	46,177	7	56 / 161
CT-RATE	CT	22,196	44,366	10	268 / 555
MR-RATE	MRI	17,155	33,285	2	247 / 528
Merlin	CT	18,071	27,539	13	377 / 825
POLAR-Synth	CXR, CT, MRI	80,626	151,367	31	-

Table 2: POLAR-Synth source datasets: full-report pairs, anatomy-conditioned examples, anatomy labels, and report length in ModernBERT tokens (mean / P99). The POLAR-Synth anatomy count (31) deduplicates one anatomy that appears in two source vocabularies.

3.5 Conditioning Triplet Assembly

From GPT-OSS annotations, we construct training triplets $(R_{\text{ref}}, R_{\text{cand}}, C)$, where C is the anatomy conditioning. We use three complementary types:

- **Full-report:** C = “full report”, with targets summing error counts across all anatomies.
- **Positive anatomy:** an anatomy present in R_{ref} (e.g., C = “heart”), with anatomy-specific error or matched findings targets > 0 .
- **Negative anatomy:** an anatomy with no associated errors or matched findings, with the target vector set to zero to teach the model to suppress unrelated errors.

Each report pair provides global and anatomy-specific supervision, with positives and negatives drawn from each dataset’s vocabulary. Per-anatomy counts are in Appendix I. Table 2 summarizes POLAR-Synth.

4 POLAR

POLAR is an encoder-based metric for radiology report evaluation that predicts polarity-aware error counts conditioned on anatomical context. Given a reference report R_{ref} , a candidate report R_{cand} , and an anatomy condition C , POLAR predicts an error count vector covering eight error categories and two matched-finding polarities (see Figure 4):

$$f(R_{\text{ref}}, R_{\text{cand}}, C) = \hat{y} \in \mathbb{R}^{10}. \quad (1)$$

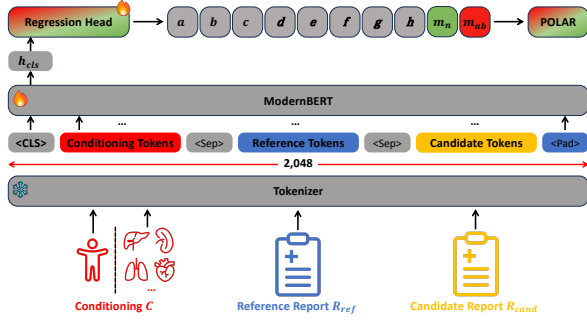


Figure 4: POLAR encodes reference, candidate, and anatomy condition to predict a 10-dimensional vector of polarity-aware error and matched-finding counts. Heads (a)–(h) cover the eight significant error categories; m_n and m_{ab} count matched normal and abnormal findings.

4.1 Input Formulation

The input sequence is constructed as

$$X = [\text{CLS}] \oplus C \oplus [\text{SEP}] \oplus R_{\text{ref}} \oplus [\text{SEP}] \oplus R_{\text{cand}} \quad (2)$$

where $\oplus$ denotes concatenation and C is a textual anatomy condition (e.g., “heart”, “liver”, “full report”). POLAR encodes X with BioClinical ModernBERT (Sounack et al., 2025), using its 2,048-token context to cover all source datasets without truncation, unlike conventional 512-token encoder-based evaluators (Table 2). The final [CLS] state $h_{\text{cls}} \in \mathbb{R}^d$ serves as a global representation of the discrepancy between R_{ref} and R_{cand} under C .

4.2 Polarity-Aware Multi-Head Regression

POLAR predicts a structured clinical error profile through ten regression heads. Eight cover the clinically significant error categories (a)–(h); the corresponding insignificant errors are excluded from supervision, as they concern normal findings without clinical impact (see Section 3.3) (Ostmeier et al., 2024). Two heads predict matched findings, split into normal (m_n) and abnormal (m_{ab}) groups, providing polarity awareness in the supervision signal.

The model outputs the 10-dimensional vector

$$\hat{y} = (\hat{y}_a, \hat{y}_b, \dots, \hat{y}_h, \hat{y}_{m_n}, \hat{y}_{m_{ab}}) \in \mathbb{R}^{10} \quad (3)$$

where each component is an error or match count conditioned on C . Heads are trained jointly with MSE loss against the GPT-OSS targets, complemented by auxiliary binary presence heads (binary cross-entropy, discarded at inference) following Cho et al. (2025). This lets POLAR distinguish clinically different types of agreement and error while retaining encoder-only efficiency.

4.3 Inference and Scoring

At inference, given an input triplet $(R_{\text{ref}}, R_{\text{cand}}, C)$, POLAR produces the prediction vector $\hat{y}$ in a single forward pass. Each head is rounded to the nearest non-negative integer before aggregation.

The POLAR score combines credit for matched findings with penalties for hallucinations and omissions, normalized by the reference’s clinical content. Let m_{ab} and m_n denote matched abnormal and normal findings, e_{fp} the false positives (head a), e_{fn} missing findings (head b), and

$$e_{\text{attr}} = \sum_{k \in \{c, d, e, f, g, h\}} \hat{y}_k \quad (4)$$

the total clinically significant attribute errors. The weighted credit and penalty terms are

$$C = m_{ab} w_{ab} + m_n w_n, \quad (5)$$

$$E = e_{fp} w_{fp} + e_{fn} w_{fn} + e_{\text{attr}} w_{\text{attr}}, \quad (6)$$

$$W_{\text{ref}} = (m_{ab} + e_{fn}) w_{ab}, \quad (7)$$

$$W_n = m_n w_n, \quad (8)$$

and the POLAR score is

$$s = \frac{C - E}{W_{\text{ref}} + W_n + E} \in [-1, 1], \quad (9)$$

with $s = 1$ for a zero denominator. We set $w_{ab} = w_{fp} = w_{fn} = 2$, $w_n = 0.5$, and $w_{\text{attr}} = 1$ to reflect radiologist-informed clinical priorities, with the scheme reviewed and approved by a board-certified radiologist. Abnormal findings, hallucinations, and omissions receive the highest weight as the most clinically consequential; attribute errors receive intermediate weight; and normal findings receive lower weight because they are often templated and less central to clinical assessment. Unlike polarity-blind metrics, this avoids length-bias artifacts when reference and candidate differ in their balance of normal and abnormal content (Roschewitz et al., 2026). Because e_{fn} enters both E and W_{ref} , omissions are penalized more than hallucinations, reflecting their higher clinical cost. Robustness to weight perturbations is evaluated in Appendix F. $C = \text{“full report”}$ yields a report-level score; anatomy strings yield localized scores. See Appendix E for score behavior.

5 Experimental Setup

5.1 Evaluation Datasets and Metrics

We evaluate POLAR on one synthetic benchmark (POLAR-Synth) and seven human-annotated

benchmarks, predominantly in CXR, covering distillation fidelity, anatomy-conditioned localization, fine-grained error counting, and expert preference alignment. Human evaluation is therefore primarily restricted to CXR, while CT and MRI experiments are limited to distillation fidelity on synthetic data, and anatomy-conditioned evaluation is performed only on synthetic data across all modalities.

Distillation fidelity. On a held-out test split of POLAR-Synth (80/10/10 train/val/test), we report MAE against GPT-OSS annotations, measuring how faithfully the encoder-based regressor reproduces the supervision signal.

Anatomy-conditioned localization. On the anatomy-conditioned split of POLAR-Synth, each test triplet is positive or negative depending on whether the conditioned anatomy has non-zero ground-truth counts. We report suppression rate (zero prediction on negatives), detection rate (non-zero on positives), and macro-averaged binary F1, for significant errors and matched findings.

Fine-grained error counting. On the radiologist-annotated benchmarks ReXVal (Yu et al., 2023b; Pollard et al., 2026) (6 radiologists), RadEval (Xu et al., 2025) (2 radiologists), and RaTE-Eval Task 1 (Zhao et al., 2024) (2 radiologists), experts assign per-category error counts to candidate-reference pairs. We report MAE and Kendall’s τ against expert counts.

Expert preference alignment. On RadPref (Baharoon et al., 2026) (3 radiologists), GreenPref (Ostmeier et al., 2024) (2 radiologists), and RaTE-Eval Task 2 (Zhao et al., 2024) (2 radiologists), radiologists rank competing candidates. We report rank correlation and preference accuracy.

Clinical stress testing. RadJudge (Baharoon et al., 2026) probes ten clinical criteria (e.g., normal-finding handling, paraphrase robustness) across 30 cases, scoring whether each metric’s ranking matches expert expectations.

5.2 Baselines

We compare POLAR against the RadEval metric collection (Xu et al., 2025), NLG metrics (e.g., BLEU (Papineni et al., 2002), BERTScore (Zhang et al., 2020)), clinically grounded metrics (e.g., CheXbert (Smit et al., 2020), RaTEScore (Zhao et al., 2024)), generative LLM-based evaluators (e.g., GREEN (Ostmeier et al., 2024), CRIMSON (Baharoon et al., 2026)), and the regression-based CREPE (Cho et al., 2025).

POLAR-LM. To compare regression and gener-

Evaluator	ReXVal Kendall $\tau \uparrow$
GPT-OSS-120B	0.78
GPT-4	0.79
Six-expert inter-reader ceiling	0.81

Table 3: Teacher validation against ReXVal expert total error counts (significant + insignificant). Counts are restricted to the six error categories ReXVal annotates. GPT-4 values follow the GREEN evaluation.

ative evaluation under matched supervision, we train POLAR-LM, a generative variant built on MedGemma 1.5 (Sjellergren et al., 2026) on the raw GPT-OSS-120B outputs from POLAR-Synth (rather than the derived regression counts). This places POLAR-LM in the same family of generative evaluators as GREEN (Ostmeier et al., 2024) and CRIMSON (Baharoon et al., 2026).

5.3 Implementation Details

POLAR uses a BioClinical ModernBERT-Large¹ backbone, a biomedical variant of ModernBERT. POLAR-LM uses MedGemma 1.5², finetuned with TRL SFT (von Werra et al., 2020). Both models are trained on the full POLAR-Synth corpus, yielding a unified model across CXR, CT, and MRI. We train POLAR on one NVIDIA H100 (94 GB) and POLAR-LM on four. Runtime is measured on a single A40 (48 GB). Hyperparameters, optimization, and runtime evaluation details are in Appendix D.

6 Results and Discussion

6.1 Teacher Validation

Because POLAR-Synth is distilled from GPT-OSS-120B, we separately validate the teacher against independent radiologist annotations on ReXVal. Following GREEN’s protocol (Ostmeier et al., 2024), we evaluate GPT-OSS-120B against expert total error counts and compare with the GPT-4 teacher result reported by GREEN. GPT-OSS-120B achieves Kendall’s $\tau = 0.78$, versus 0.79 for GPT-4 and a six-radiologist inter-reader ceiling of 0.81 (Table 3). This supports the clinical validity of the open-source teacher.

6.2 Error Count Evaluation

Table 4 reports distillation fidelity on the POLAR-Synth test split, measured as MAE against GPT-OSS targets. These results quantify how faith-

¹thomas-sounack/Bio-ModernBERT-large

²google/medgemma-1.5-4b-it

Encoder Type	Total	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	Matched Total (Norm./Abn.)
CXR (avg.)	2.92±2.94	1.02±1.49	1.15±1.59	0.22±0.47	0.23±0.51	0.07±0.29	0.11±0.45	0.02±0.15	0.10±0.34	2.62±2.14 (1.27±1.60/3.44±1.49)
BioClinical ModernBERT (Sounack et al., 2025)	0.78 ±1.17	0.33 ±0.60	0.27 ±0.55	0.11 ±0.33	0.15 ±0.39	0.08 ±0.30	0.10 ±0.39	0.02 ±0.13	0.10 ±0.32	0.48 ±0.72 (0.33±0.65/0.22±0.47)
ModernBERT (Warner et al., 2025)	<u>0.81</u> ±1.16	<u>0.34</u> ±0.60	<u>0.29</u> ±0.57	<u>0.12</u> ±0.35	<u>0.17</u> ±0.41	<u>0.08</u> ±0.29	0.12±0.41	0.02 ±0.14	0.10 ±0.32	0.46 ±0.69 (0.31±0.62/0.23±0.48)
Qwen (Zhang et al., 2025e)	1.08±1.30	0.61±0.78	0.48±0.73	0.18±0.43	0.21±0.46	0.08 ±0.29	0.12±0.45	<u>0.03</u> ±0.17	<u>0.11</u> ±0.35	1.30±1.34 (1.06±1.17/0.43±0.68)
Granite (Awasthy et al., 2025)	1.63±1.82	0.90±1.03	0.92±0.96	0.22±0.47	0.24±0.51	0.07 ±0.29	<u>0.11</u> ±0.45	0.02 ±0.15	0.10 ±0.34	1.27±1.14 (0.81±0.87/0.80±0.89)
Gemma (Vera et al., 2025)	2.53±2.73	0.77±1.03	0.99±1.43	0.30±0.67	0.30±0.55	0.16±0.41	0.23±0.58	0.05±0.22	0.17±0.41	3.58±2.82 (2.01±2.25/1.63±1.99)
PPLX (Eslami et al., 2026)	0.95±1.30	0.39±0.66	0.35±0.63	0.14±0.37	0.18±0.42	0.07 ±0.29	<u>0.11</u> ±0.40	0.02 ±0.13	0.10 ±0.33	1.10±1.36 (0.77±1.19/0.40±0.69)
MedGemma-1.5 (Søllergren et al., 2026)	0.82±1.24	0.34±0.63	0.31±0.63	0.14±0.37	0.18±0.44	0.09±0.32	0.11±0.43	0.02±0.14	0.12±0.34	0.44±0.73 (0.32±0.67/0.21±0.46)
CT-RATE (avg.)	3.70±4.74	1.11±1.98	1.77±3.00	0.25±0.69	0.23±0.70	0.03±0.19	0.03±0.23	0.16±0.55	0.13±0.43	11.60±4.85 (8.50±3.39/3.10±3.94)
BioClinical ModernBERT (Sounack et al., 2025)	0.99 ±1.60	0.45 ±0.83	0.50 ±1.00	0.13 ±0.38	0.19 ±0.56	0.03 ±0.19	0.03 ±0.19	<u>0.12</u> ±0.41	0.12 ±0.39	1.27 ±1.28 (1.05±1.14/0.44±0.73)
ModernBERT (Warner et al., 2025)	<u>1.02</u> ±1.67	<u>0.45</u> ±0.87	<u>0.53</u> ±1.07	<u>0.13</u> ±0.38	<u>0.19</u> ±0.56	0.03 ±0.19	0.03 ±0.17	0.11 ±0.40	0.12 ±0.39	1.22 ±1.19 (1.06±1.09/0.43±0.74)
Qwen (Zhang et al., 2025e)	1.96±2.37	1.09±1.05	1.03±0.69	0.24±0.66	0.24±0.67	0.03 ±0.19	<u>0.04</u> ±0.24	0.18±0.51	0.16±0.45	2.67±2.56 (1.76±1.56/1.23±1.88)
Granite (Awasthy et al., 2025)	2.11±2.38	1.08±1.39	1.17±1.67	0.28±0.68	0.24±0.67	0.03 ±0.19	0.05±0.25	0.16±0.54	0.14±0.43	2.09±1.83 (1.63±1.43/1.12±1.30)
Gemma (Vera et al., 2025)	1.58±1.99	<u>0.68</u> ±1.04	0.74±1.29	<u>0.19</u> ±0.47	0.26±0.59	<u>0.05</u> ±0.26	<u>0.04</u> ±0.21	0.13±0.43	0.15±0.43	5.96±3.82 (3.61±2.02/2.44±2.93)
PPLX (Eslami et al., 2026)	2.48±3.08	1.00±1.36	1.46±2.17	0.27±0.63	0.23 ±0.68	0.03 ±0.19	0.06±0.26	0.17±0.52	<u>0.13</u> ±0.43	3.83±4.43 (2.42±2.16/1.63±2.91)
MedGemma-1.5 (Søllergren et al., 2026)	1.08±1.88	0.47±0.99	0.54±1.15	0.15±0.42	0.18±0.60	0.03±0.20	0.03±0.20	0.12±0.42	0.12±0.39	1.23±1.24 (1.07±1.14/0.40±0.72)
MR-RATE (avg.)	2.07±3.31	0.88±1.66	0.62±1.56	0.20±0.63	0.14±0.55	0.04±0.21	0.02±0.14	0.09±0.39	0.09±0.45	18.23±11.27 (16.06±11.35/2.17±2.71)
BioClinical ModernBERT (Sounack et al., 2025)	0.70 ±1.50	0.34 ±0.81	<u>0.28</u> ±0.79	0.10 ±0.39	0.11 ±0.45	0.03 ±0.20	0.01 ±0.13	<u>0.08</u> ±0.36	0.07 ±0.41	3.67 ±4.66 (3.52±4.59/0.48±0.90)
ModernBERT (Warner et al., 2025)	<u>0.72</u> ±1.53	<u>0.35</u> ±0.85	0.27 ±0.75	0.10 ±0.39	<u>0.13</u> ±0.46	<u>0.04</u> ±0.21	<u>0.02</u> ±0.13	0.07 ±0.35	0.07 ±0.40	3.74 ±4.95 (3.59±4.88/0.50±0.93)
Qwen (Zhang et al., 2025e)	1.48±1.85	1.02±0.91	0.52±0.05	0.18±0.59	0.14±0.55	<u>0.04</u> ±0.21	0.04±0.20	0.09±0.40	0.11±0.46	5.99±6.18 (5.70±5.94/0.86±1.33)
Granite (Awasthy et al., 2025)	1.81±2.24	1.05±1.19	0.66±1.11	0.22±0.63	0.14±0.55	<u>0.04</u> ±0.21	<u>0.02</u> ±0.16	0.09±0.39	<u>0.09</u> ±0.45	5.32±4.85 (5.08±4.79/1.08±1.30)
Gemma (Vera et al., 2025)	1.28±1.84	0.71±1.17	0.38±0.92	<u>0.16</u> ±0.50	0.21±0.53	0.05±0.24	0.01 ±0.13	0.09±0.38	0.10±0.44	8.83±6.29 (7.19±5.95/1.84±2.12)
PPLX (Eslami et al., 2026)	1.48±2.33	0.72±1.15	0.56±1.16	0.17±0.52	0.14±0.53	<u>0.04</u> ±0.22	<u>0.02</u> ±0.15	0.09±0.39	<u>0.09</u> ±0.45	6.52±6.72 (5.95±6.26/0.96±1.57)
MedGemma-1.5 (Søllergren et al., 2026)	0.68±1.58	0.30±0.84	0.27±0.78	0.10±0.38	0.11±0.44	0.04±0.21	0.01±0.13	0.08±0.37	0.08±0.42	3.96±5.60 (3.82±5.56/0.42±0.86)
Merlin (avg.)	7.31±9.07	1.78±3.20	3.86±5.57	0.38±0.99	0.34±0.90	0.10±0.59	0.36±1.36	0.23±0.70	0.27±0.99	17.27±9.50 (10.43±5.12/6.84±7.12)
BioClinical ModernBERT (Sounack et al., 2025)	1.98 ±3.21	0.77 ±1.38	0.98 ±1.62	0.19 ±0.49	0.24 ±0.66	<u>0.11</u> ±0.51	<u>0.36</u> ±1.08	0.16 ±0.52	0.27 ±0.85	1.72 ±1.77 (1.44±1.51/0.92±1.22)
ModernBERT (Warner et al., 2025)	<u>2.02</u> ±3.33	0.77 ±1.40	<u>1.04</u> ±1.66	<u>0.20</u> ±0.49	0.24 ±0.66	0.12±0.51	0.35 ±1.14	0.16 ±0.52	0.27 ±0.86	1.78 ±1.79 (1.51±1.56/0.93±1.22)
Qwen (Zhang et al., 2025e)	4.11±5.08	1.62±1.88	2.40±3.16	0.36±0.90	0.35±0.87	0.10 ±0.58	0.41±1.24	0.23±0.66	<u>0.29</u> ±0.98	4.99±5.32 (2.58±2.69/3.10±4.40)
Granite (Awasthy et al., 2025)	3.79±4.54	1.51±2.21	2.08±2.57	0.48±0.90	0.39±0.86	0.10 ±0.59	0.45±1.18	0.23±0.69	0.31±0.94	3.48±3.41 (2.41±2.24/2.16±2.54)
Gemma (Vera et al., 2025)	3.05±3.86	<u>1.12</u> ±1.62	1.36±2.11	0.27±0.62	0.43±0.72	0.19±0.53	0.39±1.11	<u>0.21</u> ±0.55	0.32±0.86	8.75±6.19 (4.25±3.28/4.68±4.58)
PPLX (Eslami et al., 2026)	5.82±5.87	2.08±2.24	3.40±3.49	0.41±0.89	0.35±0.85	0.22±0.59	0.62±1.24	0.25±0.70	0.31±0.98	8.98±9.75 (4.68±4.31/4.62±6.42)
MedGemma-1.5 (Søllergren et al., 2026)	2.25±3.72	0.83±1.60	1.06±1.81	0.20±0.53	0.25±0.72	0.11±0.65	0.37±1.34	0.17±0.56	0.24±0.92	1.79±1.89 (1.52±1.69/0.90±1.21)

Table 4: Distillation fidelity on the POLAR-Synth test split: CXR (1,925), CT-RATE (1,844), MR-RATE (1,418), Merlin (1,899). MAE against GPT-OSS targets for the total error count, each category (a)–(h), and matched findings (normal/abnormal). The top row per block reports the test-set average (“Dataset avg.”); **bold/underline** mark the best/second-best encoder per dataset. MedGemma 1.5 (POLAR-LM) is reported below the line.

fully each encoder reproduces the teacher supervision. The ModernBERT variants achieve the lowest MAE across all four datasets on both total errors and matched findings. We adopt its biomedical variant, BioClinical ModernBERT (Sounack et al., 2025), as the POLAR backbone.

Metric	MAE Sig. ↓	(a)	(b)	(c)	(d)	(e)	(f)
<i>ReXVal</i>							
Dataset avg.	1.17±1.16	0.42±0.53	0.32±0.43	0.14±0.25	0.13±0.25	0.08±0.16	0.08±0.16
GREEN	0.84 ±0.99	0.38±0.51	0.53 ±0.61	0.20±0.34	<u>0.21</u> ±0.33	<u>0.10</u> ±0.21	<u>0.10</u> ±0.22
CREPE*	1.71±1.33	0.32±0.38	0.76±0.75	0.15±0.20	0.29±0.32	0.22±0.29	0.28±0.37
POLAR	<u>1.02</u> ±1.19	<u>0.35</u> ±0.44	0.61±0.74	0.17 ±0.28	<u>0.21</u> ±0.36	0.08 ±0.16	0.08 ±0.16
POLAR-LM	<u>1.02</u> ±1.27	0.34 ±0.44	<u>0.56</u> ±0.71	<u>0.19</u> ±0.34	0.20 ±0.38	<u>0.10</u> ±0.21	0.11±0.22
<i>RadEval</i>							
Dataset avg.	3.98±2.83	1.18±1.49	1.76±1.69	0.26±0.55	0.20±0.49	0.24±0.59	0.35±0.69
GREEN	1.96±1.65	1.06±1.12	1.07±1.13	<u>0.26</u> ±0.52	0.24±0.50	0.26±0.62	<u>0.37</u> ±0.67
CREPE*	2.17±1.64	0.75±0.74	1.36±1.10	0.28±0.43	0.35±0.40	0.27±0.40	0.55±0.57
POLAR	1.36 ±1.30	0.58 ±0.69	0.77 ±0.86	0.24 ±0.50	<u>0.22</u> ±0.44	<u>0.24</u> ±0.59	<u>0.37</u> ±0.64
POLAR-LM	<u>1.46</u> ±1.45	<u>0.63</u> ±0.80	<u>0.79</u> ±0.96	0.24 ±0.52	<u>0.21</u> ±0.45	0.23 ±0.58	0.33 ±0.66

Table 5: Error detection on human-annotated benchmarks (GREEN categories (a)–(f) only; POLAR predictions restricted accordingly). MAE against expert clinically significant errors. “Dataset avg.” is the mean expert count; **bold/underline** mark best/second-best per benchmark. *CREPE predicts combined significant/in-significant counts and is excluded from the ranking.

On radiologist-annotated benchmarks (Table 5), POLAR achieves the lowest MAE for significant errors on RadEval and remains competitive on ReXVal, with strong hallucination and omission detection. POLAR-LM performs comparably. This confirms that performance on our synthetic dataset

transfers reliably to human-annotated CXR benchmarks.

For comparison with metrics that do not predict explicit error counts, we report Kendall’s τ against expert error counts (Table 6). POLAR EC and POLAR-LM EC rank first and second on both ReXVal and RadEval, while remaining competitive on RaTE-Eval Task 1.

Metric	ReXVal τ	RadEval τ	RaTE1 τ (sum)	RaTE1 τ (norm)
BLEU	0.32	0.00	0.11	0.16
ROUGE	0.55	0.03	0.23	0.21
BERTScore	0.49	0.01	0.25	0.28
CheXbert	0.41	0.35	0.23	0.20
RadGraph	0.55	0.01	0.28	0.34
RadCliQ-v1	0.63	0.12	0.31	0.37
GREEN	0.64	0.19	0.41	0.32
GREEN EC	0.74	0.44	<u>0.43</u>	0.25
RaTEScore	0.51	0.05	0.31	<u>0.35</u>
SRRBert	0.46	0.17	0.10	0.10
RadEvalBERTScore	0.54	0.04	0.22	0.21
CREPE	0.74	0.50	0.41	0.27
CRIMSON	0.63	0.25	0.29	0.27
CRIMSON EC	0.74	0.55	0.29	0.19
POLAR	0.63	0.22	0.32	0.31
POLAR EC	0.76	0.62	0.41	0.26
POLAR-LM	0.59	0.21	0.39	0.31
POLAR-LM EC	<u>0.75</u>	<u>0.58</u>	0.44	0.21

Table 6: Kendall’s τ between metric scores and expert clinically significant error counts on ReXVal, RadEval, and RaTE-Eval Task 1 (RaTE1; summed and normalized). “EC” denotes error-count predictions; **bold/underline** mark best/second-best per column.

6.3 Human Preference Alignment

Table 7 reports radiologist preference alignment. POLAR achieves the best RadPref rank correlation ($\tau_{\Delta} = 0.78$), surpassing CRIMSON (0.67), and the best RadPref preference accuracy (0.88), just ahead of CRIMSON and POLAR EC (both 0.86). POLAR-LM ties for best on GreenPref (0.71). On RaTE2, several lexical and embedding metrics rank higher, indicating that preference alignment remains benchmark-dependent.

Metric	GreenPref acc	RadPref τ_{Δ} / acc	RaTE2 τ
BLEU	0.50	0.44 / 0.72	0.28
ROUGE	0.71	0.47 / 0.71	0.40
BERTScore	0.62	0.47 / 0.72	<u>0.47</u>
CheXbert	0.52	0.53 / 0.81	0.20
RadGraph	0.63	0.45 / 0.73	0.35
RadCliQ-v1	0.62	0.55 / 0.75	0.49
GREEN	<u>0.63</u>	0.59 / 0.85	0.39
GREEN EC	0.44	0.55 / 0.82	0.29
RaTEScore	0.62	0.53 / 0.78	0.46
SRRBert	0.58	0.52 / 0.81	0.21
RadEvalBERTScore	0.62	0.50 / 0.71	0.30
CREPE	<u>0.63</u>	0.51 / 0.76	0.28
CRIMSON	0.56	0.67 / <u>0.86</u>	0.26
CRIMSON EC	0.63	0.53 / 0.83	0.14
POLAR	0.58	0.78 / 0.88	0.44
POLAR EC	0.63	0.65 / 0.86	0.22
POLAR-LM	0.71	0.69 / 0.82	0.37
POLAR-LM EC	0.56	0.59 / 0.81	0.20

Table 7: Human preference alignment. “acc” = accuracy vs. expert preference; τ_{Δ} = correlation of expert and metric score differences; RaTE2 τ = correlation with expert scores. “EC” = error-count predictions; **bold/underline** = best/second-best per column.

6.4 RadJudge

Figure 5 reports RadJudge results across ten clinical criteria. CRIMSON ranks first, as expected given its alignment with the benchmark’s setup. Excluding CRIMSON, POLAR-LM leads (13/30) and POLAR ties for second (12/30), both strong on normal finding handling and false finding penalization, consistent with their polarity-aware design.

6.5 Inference Time

Figure 6 plots expert alignment against per-sample inference time, with bubble size proportional to model parameters. POLAR exceeds the alignment of CRIMSON and POLAR-LM while running $11\times$ faster than CRIMSON, confirming that encoder-based regression captures the clinical signal of generative evaluation at a fraction of the cost. This makes POLAR practical for large-scale benchmarking and training-time rewards, with the RL-rewards benchmark in Appendix H supporting online use.

RadJudge		CRIMSON	GREEN	RadGraph	CheXbert	RaTEScore	RadCliQ-v1	ROUGE	BLEU	BERTScore	SRRBert	RadEvalBERTScore	CREPE	POLAR	POLAR-LM
1	False Finding Penalization	3	0	1	3	1	3	1	0	0	0	3	3	3	3
2	Patient Context Sensitivity	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	Normal Finding Handling	3	0	0	3	0	0	0	0	0	0	0	0	2	2
4	Paraphrase Robustness	3	1	1	2	0	0	0	1	0	1	0	0	1	2
5	Clinical Practicality	2	2	0	1	1	0	0	1	0	2	0	0	2	1
6	Location Error Handling	3	3	0	2	0	0	0	1	0	2	0	1	1	2
7	Measurement Error Sensitivity	3	1	1	0	1	2	1	1	2	1	2	1	2	0
8	Diagnostic Precision	3	1	0	1	3	1	0	0	3	1	2	2	0	1
9	Clinical Significance Weighting	3	0	0	0	2	2	0	0	1	0	1	0	0	0
10	Partial Credit Assignment	3	2	2	0	0	0	0	0	0	0	1	2	1	2
Total (/30)		26	10	5	12	8	8	2	4	6	7	9	9	12	13

Figure 5: RadJudge results across ten clinical categories (three cases each, 30 total). *RadCliQ-v1 values are taken from (Baharoon et al., 2026). Each metric is scored on whether its ranking of candidate reports matches the expected ordering from expert radiologists.

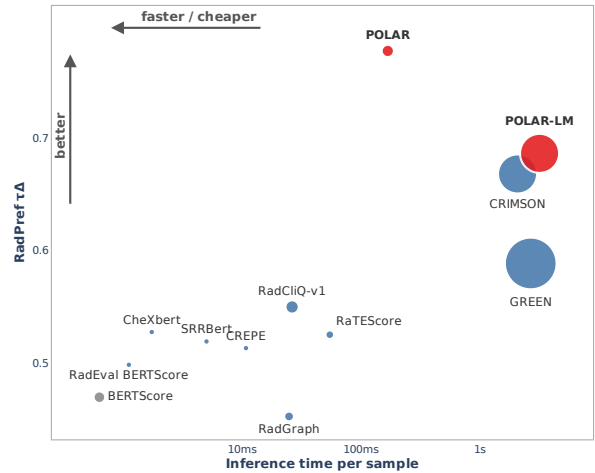


Figure 6: Performance-efficiency trade-off on RadPref. The x-axis shows inference time per sample, the y-axis reports RadPref τ_{Δ} , bubble size denotes model size.

6.6 Anatomy-Conditioned Error Localization

Table 8 reports how well POLAR localizes errors and matched findings to the conditioned anatomy. POLAR’s suppression rates for significant errors are consistently high (91-98%), confirming that it correctly predicts zero for absent anatomies. Detection rates are lower (66-89%), reflecting the harder task of producing non-zero counts on the correct anatomy. Macro-F1 reaches 68-87% for significant errors and 86-92% for matched findings across CXR, CT, and MRI reports. POLAR-LM tends to achieve higher detection while POLAR tends to be stronger on suppression, though the trade-off varies across datasets. Per-anatomy results are reported in Appendix I.

Model	Supp. Sig (%)	Det. Sig (%)	F1 Sig (%)	Supp. Match (%)	Det. Match (%)	F1 Match (%)
<i>CXR</i>						
POLAR	96.2	75.3	75.6	95.9	86.0	87.7
POLAR-LM	95.1	80.0	80.3	95.8	84.7	86.3
<i>CT-RATE</i>						
POLAR	97.7	74.2	75.8	96.1	91.8	92.3
POLAR-LM	95.3	80.1	73.0	92.4	93.3	91.3
<i>MR-RATE</i>						
POLAR	97.7	89.2	86.5	84.3	94.4	87.8
POLAR-LM	96.7	90.5	82.3	85.6	93.2	87.2
<i>Merlin</i>						
POLAR	91.2	65.6	68.3	84.5	86.2	86.2
POLAR-LM	75.0	84.6	70.8	61.8	92.9	84.6

Table 8: Anatomy-conditioning evaluation on POLAR-Synth. Supp. (%): correct zero prediction on negatives; Det. (%): non-zero prediction on positives; F1 (%): macro-averaged binary F1 across anatomies. Sig = significant errors; Match = matched findings.

6.7 Component Analyses

Anatomy conditioning. To isolate the effect of anatomy conditioning, we retrain POLAR using only full-report samples (C = “full report”), excluding all anatomy-conditioned examples. Without anatomy conditioning, region-level MAE increases sharply across all datasets, indicating that localized evaluation does not emerge from full-report supervision alone (Table 9). Whole-report performance also degrades slightly, suggesting that anatomy-conditioned examples provide a useful auxiliary learning signal for global evaluation.

Dataset	Region-conditioned		Whole-report	
	POLAR	—anat.	POLAR	—anat.
CXR	0.20	2.33	0.78	0.82
MR-RATE	0.28	1.40	0.70	0.75
CT-RATE	0.13	3.23	0.99	1.03
Merlin	0.25	7.47	1.98	2.08

Table 9: Anatomy-conditioning ablation on POLAR-Synth (MAE $\downarrow$). —anat. denotes POLAR retrained on full-report examples only.

Context length. To isolate the effect of context length, we train and evaluate POLAR with the same backbone using a 512-token instead of a 2,048-token context window. Performance is nearly unchanged on CXR, whose P99 report length is only 161 tokens, but degrades consistently on the longer CT and MRI datasets, where jointly encoding the reference and candidate reports leads to truncation at 512 tokens (Table 10). This shows that longer context is particularly important for longer reports.

Polarity-aware scoring. Because polarity is encoded in the POLAR-Synth supervision, we assess

Dataset	P99	Context length	
		2,048 tok.	512 tok.
CXR	161	0.78	0.80
MR-RATE	528	0.70	0.91
CT-RATE	555	0.99	1.30
Merlin	825	1.98	3.47

Table 10: Context-length ablation on POLAR-Synth (full-report MAE $\downarrow$). Both models use the same backbone and are trained and evaluated at their respective context lengths. P99 denotes the 99th percentile report length in ModernBERT tokens.

its effect at the score level by equalizing normal and abnormal match weights while keeping POLAR’s predictions fixed. This reduces RadPref τ_{Δ} from 0.78 to 0.74 and RaTE-Eval Task 2 τ from 0.44 to 0.43, while GreenPref accuracy increases from 0.58 to 0.65. Although gains vary across benchmarks, polarity awareness remains clinically motivated; Appendix G provides a motivating example where polarity-blind scoring treats clinically unequal outcomes as equivalent.

7 Conclusion

We introduced POLAR, a fast encoder-based metric for radiology report evaluation that combines polarity-aware multi-head regression, anatomy conditioning, and long-context encoding. Trained on POLAR-Synth, a dataset distilled from a locally hosted LLM, POLAR aligns well with radiologist judgments across error counting, preference, and stress-test benchmarks while running faster than generative evaluators. We release the full pipeline for data generation, training, and inference as open source, enabling transparent, scalable, and privacy-preserving evaluation across CXR, CT, and MRI, and supporting extension to new modalities and anatomies. POLAR provides both a practical evaluation metric and a candidate reward signal for training report generation models.

Code and Data Availability

The source code for POLAR is publicly available at <https://github.com/lbuess/POLAR>, while the corresponding trained model and dataset are hosted on Hugging Face at <https://huggingface.co/collections/lbuess/polar>.

Limitations

POLAR has several limitations. First, our training targets are distilled from GPT-OSS-120B

rather than produced directly by expert radiologists. While this supports an open and privacy-preserving pipeline, the teacher has only been independently validated against CXR expert annotations. On ReXVal, its total-error correlation ($\tau = 0.78$) approaches that of the proprietary GPT-4 teacher reported by GREEN ($\tau = 0.79$). We further rely on downstream alignment with radiologists across seven independent benchmarks (Section 6) as indirect evidence that the supervision is clinically informative, but the quality of the distilled supervision remains dependent on the teacher model. Stronger proprietary models such as GPT-5 could yield higher-quality supervision at the cost of openness. Second, available expert-annotated benchmarks cover only CXR, so generalization to CT and MRI is currently evaluated against synthetic targets and remains to be validated against human annotations. Furthermore, while POLAR clearly reduces MAE relative to the dataset average on CT-RATE, MR-RATE, and Merlin (Table 4), the absolute errors are still substantial; POLAR is best understood as a first step toward reliable long-report evaluation rather than a source of exact error counts on volumetric modalities. Third, anatomy supervision is constructed from dataset-specific taxonomies. Although POLAR consumes free-text anatomy conditions and is trained with synonymous anatomy names, robustness to unseen institutional terminology, different levels of anatomical granularity, and entirely new taxonomies has not been systematically evaluated. Moreover, anatomy-conditioned detection remains challenging, with region-specific findings still being missed. We therefore view anatomy-conditioned evaluation as a first step toward reliable localized report assessment. Fourth, full anatomy-resolved evaluation requires one forward pass per condition, making it more expensive than a single global score; POLAR-LM avoids this but is substantially slower. Fifth, reducing report quality to a single score can obscure distinct clinical error profiles, as reports with different types or locations of errors may receive similar scores. POLAR’s per-category and anatomy-conditioned predictions provide finer-grained information, but the aggregate score necessarily compresses this structure. Sixth, as a regression-based model, POLAR outputs only structured counts, while generative evaluators such as GREEN also produce free-text explanations of detected errors. POLAR-LM retains this capability; POLAR trades it for speed. Finally, unlike CRIMSON, POLAR does not predict

fine-grained severity levels for individual errors, which can be informative for downstream clinical triage.

Potential risks. POLAR is intended for research benchmarking and training-time feedback, not clinical decision-making. Reported POLAR scores should not substitute for radiologist evaluation when assessing models intended for clinical use.

Acknowledgments

The authors gratefully acknowledge the scientific support and HPC resources provided by the Erlangen National High Performance Computing Center (NHR@FAU) of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU). The hardware is funded by the German Research Foundation (DFG).

References

- Asad Aali, Vasiliki Bikia, Maya Varma, Nicole Chiou, Sophie Ostmeier, Arnav Singhvi, Magdalini Paschali, Ashwin Kumar, Andrew Johnston, Karimar Amador-Martinez, and 1 others. 2026. [Toward expert-level medical text validation with language models](#). *npj Digital Medicine*.
- Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, and 1 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.
- Parul Awasthy, Aashka Trivedi, Yulong Li, Meet Doshi, Riyaz Bhat, Vishwajeet Kumar, Yushu Yang, Bhavani Iyer, Abraham Daniels, Rudra Murthy, and 1 others. 2025. Granite embedding r2 models. *arXiv preprint arXiv:2508.21085*.
- Mohammed Baharoon, Thibault Heintz, Siavash Raissi, Mahmoud Alabbad, Mona Alhammad, Hassan AlOmaish, Sung Eun Kim, Oishi Banerjee, and Pranav Rajpurkar. 2026. Crimson: A clinically-grounded llm-based metric for generative radiology report evaluation. *arXiv preprint arXiv:2603.06183*.
- Satanjeev Banerjee and Alon Lavie. 2005. [METEOR: An automatic metric for MT evaluation with improved correlation with human judgments](#). In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.
- Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Anton Schwaighofer, Anja Thieme, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, and 1 others. 2024. Maira-

- 2: Grounded radiology report generation. *arXiv preprint arXiv:2406.04449*.
- Louis Blankemeier, Ashwin Kumar, Joseph Paul Cohen, Jiaming Liu, Longchao Liu, Dave Van Veen, Syed Jamal Safdar Gardezi, Hongkun Yu, Magdalini Paschali, Zhihong Chen, and 1 others. 2026. [Merlin: a computed tomography vision–language foundation model and dataset](#). *Nature*, 652(8112):1318–1328.
- Lukas Buess, Matthias Keicher, Nassir Navab, Andreas Maier, and Soroosh Tayebi Arasteh. 2025. [From large language models to multimodal ai: a scoping review on the potential of generative ai in medicine](#). *Biomedical Engineering Letters*, 15(5):845–863.
- Lukas Buess, Robert Kurin, Adarsh Bhandary Panambur, Tomas Arias-Vergara, and Andreas Maier. 2026. [Alo: Addressing class imbalance in radiology report generation through anatomy-level oversampling](#). In *Proceedings of The 9th International Conference on Medical Imaging with Deep Learning*, volume 315 of *Proceedings of Machine Learning Research*, pages 2998–3017. PMLR.
- Amos Calamida, Farhad Nooralahzadeh, Morteza Rohanian, Koji Fujimoto, Mizuho Nishio, and Michael Krauthammer. 2023. Radiology-aware model-based evaluation metric for report generation. *arXiv preprint arXiv:2311.16764*.
- Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, Emily Tsai, Andrew Johnston, Cameron Olsen, Tanishq Mathew Abraham, Sergios Gatidis, Akshay S Chaudhari, and Curtis Langlotz. 2024. [Chexagent: Towards a foundation model for chest x-ray interpretation](#). In *AAAI 2024 Spring Symposium on Clinical Foundation Models*.
- Gihun Cho, Seunghyun Jang, Hanbin Ko, Inhyeok Baek, and Chang Min Park. 2025. [CREPE: Rapid chest X-ray report evaluation by predicting multi-category error counts](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 21738–21755, Suzhou, China. Association for Computational Linguistics.
- Jean-Benoit Delbrouck, Pierre Chambon, Christian Bluethgen, Emily Tsai, Omar Almusa, and Curtis Langlotz. 2022. [Improving the factual correctness of radiology report generation with semantic rewards](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4348–4360, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Sedigheh Eslami, Maksim Gaiduk, Markus Krimmel, Louis Mark Milliken, Bo Wang, and Denis Bykov. 2026. [Diffusion-pretrained dense and contextual embeddings](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 990–1004, San Diego, California, USA. Association for Computational Linguistics.
- Forithmus. 2026. [MR-RATE: A Vision-Language Foundation Model and Dataset for Magnetic Resonance Imaging](#). Hugging Face.
- Ibrahim Ethem Hamamci, Sezgin Er, Suprosanna Shit, Hadrien Reynaud, Bernhard Kainz, and Bjoern Menze. 2025. [CRG score: A distribution-aware clinical metric for radiology report generation](#). In *Medical Imaging with Deep Learning - Short Papers*.
- Ibrahim Ethem Hamamci, Sezgin Er, Chenyu Wang, Furkan Almas, Ayse Gulnihhan Simsek, Sevval Nil Esirgun, Irem Dogan, Omer Faruk Durugol, Benjamin Hou, Suprosanna Shit, and 1 others. 2026. [Generalist foundation models from a multimodal dataset for 3d computed tomography](#). *Nature Biomedical Engineering*, pages 1–19.
- Alyssa Huang, Oishi Banerjee, Kay Wu, Eduardo Pontes Reis, and Pranav Rajpurkar. 2024. [Fineradscore: A radiology report line-by-line evaluation technique generating corrections with severity scores](#). *arXiv preprint arXiv:2405.20613*.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Johannes Moll, Louisa Fay, Asfandiyar Azhar, Sophie Ostmeier, Sergios Gatidis, Tim Lueth, Curtis P. Langlotz, and Jean-Benoit Delbrouck. 2025. [Structuring radiology reports: Challenging LLMs with lightweight models](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 7707–7724, Suzhou, China. Association for Computational Linguistics.
- Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson, Michael Moseley, Curtis Langlotz, Akshay S Chaudhari, and Jean-Benoit Delbrouck. 2024. [GREEN: Generative radiology report evaluation and error notation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 374–390, Miami, Florida, USA. Association for Computational Linguistics.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Chantal Pellegrini, Ege Özsoy, Benjamin Busam, Benedikt Wiestler, Nassir Navab, and Matthias Keicher. 2025. [Radialog: Large vision-language models for x-ray reporting and dialog-driven assistance](#). In *Medical Imaging with Deep Learning*.

- Tom Pollard, Benjamin E Moody, Li-wei H Lehman, Brian J Gow, Chrystinne Fernandes, Chen Xie, Alistair Johnson, Roger G Mark, and Thomas Heldt. 2026. Physionet as a global platform for biomedical research. *Nature Health*, pages 1–4.
- Mélanie Roschewitz, Kenneth Styppa, Yitian Tao, Jiwoong Sohn, Jean-Benoit Delbrouck, Benjamin Gundersen, Nicolas Deperrois, Christian Bluethgen, Julia E Vogt, Bjoern Menze, and 1 others. 2026. Radaagent: A tool-using ai agent for stepwise interpretation of chest computed tomography. *arXiv preprint arXiv:2604.15231*.
- Andrew Sellergren, Chufan Gao, Fereshteh Mahvar, Timo Kohlberger, Fayaz Jamil, Madeleine Traverse, Alberto Tono, Bashir Sadjad, Lin Yang, Charles Lau, and 1 others. 2026. Medgemma 1.5 technical report. *arXiv preprint arXiv:2604.05081*.
- Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Ng, and Matthew Lungren. 2020. [Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1500–1519, Online. Association for Computational Linguistics.
- Thomas Sounack, Joshua Davis, Brigitte Durieux, Antoine Chaffin, Tom J Pollard, Eric Lehman, Alistair EW Johnson, Matthew McDermott, Tristan Naumann, and Charlotta Lindvall. 2025. Bioclinical modernbert: A state-of-the-art long-context encoder for biomedical and clinical nlp. *arXiv preprint arXiv:2506.10896*.
- Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panayam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, and 1 others. 2025. Embeddinggemma: Powerful and lightweight text representations. *arXiv preprint arXiv:2509.20354*.
- Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Galouédec. 2020. [TRL: Transformers Reinforcement Learning](#).
- Pengyu Wang, Shuchang Ye, Usman Naseem, and Jinman Kim. 2025. Mrg-r1: Reinforcement learning for clinically aligned medical report generation. *arXiv preprint arXiv:2512.16145*.
- Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. [Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.
- Justin Xu, Xi Zhang, Javid Abderezaei, Julie Bauml, Roger Boodoo, Fatemeh Haghighi, Ali Ganjizadeh, Eric Brattain, Dave Van Veen, Zaiqiao Meng, David W Eyre, and Jean-Benoit Delbrouck. 2025. [RadEval: A framework for radiology text evaluation](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 546–557, Suzhou, China. Association for Computational Linguistics.
- Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y. Ng, Curtis P. Langlotz, Vasantha Kumar Venugopal, and Pranav Rajpurkar. 2023a. [Evaluating progress in automatic chest x-ray radiology report generation](#). *Patterns*, 4(9):100802.
- Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Lee, Zahra Shakeri, Andrew Ng, Curtis Langlotz, Vasantha Kumar Venugopal, and Pranav Rajpurkar. 2023b. [Radiology Report Expert Evaluation \(ReXVal\) Dataset](#). *PhysioNet*. Version 1.0.0.
- Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Sheng Zhang, Fei Wang, Yujia Xie, Mahmoud Khademi, Ziyi Yang, and 1 others. 2025. [A clinically accessible small multimodal radiology model and evaluation metric for chest x-ray findings](#). *Nature Communications*, 16(1):3108.
- Kai Zhang, Christopher Malon, Lichao Sun, and Martin Renqiang Min. 2025a. [EditGRPO: Reinforcement learning with post-rollout edits for clinically accurate chest X-ray report generation](#). In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 304–316, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.
- Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. [Bertscore: Evaluating text generation with bert](#). In *International Conference on Learning Representations*.
- Xiaoman Zhang, Julian Nicolas Acosta, Xiaoli Yang, Subathra Adithan, Luyang Luo, Hong-Yu Zhou, Joshua Miller, Ouwen Huang, Zongwei Zhou, Ibrahim Ethem Hamamci, and 1 others. 2025b. [Automated chest x-ray report generation remains unsolved](#). In *Biocomputing 2026: Proceedings of the Pacific Symposium*, pages 236–250. World Scientific.
- Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Jiayu Lei, Weiwei Tian, Ya Zhang, Weidi Xie, and Yanfeng Wang. 2025c. [Development of a large-scale grounded vision language dataset for chest ct analysis](#). *Scientific Data*, 12(1):1636.

- Xiaoman Zhang, Hong-Yu Zhou, Xiaoli Yang, Oishi Banerjee, Julián N. Acosta, Josh Miller, Ouwen Huang, and Pranav Rajpurkar. 2025d. [Rexrank: A public leaderboard for ai-powered radiology report generation](#). In *Proceedings of The First AAAI Bridge Program on AI for Medicine and Healthcare*, volume 281 of *Proceedings of Machine Learning Research*, pages 90–99. PMLR.
- Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, and 1 others. 2025e. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.
- Weike Zhao, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. [RaTEScore: A metric for radiology report generation](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15004–15019, Miami, Florida, USA. Association for Computational Linguistics.
- Hong-Yu Zhou, Julián Nicolás Acosta, Subathra Adithan, Suvrankar Datta, Eric J Topol, and Pranav Rajpurkar. 2026. [Medversa: a generalist foundation model for diverse medical imaging tasks](#). *NEJM AI*, 3(4):AIoa2500595.

A Candidate Synthesis

This section lists the prompt templates used to synthesize candidate reports for POLAR-Synth (heuristics 2, 4, and 5 from Section 3.2). Each prompt is passed to GPT-OSS-120B with a reference report as input. For synthetic error insertion, a random subset of the eight error categories is sampled.

Synthetic Error Insertion Prompt

[Objective]:
Create a candidate radiology report that subtly integrates specific errors based on the provided reference report.

Process Overview:
You will be presented with:

1. Style of errors.
2. A reference radiology report to base your candidate report on.
3. The desired format for your candidate report.

Note: Be short in your response!

Style of errors:
Introduce errors related to {error_types}. The errors should be woven into the report as if they were genuine observations from a medical image, without any meta-commentary on their accuracy.

Reference Report:
{report}

Desired format for your candidate report:
[Candidate]: <Candidate Report>

Figure 7: Prompt template for synthetic error insertion. A random subset of the eight error categories is sampled per reference report.

Meaning-Preserving Rewrite Prompt

[Objective]:
Create a candidate radiology report that has the same clinical meaning but is slightly rephrased.

Process Overview:
You will be presented with:

1. A reference radiology report to base your candidate report on.
2. The desired format for your candidate report.

Note: Be short in your response!
In this clinical setup, prefer clinically appropriate synonyms and phrase variation while preserving semantic similarity to the reference report.

Reference radiology report:
{report}

Desired format for your candidate report:
[Candidate]: <Candidate Report>

Figure 8: Prompt template for meaning-preserving rewrites, producing lexically varied but clinically equivalent candidates.

B Anatomy List

Table 11 lists the anatomy vocabulary used for each source dataset during annotation. These vocabularies are inserted into the {anatomy_str} placeholder in the POLAR analysis prompt (Appendix C) and provide the label space for both error localization and matched-finding assignment. Vocabularies follow the structures of Moll et al. (2025) for CXR, RadGenome (Zhang et al., 2025c) for CT-RATE,

Sentence Reordering Prompt

[Objective]:
Create a candidate radiology report by reordering sentence structure and optionally removing one sentence.

Process Overview:
You will be presented with:

1. A reference radiology report to base your candidate report on.
2. The desired format for your candidate report.

Note: Be short in your response!

Instructions:
- Reorder the report’s sentence order to change presentation flow.
- Optionally remove one sentence.
- Do not add new findings that are not present in the reference report.
- Keep the output clinically plausible as a candidate report.

Reference radiology report:
{report}

Desired format for your candidate report:
[Candidate]: <Candidate Report>

Figure 9: Prompt template for sentence-level augmentation. Sentences are reordered, with optional removal of one sentence.

and the original works for MR-RATE and Merlin (Forithmus, 2026; Blankemeier et al., 2026). During training, we randomly sample synonyms for each anatomy label to expose POLAR to lexical variation in the conditioning string.

Dataset	Anatomies
CXR	Lungs and Airways; Pleura; Cardiovascular; Hila and Mediastinum; Tubes and Catheters and Support Devices; Musculoskeletal and Chest Wall; Abdominal
CT-RATE	Lung; Trachea and Bronchie; Mediastinum; Heart; Esophagus; Pleura; Bone; Thyroid; Breast; Abdomen
MR-RATE	Brain; Spine
Merlin	Lower Thorax; Liver and Biliary Tree; Gallbladder; Spleen; Pancreas; Adrenal Glands; Kidneys and Ureters; Gastrointestinal Tract; Peritoneal Cavity; Pelvic Organs; Vasculture; Lymph Nodes; Musculoskeletal

Table 11: Dataset-specific anatomy vocabularies used in the POLAR-Synth annotation prompt.

C POLAR Analysis

This section details the prompt used to generate structured annotations with GPT-OSS-120B, which form the supervision targets for POLAR-Synth. The prompt (Figure 10) specifies the eight error categories, the clinical significance rules, the matched-finding splitting rule, and the fixed output format with per-item anatomy tags drawn from the dataset-specific vocabulary listed in Appendix B. Figure 11 shows an example of the corresponding GPT-OSS-120B output for a CXR reference-candidate pair.

POLAR Analysis Prompt

Objective: Evaluate the accuracy of a candidate radiology report in comparison to a reference radiology report composed by expert radiologists.

Process Overview:

You will be presented with:

1. The criteria for making a judgment.
2. The reference radiology report.
3. The candidate radiology report.
4. The desired format for your assessment.
5. Additional annotation instructions.

1. Criteria for Judgment

For each candidate report, determine:

- The count of clinically significant errors.
- The count of clinically insignificant errors.

Errors can fall into one of these categories:

- a) False report of a finding in the candidate.
- b) Missing a finding present in the reference.
- c) Misidentification of a finding's anatomic location, laterality, or position.
- d) Misassessment of the severity or extent of a finding.
- e) Mentioning a comparison that isn't in the reference.
- f) Omitting a comparison detailing a change from a prior study.
- g) Misstatement of quantitative details such as size, number, or measurements.
- h) Misstatement of descriptive attributes (e.g., morphology, shape/contour, margin, overall appearance, ...).

Note: Focus on clinical findings, not writing style.

2. Reference Report: {text1}

3. Candidate Report: {text2}

4. Reporting Your Assessment

Follow this specific format for your output, even if no errors are found. For every finding, assign one anatomy from this list: {anatomy_str}. Use exactly one anatomy from this list for each listed item. If anatomy cannot be determined confidently, use <a>Other.

[Explanation]:

<Explanation>

[Clinically Significant Errors]:

(a) <Error Type>: <The number of errors>. <Error 1> <a>ANATOMY_HERE; <Error 2> <a>ANATOMY_HERE; ... ; <Error n> <a>ANATOMY_HERE
...

(h) <Error Type>: <The number of errors>. <Error 1> <a>ANATOMY_HERE; <Error 2> <a>ANATOMY_HERE; ... ; <Error n> <a>ANATOMY_HERE

[Clinically Insignificant Errors]:

(a) <Error Type>: <The number of errors>. <Error 1> <a>ANATOMY_HERE; <Error 2> <a>ANATOMY_HERE; ... ; <Error n> <a>ANATOMY_HERE
...

(h) <Error Type>: <The number of errors>. <Error 1> <a>ANATOMY_HERE; <Error 2> <a>ANATOMY_HERE; ... ; <Error n> <a>ANATOMY_HERE

[Matched Findings]:

- Normal findings: <The number of findings>. <Finding 1> <a>ANATOMY_HERE; <Finding 2> <a>ANATOMY_HERE; ... ; <Finding n> <a>ANATOMY_HERE
- Abnormal findings: <The number of findings>. <Finding 1> <a>ANATOMY_HERE; <Finding 2> <a>ANATOMY_HERE; ... ; <Finding n> <a>ANATOMY_HERE

5. Additional annotation instructions

5.1 Finding splitting rule (required)

- If one sentence contains multiple distinct findings, split into separate items.
- Example: "Mild cardiomegaly with a small pericardial effusion" -> two findings.

5.2 Clinical significance

- Clinically significant: omission/addition/mischaracterization of an abnormal (positive) finding that could affect interpretation or management.
- Clinically significant: any medical/support device discrepancy (presence, type, course, tip position, or interval change), even if described as expected.
- Clinically insignificant: discrepancies limited to explicitly negated common-normal statements (e.g., "no pleural effusion", "no pneumothorax"), unless the negation contradicts a true abnormal reference finding.
- For attribute mismatches (location, severity, morphology/appearance): significant by default; insignificant only if minor (e.g., slight position offset or small size delta) and clinically irrelevant.

5.3 Matched finding rules

- A finding matches if the same core finding appears in both reports.
- Attribute mismatches do NOT break core-finding matching (e.g., location, severity, morphology, measurement, temporal status): include the finding in [Matched Findings] and also count one error in the relevant attribute category.
- If reference is unspecific and candidate adds specific detail for an attribute (e.g., location): acceptable (no attribute error for added specificity alone).
- If reference is specific and candidate is unspecific for an attribute (e.g., location): still matched, but count one attribute error in the relevant category.
- In [Matched Findings], place medical/support device findings under '- Abnormal findings' (even when described as expected position).
- Provide the '- Normal findings' and '- Abnormal findings' breakdown ONLY for [Matched Findings]. Do NOT add a normal/abnormal breakdown anywhere else.

Figure 10: Full annotation prompt used with GPT-OSS-120B to produce structured error, matched-finding, and anatomy-tagged outputs for POLAR-Synth. Placeholders {text1 }, {text2}, and {anatomy_str} are filled with the reference report, candidate report, and dataset-specific anatomy list (Appendix B).

GPT-OSS 120B POLAR Analysis

```
[Explanation]:
The candidate report correctly identifies a left basilar opacity consistent with the reference, but it omits the right-side component, adds a false
right pleural effusion, and incorrectly reports tubes that were not mentioned. It also describes the opacity as "patchy" rather than "linear," which
is a descriptive mismatch. These constitute clinically significant errors; no clinically insignificant errors are present.

[Clinically Significant Errors]:
(a) False report of a finding: 2. Endotracheal and enteric tubes are in standard positions <a>Tubes and Catheters and Support Devices</a>; Probable
trace right pleural effusion <a>Pleura</a>
(b) Missing a finding: 1. Right basilar opacity (bibasilar opacities) <a>Lungs and Airways</a>
(c) Misidentification of a finding's anatomic location, laterality, or position: 0.
(d) Misassessment of the severity or extent of a finding: 0.
(e) Mentioning a comparison that isn't in the reference: 0.
(f) Omitting a comparison detailing a change from a prior study: 0.
(g) Misstatement of quantitative details such as size, number, or measurements: 0.
(h) Misstatement of descriptive attributes: 1. Morphology described as patchy rather than linear for basilar opacity <a>Lungs and Airways</a>

[Clinically Insignificant Errors]:
(a) False report of a finding: 0.
(b) Missing a finding: 0.
(c) Misidentification of a finding's anatomic location, laterality, or position: 0.
(d) Misassessment of the severity or extent of a finding: 0.
(e) Mentioning a comparison that isn't in the reference: 0.
(f) Omitting a comparison detailing a change from a prior study: 0.
(g) Misstatement of quantitative details such as size, number, or measurements: 0.
(h) Misstatement of descriptive attributes: 0.

[Matched Findings]:
- Normal findings: 0.
- Abnormal findings: 1. Left basilar patchy opacity, likely atelectasis <a>Lungs and Airways</a>
```

Figure 11: Worked example of a GPT-OSS-120B POLAR analysis. **Reference:** “Linear bibasilar opacities suggestive of atelectasis or scarring. No acute cardiopulmonary process.” **Candidate:** “1. Endotracheal and enteric tubes are in standard positions. 2. Left basilar patchy opacity, likely atelectasis. 3. Probable trace right pleural effusion.” The model produces a free-text explanation, per-category counts of clinically significant and insignificant errors with anatomy tags, and a polarity-split list of matched findings. These structured outputs form the supervision targets for POLAR.

D Training Hyperparameters

This section details the training setup for POLAR and POLAR-LM. Both models are trained on the POLAR-Synth dataset described in Section 3.

D.1 POLAR

We finetune thomas-sounack/Bio-ModernBERT-large (395M parameters) using AdamW with a learning rate of 2×10^{-5} and a linear scheduler. Training runs for 10 epochs with an effective batch size of 128. Inputs are formatted as Condition: {region} [SEP] Reference: {reference} [SEP] Candidate: {candidate} with a 2,048-token context window. The checkpoint with the lowest validation loss is retained. Training was run on a single NVIDIA H100 GPU.

D.2 POLAR-LM

We finetune google/medgemma-1.5-4b-it (4B parameters) using supervised finetuning with TRL’s SFTTrainer, a learning rate of 2×10^{-5} , cosine decay, and a warmup ratio of 0.05. Training runs for 10 epochs with an effective batch size of 2,048 (8 per device, 4 GPUs, 64 gradient accumulation steps), using bfloat16 precision and a

maximum sequence length of 4,096 tokens. We evaluate and save checkpoints at each epoch, retain the checkpoint with the lowest validation loss, and run training on a single node with four NVIDIA H100 GPUs using DeepSpeed ZeRO-3.

D.3 Embedding Models

For the encoder comparison in Table 4, we benchmark four recent embedding models against the ModernBERT backbones: google/embeddinggemma-300m (Gemma), ibm-granite/granite-embedding-english-r2 (Granite), perplexity-ai/pplx-embed-v1-0.6b (PPLX), and Qwen/Qwen3-Embedding-0.6B (Qwen). All models are fine-tuned on POLAR-Synth under the same setup as POLAR.

E POLAR Score Behavior

Table 12 illustrates the score defined in Eq. (9) for representative report scenarios using the default weights ($w_{ab} = w_{fp} = w_{fn} = 2$, $w_n = 0.5$, $w_{attr} = 1$).

Three properties of the score are apparent from the table. First, false negatives are penalized more heavily than false positives: a single missed ab-

Scenario	m_{ab}	e_{fn}	e_{fp}	e_{attr}	Score
<i>Abnormal reference (3 findings)</i>					
Perfect report	3	0	0	0	+1.00
All matched, 2 attribute errors	3	0	0	2	+0.50
All matched, 1 false positive	3	0	1	0	+0.50
2/3 matched, no errors	2	1	0	0	+0.25
2/3 matched, 1 false positive	2	1	1	0	0.00
1/3 matched, no errors	1	2	0	0	-0.20
Blank report	0	3	0	0	-0.50
<i>Normal reference ($m_n = 2$)</i>					
Correctly reported as normal	0	0	0	-	+1.00
1 hallucinated finding	0	0	1	-	-0.33
3 hallucinated findings	0	0	3	-	-0.71

Table 12: POLAR score for representative scenarios. The upper block assumes a reference with three abnormal findings and no normal findings ($m_n = 0$); the lower block assumes a normal reference with two matched normal findings ($m_n = 2$).

normality reduces the score by 0.75 from a perfect report, whereas a single false positive reduces it by only 0.50. This asymmetry arises because e_{fn} enters both the penalty term E and the reference normalizer W_{ref} , while e_{fp} contributes only to E . Second, a blank report scores exactly $-w_{fn}/(w_{ab} + w_{fn}) = -0.5$ regardless of the number of missed findings, leaving room on the negative half of the scale to distinguish empty reports from actively harmful ones with hallucinated findings. Third, one false positive is equivalent in penalty to two attribute errors ($w_{fp} = 2w_{attr}$), reflecting the higher clinical severity of hallucinated findings relative to imprecise descriptions of correctly identified ones.

F Scoring Weight Sensitivity

To assess sensitivity to the manually specified weights in Eq. (9), we perturb each weight independently by $\pm 50\%$. On RadPref, τ_Δ remains between 0.76 and 0.78 (default: 0.78), while rankings remain highly correlated with the default score (Kendall’s $\tau = 0.90$ -0.99), indicating robust behavior across substantial weight changes.

G Polarity Blindness

A general limitation of polarity-blind scoring is that it can treat clinically unequal outcomes as equivalent. For the reference “*no pleural effusion; cardiomegaly present*”, a candidate containing only “*no pleural effusion*” and one containing only “*cardiomegaly present*” can receive the same score, despite the latter preserving the clinically more important abnormal finding. With the default POLAR

weighting, the former scores -0.33 , whereas the latter scores $+1.00$.

H RadEval RL Benchmark

RL training calls the reward metric once per rollout, requiring both speed and clinical sensitivity. We evaluate both dimensions, extending the RadEval benchmark³ (Xu et al., 2025) with POLAR and POLAR-LM.

Metric	Init (s)	Per sample (ms)	Peak VRAM (MB)
BLEU	0.00	0.10	-
ROUGE	0.02	<u>0.21</u>	-
BERTScore	0.95	0.54	259
RadGraph	0.37	25.24	890
RadEvalBERTScore	1.09	1.06	606
F1CheXBERT	2.17	1.48	434
RaTE-Score	2.97	49.01	1,130
SRRBert	0.76	4.56	420
RadCliQ-v1	2.49	26.52	1,176
CREPE	0.90	10.67	477
CRIMSON	3.52	1,965.52	10,545
GREEN	6.00	2,668.74	23,857
POLAR	0.89	170.86	1,888
POLAR-LM	2.94	3,175.27	10,284

Table 13: Inference speed on 20 chest X-ray reports. Init: model load time after weights are cached on disk. Per sample: median latency over 3 warm passes (1 warm-up discarded). VRAM: peak GPU memory delta; CPU-only metrics show -. **Bold/underline** = fastest/second-fastest per sample.

Setup. We follow the RadEval benchmark protocol (Xu et al., 2025): 20 chest X-ray reports from the RadEval speed fixture are processed on a single NVIDIA A40 GPU, with metrics loaded, benchmarked, and torn down in sequence within a single process (del + gc.collect + torch.cuda.empty_cache) to prevent cross-metric memory interference. For each metric, one warm-up pass is discarded before three timed passes. The protocol reports median per-sample latency and peak torch.cuda.max_memory_allocated delta as a GPU memory overhead approximation.

Speed. Table 13 summarises the results. Encoder-based clinical metrics (F1CheXBERT, SRRBert, CREPE) are readily usable in an online loop at 1-11 ms and under 700 MB of VRAM. POLAR runs at ≈ 170 ms with 1.9 GB, offering a favourable accuracy-throughput trade-off. Generative metrics (CRIMSON, GREEN, POLAR-LM) require 10-

³https://github.com/jbdel/RadEval/blob/main/docs/trl_rewards_benchmarks.md

Reference	Candidate rollout	BLEU ↑	BERTScore ↑	RadGraph ↑	RadCliQ ↓	CRIMSON ↑	POLAR ↑	POLAR-LM ↑
No acute cardiopulmonary process.	No acute cardiopulmonary process.	1.00	1.000	1.00	9.37	1.000	1.00	1.00
Mild cardiomegaly.	The cardiac silhouette is mildly enlarged.	0.00	0.215	0.00	10.67	1.000	1.00	1.00
No pleural effusion.	Pleural effusion.	0.00	0.893	0.50	11.06	−0.333	−1.00	−1.00
Mild pulmonary edema.	Severe pulmonary edema.	0.00	0.955	0.67	10.20	0.500	0.33	0.33
Stable bibasilar atelectasis without consolidation.	New right lower lobe consolidation.	0.00	0.355	0.00	11.12	−0.333	−0.67	−0.67

Table 14: Divergence gallery: five curated (reference, candidate) pairs scored by each metric. Row 3 (negation flip, **bold**) illustrates the key failure mode: BERTScore rewards a clinically wrong output at 0.893, while CRIMSON and POLAR correctly assign a negative score.

24 GB and 2-3 s per sample, making them best suited to offline reward labelling.

Clinical sensitivity. Table 14 scores five curated reference and candidate pairs covering clinically relevant failure modes: negation flip, missing finding, hallucination, severity flip, and opposite clinical conclusion. On the negation flip (“*No pleural effusion*” to “*Pleural effusion*”), BERTScore assigns 0.893, higher than its score for a correct paraphrase (0.215), while POLAR, POLAR-LM, and CRIMSON correctly penalise the error; a policy optimised against BERTScore alone could learn to invert negations without penalty.

I Anatomy Conditioning Details

This appendix details the anatomy-conditioned split of POLAR-Synth used in Section 6.6. Table 19 reports the per-anatomy positive and negative sample counts in the training split, where a positive example has at least one significant error or matched finding associated with the conditioned anatomy and a negative example has all targets set to zero. The distribution is notably imbalanced: anatomies that are nearly always described in their respective modality (e.g., Lung in CT-RATE, Brain in MR-RATE) are dominated by positives, while peripheral anatomies (e.g., Thyroid, Breast, Spine) are dominated by negatives. This imbalance biases the regression model toward zero predictions on rare positives, consistent with POLAR’s higher suppression and lower detection rates observed in Section 6.6.

Tables 15–18 report per-anatomy binary F1 scores on the test split for POLAR and POLAR-LM, separately for significant errors (any error present vs. absent) and matched findings (any match present vs. absent). POLAR and POLAR-LM trade off across anatomies, with neither model dominating consistently across modalities.

Anatomy	Err F1 (%)	Err F1 (%)	Match F1 (%)	Match F1 (%)
	POLAR	POLAR-LM	POLAR	POLAR-LM
Abdominal	82.9	83.3	90.8	89.1
Cardiovascular	85.1	86.6	91.0	89.5
Hila and Mediastinum	80.8	80.2	88.0	82.3
Lungs and Airways	91.2	89.0	91.1	89.7
Musculoskeletal and Chest Wall	79.5	80.8	83.5	81.1
Other	4.4	37.4	74.0	72.7
Pleura	90.8	93.0	93.0	92.2
Tubes and Catheters and Support Devices	90.3	92.2	90.4	94.2
Macro avg	75.6	80.3	87.7	86.3

Table 15: Per-anatomy binary F1 scores on CXR (anatomy-conditioned). Err F1: any significant errors present vs. absent. Match F1: any matched findings present vs. absent.

Anatomy	Err F1 (%)	Err F1 (%)	Match F1 (%)	Match F1 (%)
	POLAR	POLAR-LM	POLAR	POLAR-LM
Abdomen	69.1	68.5	93.5	89.1
Bone	51.0	55.6	93.5	89.1
Breast	92.3	90.6	90.9	93.0
Esophagus	69.6	55.2	95.2	96.6
Heart	86.2	86.7	93.5	91.5
Lung	92.9	93.0	94.6	94.2
Mediastinum	69.3	68.5	96.5	94.3
Other	54.1	62.4	69.3	75.7
Pleura	90.9	89.7	93.2	88.3
Thyroid	91.7	92.8	98.1	97.2
Trachea and Bronchie	66.7	40.0	97.1	95.8
Macro avg	75.8	73.0	92.3	91.3

Table 16: Per-anatomy binary F1 scores on CT-RATE (anatomy-conditioned). Err F1: any significant errors present vs. absent. Match F1: any matched findings present vs. absent.

Anatomy	Err F1 (%)	Err F1 (%)	Match F1 (%)	Match F1 (%)
	POLAR	POLAR-LM	POLAR	POLAR-LM
Brain	95.0	95.8	99.7	99.6
Other	72.1	73.1	80.9	80.9
Spine	92.3	78.0	82.9	80.9
Macro avg	86.5	82.3	87.8	87.2

Table 17: Per-anatomy binary F1 scores on MR-RATE (anatomy-conditioned). Err F1: any significant errors present vs. absent. Match F1: any matched findings present vs. absent.

Anatomy	Err F1 (%)	Err F1 (%)	Match F1 (%)	Match F1 (%)
	POLAR	POLAR-LM	POLAR	POLAR-LM
Adrenal Glands	80.0	90.9	92.1	86.7
Gallbladder	67.9	77.9	85.4	86.8
Gastrointestinal Tract	77.1	69.1	87.3	84.6
Kidneys and Ureters	75.3	70.7	94.6	90.7
Liver and Biliary Tree	79.7	77.6	89.3	91.6
Lower Thorax	87.7	84.5	93.1	90.4
Lymph Nodes	70.6	67.9	91.3	83.3
Musculoskeletal	62.2	65.1	93.2	81.2
Other	25.9	43.0	41.9	56.4
Pancreas	56.3	70.0	91.2	92.1
Pelvic Organs	50.6	65.6	83.8	80.6
Peritoneal Cavity	70.6	69.2	82.6	81.9
Spleen	80.0	69.8	91.5	91.4
Vasculature	72.4	70.4	89.6	86.3
<i>Macro avg</i>	68.3	70.8	86.2	84.6

Table 18: Per-anatomy binary F1 scores on Merlin (anatomy-conditioned). Err F1: any significant errors present vs. absent. Match F1: any matched findings present vs. absent.

Anatomy	Total	Positive	Negative
<i>CXR</i>			
Abdominal	3,485	1,068	2,417
Cardiovascular	5,091	2,673	2,418
Hila and Mediastinum	5,058	2,640	2,418
Lungs and Airways	5,087	2,673	2,414
Musculoskeletal and Chest Wall	4,817	2,400	2,417
Other	4,771	2,354	2,417
Pleura	5,090	2,672	2,418
Tubes and Catheters and Support Devices	5,090	2,672	2,418
<i>Total</i>	38,489	19,152	19,337
<i>CT-RATE</i>			
Abdomen	2,725	1,857	868
Bone	2,727	1,857	870
Breast	3,942	825	3,117
Esophagus	3,491	1,857	1,634
Heart	2,702	1,857	845
Lung	1,981	1,857	124
Mediastinum	2,279	1,857	422
Other	4,974	1,857	3,117
Pleura	4,968	1,857	3,111
Thyroid	4,068	951	3,117
Trachea and Bronchie	3,116	1,857	1,259
<i>Total</i>	36,973	18,489	18,484
<i>MR-RATE</i>			
Brain	6,600	6,588	12
Other	11,082	6,588	4,494
Spine	10,078	1,130	8,948
<i>Total</i>	27,760	14,306	13,454
<i>Merlin</i>			
Adrenal Glands	1,349	1,017	332
Gallbladder	1,403	1,017	386
Gastrointestinal Tract	1,400	1,017	383
Kidneys and Ureters	1,348	1,017	331
Liver and Biliary Tree	1,295	1,016	279
Lower Thorax	1,448	1,017	431
Lymph Nodes	1,445	1,017	428
Musculoskeletal	1,586	1,017	569
Other	3,306	1,017	2,289
Pancreas	1,347	1,017	330
Pelvic Organs	1,487	1,017	470
Peritoneal Cavity	1,539	1,017	522
Spleen	1,346	1,017	329
Vasculature	1,442	1,017	425
<i>Total</i>	21,741	14,237	7,504

Table 19: Positive and negative sample counts per anatomy in the POLAR-Synth anatomy-conditioned training split. Positive: anatomy has any significant errors or matched findings. Negative: anatomy absent from reference (all targets zero).