A STRUCTURED COMMITMENT SPACE FOR FACTUAL POLICY LEARNING IN CHEST RADIOGRAPHY

Anonymous authors
Paper under double-blind review

ABSTRACT

Clinical vision-language models can identify the correct finding yet misstate whether it is present or how severe it is. Aggregate reinforcement-learning rewards provide little explicit control over how these different factual errors are penalized. We introduce Fact-Commitment Space, a reward for factual policy learning in chest radiography. It represents free-form clinical text as commitments, each specifying a clinical entity, its assertion state, and relevant attributes. A Clinical Fact Parser (CFP) extracts these commitments, and a deterministic compiler normalizes them. Matching response and reference commitments then assigns separate penalties to omissions, unsupported findings, attribute errors, and contradictions, without an online language-model judge. Under stated conditions on structured commitments, omitting an attribute has lower expected loss than guessing in the zero-confidence limit, reversing an assertion scores below omission, and adding an unsupported commitment lowers the reward. On 600 human-reviewed responses, CFP improves entity–assertion F1 from 13.82% to 75.32% over its Qwen3.5-2B base, and its rewards preserve 72.78% of non-tied reference pair orderings. Training with this reward using Group Relative Policy Optimization (GRPO) yields VeraRad-2B. On 3,000 clinical visual questions, it achieves 75.53% strict accuracy versus 66.70% for its Qwen3.5-2B base, with thinking enabled for both models. On report generation, a task absent from training, it leads the primary quality metric on two of three datasets and produces reports 43.7% shorter than its base in the same mode. Image-swap and no-image interventions assess dependence on study-specific images, complemented by spatial attribution analysis of a cardiomegaly case.

1 INTRODUCTION

Reinforcement learning for open-ended clinical answers requires deciding how individual factual errors should affect the reward. A response can contain several supported findings alongside an unsupported severity grade (Ostmeier et al., 2024). We study chest radiography, where a clinically meaningful ontology, explicit assertion states, and bounded attributes allow us to assign explicit error costs and examine their consequences for policy learning.

Aggregate similarity and preference scores do not directly specify separate costs for omissions, unsupported findings, and attribute mismatches (Yu et al., 2023). Online language-model judges can reason about these errors (Liu et al., 2023; Zheng et al., 2023), but add rollout costs without guaranteeing reproducible scores across training steps. Our reward design addresses two requirements: decomposing responses into identifiable commitments and assigning explicit costs to each error type before policy training.

We propose Fact-Commitment Space (Figure 1), representing each commitment by a clinical entity, an assertion state (present, absent, or uncertain), and bounded attribute-value bindings. A Clinical Fact Parser (CFP) extracts candidates under schema-guided decoding; a deterministic compiler normalizes and validates them. An asymmetric bipartite evaluator matches response and reference commitments, pricing omission, unsupported insertion, attribute mismatch, and contradiction without an online language-model judge. Under stated parameter conditions, the evaluator on structured commitments satisfies three properties: honest attribute omission has lower expected cost than guessing in the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

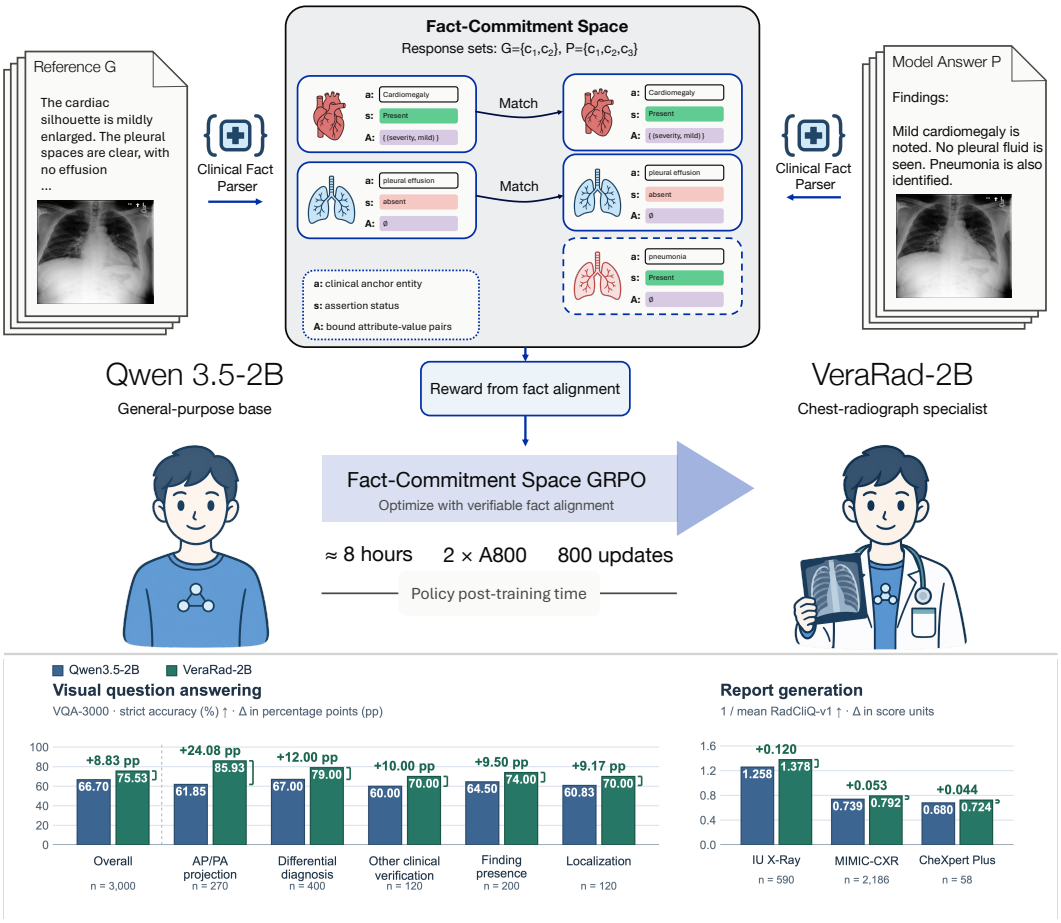


Figure 1: Fact-Commitment Space training and downstream evaluation. Parsed reference and response commitments supply the policy reward. Lower panels compare Qwen3.5-2B and VeraRad-2B with thinking enabled; the five VQA categories have the largest gains among prespecified types with at least 100 questions.

zero-confidence limit (Theorem 1); assertion-state inversion scores below total omission (Theorem 3); and unsupported insertion strictly decreases reward (Theorem 4).

On 600 reviewed responses, CFP improves entity–assertion F1 by 61.50 percentage points over its Qwen3.5-2B base. We train VeraRad-2B with Group Relative Policy Optimization (GRPO) (Shao et al., 2024) on image-question tasks using this reward. With thinking enabled, VeraRad-2B achieves 75.53% strict accuracy on 3,000 clinical visual questions, highest among nine configurations and 8.83 points above its base in the same mode. On report generation, absent from training, it leads the primary metric on two of three datasets with shorter reports. Image-swap, no-image, and spatial attribution analyses examine its dependence on study-specific visual evidence.

Our contributions are a structured fact representation with an asymmetric reward evaluator; sufficient conditions for three scorer-level properties; and an empirical evaluation connecting parsing fidelity and reward preservation to downstream VQA, report generation, and image-intervention controls.

2 Related Work

Clinical factuality measures and semantic rewards. CheXbert extracts fixed finding labels (Smit et al., 2020); RadGraph represents findings and anatomical relations as graphs (Jain et al., 2021), whose overlap also serves as a report-generation reward (Delbrouck et al., 2022). RaTEScore weights medical entity similarity while accounting for negation and entity type (Zhao et al., 2024); GREEN scores and explains clinical errors (Ostmeier et al., 2024); RadCliQ combines metrics to track

radiologist-rated errors (Yu et al., 2023). Building on structured evaluation, Fact-Commitment Space pairs instance-bound attributes and explicit assertion states with predefined asymmetric costs for omission, insertion, and mismatch. Conditions on these costs establish scorer-level properties before training.

Atomic factuality evaluation. FActScore decomposes text into atomic propositions verified against knowledge sources (Min et al., 2023). Related approaches address factual consistency in summarization and retrieval-based revision (Yang et al., 2024; Gao et al., 2023). We use clinically constrained commitments with three assertion states and typed attributes such as laterality, severity, and temporal change. After response and reference commitments are compiled, matching and cost calculation are deterministic and require neither retrieval nor a language-model judge.

Reinforcement learning for medical vision-language models. Medical VLM reinforcement learning uses outcome rewards from reference answers (Pan et al., 2025) or combines answer correctness with process rewards based on report-derived reasoning traces (Fan et al., 2026). VeraRad-2B receives rewards for commitments extracted from final free-form answers, without scoring intermediate reasoning. Its references are compiled from task answers or source reports, supporting open-ended clinical questions. The reward assigns explicit costs to factual errors at the entity, assertion-state, and attribute levels.

3 METHODOLOGY

3.1 TASK, REPRESENTATION, AND PARSING

To distinguish factual errors, we extract claims, normalize equivalent expressions, and compare them with the reference (Figure 1).

Given a radiograph and question or instruction $q = (I, x_{\text{prompt}})$, the policy generates $y \sim \pi_\theta(y \mid q)$. The task $x = (q, G, \mathcal{O})$ supplies reference commitments G; a fixed mask $\mathcal{O}$ limits scoring to clinically determinable entities and attributes. The Clinical Fact Parser (CFP) Φ_ϕ extracts candidates $\tilde{P}$; compiler $\mathcal{N}$ normalizes them into commitments P:

$$y \xrightarrow{\Phi_\phi} \tilde{P} \xrightarrow{\mathcal{N}} P \xrightarrow{L(G,P)} R_{\text{fact}}(P, G, \mathcal{O}) \in (-\infty, 1]. \tag{1}$$

Frozen offline extraction supplies G; unresolved references are quarantined before scoring.

Fact-Commitment Space $\mathcal{S}_{\text{fc}}$ binds each finding to its state and attributes to locate errors within each commitment:

$$c = \langle a,\ s,\ A,\ m,\ e \rangle, \tag{2}$$

where $a \in \mathcal{V}_{\text{anchor}}$ identifies a clinical entity and $s \in \mathcal{S}_{\text{assertion}} = \{\text{present}, \text{absent}, \text{uncertain}\}$ gives its state. Attributes $A = \{(k, v_k)\}_{k \in \mathcal{K}_c}$ describe this instance (e.g., laterality, severity, temporal change); they cannot transfer across instances. A fixed vocabulary registry defines entities, 25 attribute axes $\mathcal{K}$, and each axis's allowed values $\mathcal{V}_k$ (Appendix A). Mention m and provenance e support traceability but are excluded from scoring. The absent state includes explicit negation and normal statements implying absence.

CFP is initialized from Qwen3.5-2B (Qwen Team, 2026a), trained on 19,472 text–commitment pairs, and frozen during policy optimization (Figure 2; Appendix B.1). It uses schema-constrained decoding. The compiler then resolves equivalent names, splits composite entities, checks attribute bindings, and removes duplicates. Out-of-vocabulary entities, illegal values, unresolvable ambiguities, and binding conflicts become unresolved records; relations between entities are retained separately for scoring.

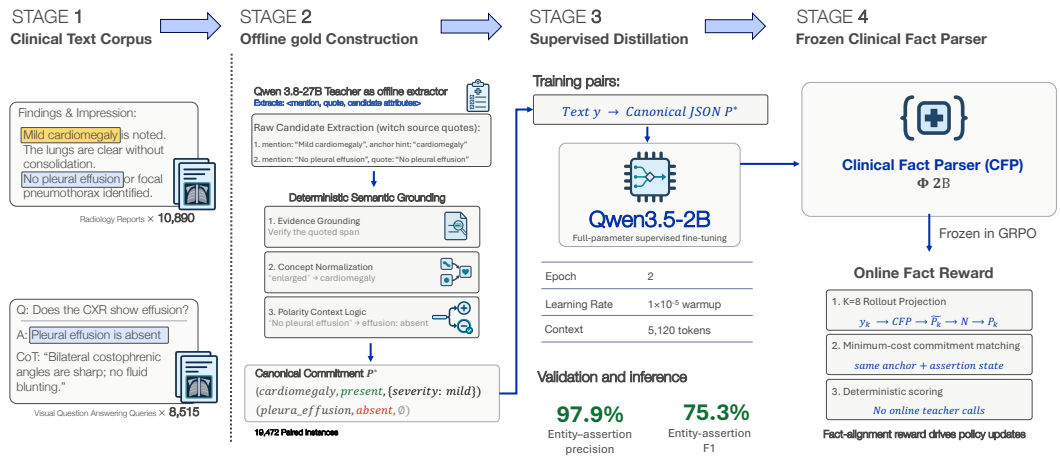


Figure 2: Clinical Fact Parser construction and deployment. Qwen3.8-27B annotations (Qwen Team, 2026c) provide 10,890 report examples, 8,515 question-answer examples, and 67 no-claim controls. The trained parser is frozen to supply commitments for GRPO rewards.

3.2 Asymmetric Fact-Commitment Evaluator

Missing facts, unsupported facts, and attribute errors incur separate costs, fixed before training.

Omission and insertion. Omission and insertion incur fact and attribute costs:

$$D(g) = b_a + \sum_{k \in A_g \cap \mathcal{O}} w_k, \qquad I(p) = \kappa \left(b_a + \sum_{k \in A_p \cap \mathcal{O}} w_k \right), \tag{3}$$

where $b_a > 0$ is the base fact cost (default 1.0), $w_k > 0$ is the frozen axis weight, and $\kappa > 0$ scales insertion relative to omission (production: $\kappa = 1.0$).

Pairing and attribute errors. To keep attributes from offsetting a wrong finding or state, pairing requires:

$$\text{Eligible}(g, p) \iff (a_g = a_p) \wedge (s_g = s_p). \tag{4}$$

For eligible pairs, $C(g,p) = \sum_{k \in (A_g \cup A_p) \cap \mathcal{O}} \ell_k(g,p)$ sums attribute errors:

$$\ell_k(g,p) = \begin{cases} 0, & v_g = v_p \\ w_k, & k \in A_g,\ k \notin A_p \quad \text{(omission)} \\ \kappa w_k, & k \notin A_g,\ k \in A_p \quad \text{(spurious)} \\ (1+\kappa) w_k \big[\tau_k + (1-\tau_k) d_k(v_g, v_p)\big], & k \in A_g \cap A_p,\ v_g \neq v_p \quad \text{(mismatch)} \end{cases} \tag{5}$$

where $d_k(v_g, v_p) \in [0,1]$ is a normalized attribute distance (Appendix A) and $\tau_k \in (0,1]$ is the mismatch penalty floor; admissible configurations require $\tau_k > 1/(1+\kappa)$ (production: $\tau_k = 0.60$).

Matching. One-to-one matching over eligible pairs prevents prediction reuse (Munkres, 1957); unmatched facts incur omission or insertion costs:

$$L(G,P) = \min_{M \subseteq G \times P} \left[\sum_{(g,p) \in M} C(g,p) + \sum_{g \in G \setminus \text{dom}(M)} D(g) + \sum_{p \in P \setminus \text{img}(M)} I(p) \right] \tag{6}$$

Fact reward. We penalize errors left outside matching and normalize by the reference omission cost:

$$R_{\text{fact}}(P, G, \mathcal{O}) = 1 - \frac{L(G,P) + H(P) + U(P) + S(P)}{Z_G}, \tag{7}$$

$$Z_G = \max\left(1, \sum_{g \in G} D(g)\right). \tag{8}$$

Here $H(P) = b_a N_{\text{conflict}}(P)$ penalizes contradictions: the same entity asserted present and absent in overlapping scopes. $U(P) = b_a N_{\text{unresolved}}(P)$ penalizes prediction-side unresolved records from compilation; $S(P) = \kappa b_a N_{\text{unsup_rel}}(P)$ penalizes `suggestive_of` relations without reference support. Each N counts the corresponding events. The reference alone fixes Z_G; rewards are unbounded below.

3.3 Scorer-Level Properties

The costs should discourage unsupported guessing. For fixed references and parameters, the evaluator on structured commitments has three properties (full statements and proofs in Appendix A.1). On a matched eligible pair, honest omission of an active attribute has lower expected loss than guessing in the zero-confidence limit, uniformly over normalized distances including their infimum, exactly when $\tau_k > 1/(1+\kappa)$ (Theorem 1). For a single attribute-free reference and prediction with the same entity, $\kappa \geq 1$, $b_a > 0$, and zero auxiliary penalties, predicting an assertion state different from the reference scores below total omission (Theorem 3). Adding a new commitment with no eligible reference edge strictly lowers reward when auxiliary penalties do not decrease (Theorem 4).

3.4 GRPO Integration

To favor answers with better fact rewards, GRPO (Shao et al., 2024) samples $K = 8$ completions per query from $\pi_{\theta_{\text{old}}}$; we score their final answers, excluding reasoning traces (Figure 3). A standard-deviation floor $\sigma_{\text{floor}} = 0.25$ limits amplification of small reward differences; groups with reward range below 0.15 receive zero advantages. The clipped objective and hyperparameters (ϵ, $\beta = 0$, LoRA rank) appear in Appendix B.2.

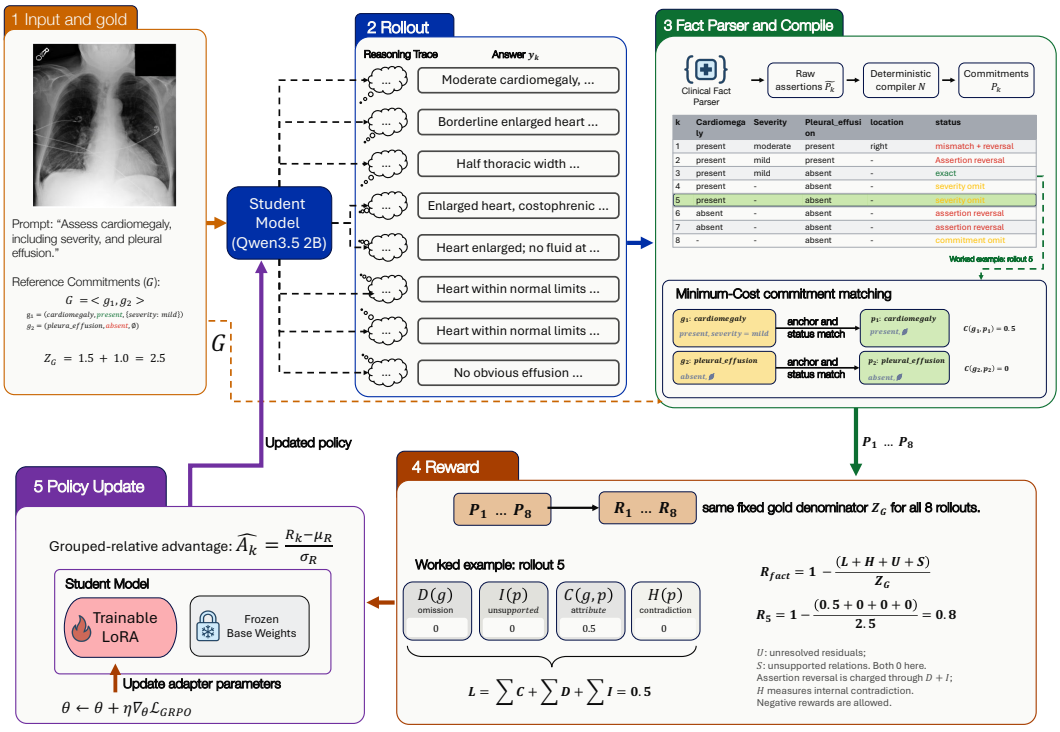


Figure 3: Reward computation for eight GRPO completions sharing reference commitments. In completion 5, omitted severity costs 0.5; the fixed reference denominator is 2.5, giving $R_{\text{fact}} = 0.8$. Group-relative rewards update the policy adapter.

4 Experiments

Shared experimental setup. VeraRad-2B is Qwen3.5-2B (Qwen Team, 2026a) trained for 800 GRPO steps with rank-32 LoRA and frozen checkpoint-2434 CFP, using image-question tasks only.

Evaluation scores final answers, excluding reasoning traces; thinking-enabled configurations have a 1,024-token reasoning budget. Task-specific protocols appear in Appendix C.

4.1 Clinical Fact Parsing and Reward Reliability

We evaluate CFP on 600 responses to 80 tasks across four genres, covering 472 distinct context–response inputs from general-purpose and medical models, VeraRad, and policy checkpoints. QA inputs include questions and options; report inputs contain only text. A human corrects 27B-assisted annotations: $A(y)$ records each response's assertions, including errors; G is independently annotated from benchmark answers or source reports.

Fact Extraction. CFP and Qwen3.5-2B share prompts, schemas, and decoding settings. All 600 responses (3,994 reference facts) are scored, retaining complete schema-valid facts from truncated outputs and discarding unfinished objects and duplicate tuples. Unrecoverable outputs yield no predictions but retain reference facts in recall denominators. Entity–assertion metrics are micro-averaged; bound-attribute F1 uses one-to-one instance matching shared across axes; complete-tuple F1 requires all attributes to match.

Table 1: Fact extraction against the human-reviewed reference.

Model	Extracted facts	Truncated (%)	Entity–assertion precision (%)	Recall (%)	F1 (%)	Bound-attribute F1 (%)	Complete-tuple F1 (%)
Qwen3.5-2B	1,301	7.50	28.13	9.16	13.82	0.31	0.72
CFP	2,498	**0.00**	**97.88**	**61.22**	**75.32**	**35.10**	**31.64**

CFP gains 61.50 percentage points in entity–assertion F1 (Table 1). Precision exceeds recall, making missed commitments the larger extraction gap; bound-attribute F1 remains lower, at 35.10%.

Reward Reliability. Holding G, $\mathcal{O}$, and scoring fixed, we compare rewards from compiled human-reviewed $A(y)$ and CFP outputs. Across 80 groups, CFP preserves 1,035/1,422 unequal-reference-reward orderings (72.78% concordance; predicted ties do not count as concordant) and 463/600 advantage signs (77.17%). Removing attributes from both paths yields 75.97% concordance.

Controlled Changes to Fact Commitments. From 100 reviewed responses across 77 tasks, the structured path scores edited commitments directly; the text path edits responses before CFP parsing (Figure 4). Compiler-input aliases preserve rewards in all 97 eligible seeds across six configurations. Unsupported additions lower reward in 98/98 cases; opposite-state additions do so in 88/92 after excluding qualifier and temporal-scope conflicts. Verbatim repetition of 37 report or multi-fact answers preserves all structured rewards, but only 22/37 after CFP parsing, separating scorer invariance from parsing sensitivity.

4.2 Clinical Visual Question Answering

We test nine configurations on 3,000 questions: 1,200 ReXVQA (Pal et al., 2025), 1,200 MIMIC-Ext-VQA (Bae et al., 2024), and 600 Medical-Diff-VQA (Hu et al., 2023), including 300 temporal questions converted from open answers and paired reports with automated label review. Table 2 compares VeraRad and Qwen3.5 models (Qwen Team, 2026b) in both thinking modes with MedGemma (Sellergren et al., 2025) and CheXagent (Chen et al., 2024); task allocations and decoding are detailed in Appendix C.2.1.

Strict accuracy requires the correct answer in the requested single-label format. **Answer-extraction accuracy** additionally uses a reference-blind deterministic extractor for explicit answer declarations, standalone labels, or exact option text; conflicting or incomplete outputs remain incorrect. Both use the fixed 3,000-question denominator (Appendix C.2.2).

VeraRad-2B (On) leads strict accuracy at 75.53%, gaining 8.83 points overall and 5.50–14.67 across sources over its base in the same mode. Answer extraction changes its score little (75.73%), whereas Qwen3.5-4B (On) leads that measure at 79.13%. Gains in differential diagnosis (+12.00 points vs. base) and conjunction questions (+6.04) concern competing diagnoses and co-occurring findings, consistent with the reward's treatment of assertion state and unsupported commitments (Tables 4–5).

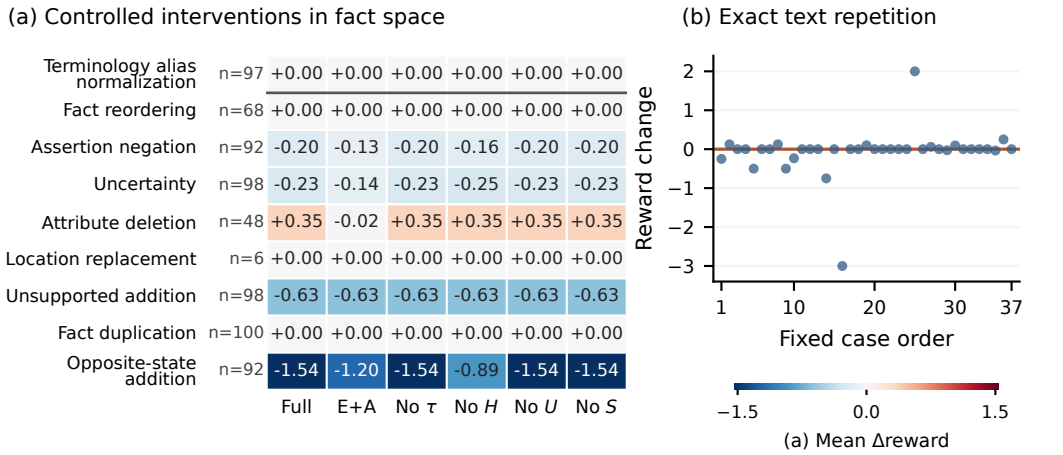


Figure 4: **Reward invariance and sensitivity.** (a) Mean reward change (intervened − original); n: eligible seeds out of 100. Full: complete reward; E+A: entities and assertion states only; No $\tau/H/U/S$: without mismatch floor, contradiction, unresolved, or relation penalty. All share Full's normalization. The terminology alias probe replaces entity names at compiler input (outside CFP's closed output schema) to test normalization invariance, not paraphrase extraction. (b) Complete-reward changes after CFP parses 37 verbatim-repeated answers, in fixed order; the horizontal zero line denotes structured-path invariance.

Table 2: Strict accuracy by source and overall, with answer-extraction accuracy on the full set. Values are in percent. Bold marks the highest value in each column; underline marks the second-highest distinct value. N/A indicates models without a separate thinking mode. ReXVQA and MIMIC-Ext-VQA each contain 1,200 questions; Medical-Diff-VQA contains 600.

Model	Thinking	ReXVQA	MIMIC-Ext	Medical-Diff	Overall	Answer extraction
Qwen3.5-2B	Off	62.50	63.25	13.17	52.93	68.20
Qwen3.5-2B	On	69.92	66.33	61.00	66.70	73.27
Qwen3.5-4B	Off	78.33	67.08	6.67	59.50	74.93
Qwen3.5-4B	On	79.00	68.83	**78.00**	<u>74.73</u>	**79.13**
MedGemma-4B-it	N/A	**82.42**	66.42	59.17	71.37	72.23
MedGemma-27B-it	N/A	14.17	43.58	29.17	28.93	65.03
CheXagent-2-3b	N/A	70.83	**81.67**	63.17	73.63	73.63
VeraRad-2B	Off	77.25	64.67	64.67	69.70	69.77
VeraRad-2B	On	<u>79.17</u>	<u>71.83</u>	<u>75.67</u>	**75.53**	<u>75.73</u>

4.3 Radiology Report Generation

After excluding training-library and question-bank overlap, Stanford ReXRank (Zhang et al., 2025) supplies 2,962 studies; 2,834 have Findings references: 2,186 MIMIC-CXR (Johnson et al., 2019), 590 IU X-Ray (Demner-Fushman et al., 2016), and 58 CheXpert Plus (Chambon et al., 2024). Report generation was absent from training. A stratified 300-study intervention panel (150/100/50) uses distinct patients where identifiable and distinct IU X-Ray case groups. Prompts, decoding, and retries are detailed in Appendix C.3.

Scoring. ReXRank's primary Findings metric is inverse mean RadCliQ-v1, $Q(S) = [|S|^{-1} \sum_{i \in S} \text{RadCliQ-v1}_i]^{-1}$ (Yu et al., 2023). We also report RadGraph Combined F1 (Jain et al., 2021); both are higher-is-better and independent of CFP. Table 7 gives all six metrics.

Full-Cohort Report Quality. VeraRad-2B (On) leads Q on MIMIC-CXR and CheXpert Plus and exceeds its matched base on all three datasets (Table 3a). On IU X-Ray, the base (Off) leads Q, while VeraRad (On) leads RadGraph Combined F1 and RaTEScore. Relative to base, VeraRad's reports are 43.7% shorter with thinking On (44.2% Off), using 35.4% fewer completion tokens On (paired $N = 2,957$). Shorter reports with improved composite scores are consistent with rewarding

Table 3: Report quality and sensitivity to image correspondence. (a) Q: inverse mean RadCliQ-v1; RadGraph Combined F1 is also reported. Bold and underline mark the highest and second-highest distinct quality scores. (b) $\Delta_{\text{specific}} = Q_{\text{correct}} - Q_{\text{shuffled}}$ and $\Delta_{\text{total}} = Q_{\text{correct}} - Q_{\text{no image}}$, computed from unrounded scores on the separate intervention panel. Study counts follow dataset order. Full results appear in Tables 7 and 9.

	MIMIC-CXR		IU X-Ray		CheXpert Plus	
(a) Full-cohort report quality (2,186/590/58 studies)						
Model	$Q \uparrow$	RadGraph Combined F1 $\uparrow$	$Q \uparrow$	RadGraph Combined F1 $\uparrow$	$Q \uparrow$	RadGraph Combined F1 $\uparrow$
Qwen3.5-2B (Off)	0.752	**0.187**	**1.529**	0.263	0.657	0.168
Qwen3.5-2B (On)	0.739	0.179	1.258	<u>0.270</u>	<u>0.680</u>	**0.186**
Qwen3.5-4B (On)	0.735	<u>0.183</u>	1.202	0.257	0.648	0.168
MedGemma-4B-it	0.725	0.163	1.234	0.245	0.628	0.138
MedGemma-27B-it	0.588	0.140	0.888	0.244	0.548	0.147
CheXagent-2-3b	0.613	0.118	0.657	0.118	0.629	0.178
VeraRad-2B (Off)	<u>0.768</u>	0.173	<u>1.455</u>	0.238	0.670	0.169
VeraRad-2B (On)	**0.792**	0.179	1.378	**0.282**	**0.724**	<u>0.184</u>
(b) Image-intervention differences (150/100/50 shared studies)						
Model	Δ_{specific}	Δ_{total}	Δ_{specific}	Δ_{total}	Δ_{specific}	Δ_{total}
Qwen3.5-2B (Off)	+0.1644	+0.1645	−0.0445	+0.5620	+0.1060	+0.1156
Qwen3.5-2B (On)	+0.1556	+0.2662	+0.0837	+0.6950	+0.1053	+0.2066
VeraRad-2B (Off)	+0.1115	+0.0833	−0.1162	−0.1039	+0.0868	+0.0771
VeraRad-2B (On)	+0.1656	+0.3344	+0.0274	+0.8498	+0.1308	+0.2684

supported assertions and penalizing unsupported elaboration; length and checkpoint results appear in Appendix C.3.2 and C.3.4.

Sensitivity to Image Correspondence. Each panel study uses its own image, another study's image from the same dataset (Shuffled), or no image. Donors differ by patient where identifiable and match projection where recorded, without diagnosis-based selection. Table 3b contrasts scores on studies valid across all models and conditions; Table 9 gives all eight models.

VeraRad (On) loses reference agreement after image replacement on MIMIC-CXR and CheXpert Plus, with larger Δ_{specific} than base (On). On IU X-Ray, its own and donor images yield similar scores above the no-image condition; VeraRad (Off) has negative Δ_{specific}.

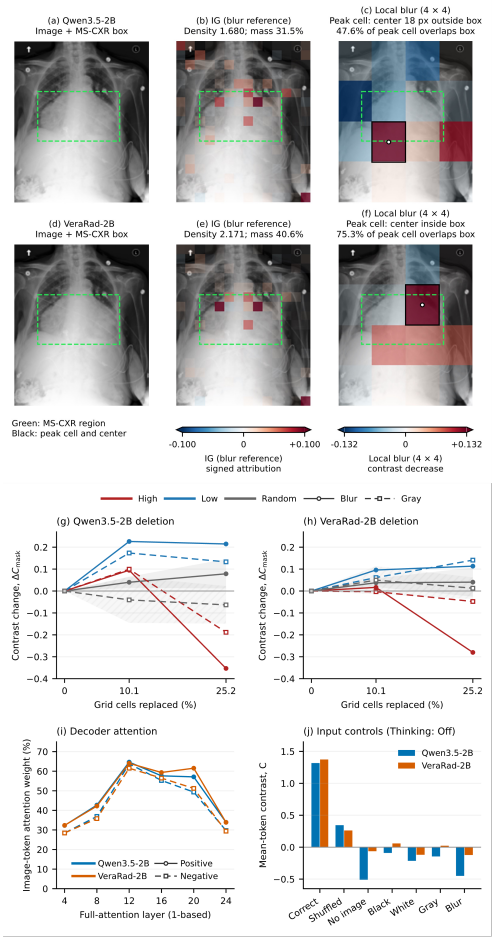
4.4 Visual Evidence Use in a Cardiomegaly Case

In the cardiomegaly case in Figure 5, blurred-reference integrated gradients place 40.6% of VeraRad's positive attribution inside the MS-CXR box versus 31.5% for base (box area: 18.7%). Thinking-Off probes compare length-normalized log-likelihoods of cardiomegaly and its negation; replacing the image reduces VeraRad's positive preference more sharply. Across seven annotated cases in the 32-case pilot, both models' attribution peaks fall inside reference boxes in five. In Thinking-On report generation, VeraRad gives explicit missing-image refusals in 26/32 cases versus 15/32 for base, but five VeraRad outputs assert ungrounded findings without acknowledgment. Its correct-image report identifies cardiomegaly while omitting possible retrocardiac pneumonia and effusion. Appendix C.4 gives case audits and probes.

5 Discussion and Limitations

Our policy comparisons evaluate GRPO with the Fact-Commitment Space reward as a whole. Matched-budget GRPO runs with alternative rewards, such as RadGraph-based scores or language-model judgments, are needed to isolate the contribution of the reward design.

CFP can miss supported facts, overlook unsupported assertions, or misbind attributes, distorting rewards despite the scorer's structured-input guarantees. Valid findings absent from the reference can incur insertion penalties within the scoring scope. Shorter reports do not establish reduced hallucination: the cardiomegaly example omits reference findings.



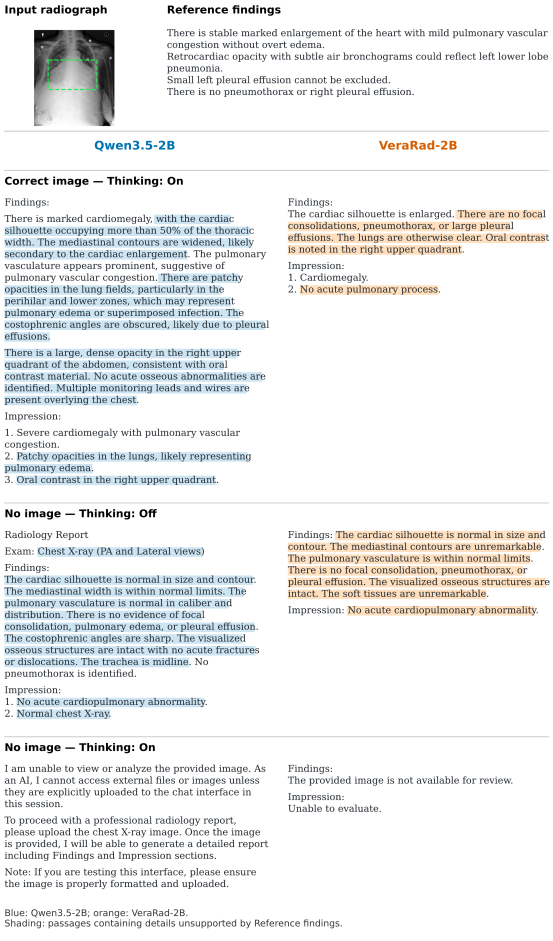


Figure 5: Visual evidence use in MIMIC-CXR 59067739. (a–f) Radiograph, integrated gradients (blurred reference), and local-blur occlusion with shared scales. Green: MS-CXR region; black: highest-drop cell. (g–h) Masking-induced contrast changes (cardiomegaly vs. its negation); random bands: sample SD over five masks. (i) Image-token attention. (j) Input controls. Probes use thinking Off and 238 visual tokens; reports use the original-resolution protocol. Blue/orange mark Qwen/VeraRad passages with reference-unsupported details, not independently adjudicated image errors.

Similar correct- and shuffled-image scores for VeraRad (On) on IU X-Ray indicate limited pairing sensitivity under this metric; donor images can share findings with targets. Transfer beyond chest radiography requires validating and potentially extending the vocabulary, attributes, and parser. Training used $\kappa = 1$; the effects of other insertion-to-omission cost ratios remain untested.

6 Conclusion

Fact-Commitment Space makes the cost of factual errors an explicit design choice for policy learning in chest radiography. It separates fact extraction from deterministic scoring, so reward rules can be examined before training and applied without an online language-model judge. Scorer-level analysis and controlled interventions examine the scoring rules, while human-reviewed evaluation measures how well CFP preserves claims and their rewards. Trained with this reward and GRPO, VeraRad-2B improves clinical VQA over its base and transfers to report generation without report training.

Reproducibility Statement

Ethics Statement

AI Use Statement

Generative AI tools were used to assist with language editing, improving the clarity and readability of the manuscript, LaTeX formatting, and software development and debugging. They were also used for limited assistance with literature search and figure preparation. The research questions, methodological design, experimental setup, analysis of results, and scientific conclusions were developed and validated by the authors. All AI-assisted text, code, figures, and other outputs were reviewed and verified by the authors. The authors take full responsibility for the content of this work.

References

Seongsu Bae, Daeun Kyung, Jaehee Ryu, Eunbyeol Cho, Gyubok Lee, Sunjun Kweon, Jungwoo Oh, Lei Ji, Eric Chang, Tackeun Kim, and Edward Choi. MIMIC-Ext-MIMIC-CXR-VQA: A complex, diverse, and large-scale visual question answering dataset for chest x-ray images, 2024. URL https://doi.org/10.13026/deqx-d943. Version 1.0.0.

Pierre Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven Q. H. Truong, Chu The Chuong, and Curtis P. Langlotz. CheXpert Plus, 2024. URL https://doi.org/10.71718/6nvz-pm34.

Zhihong Chen, Maya Varma, Justin Xu, Magdalini Paschali, Dave Van Veen, Andrew Johnston, et al. A vision-language foundation model to enhance efficiency of chest x-ray interpretation, 2024. URL https://doi.org/10.48550/arXiv.2401.12208.

Jean-Benoit Delbrouck, Pierre Chambon, Christian Bluethgen, Emily Tsai, Omar Almusa, and Curtis Langlotz. Improving the factual correctness of radiology report generation with semantic rewards. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 4348–4360. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.findings-emnlp.319. URL https://doi.org/10.18653/v1/2022.findings-emnlp.319.

Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer Antani, George R. Thoma, and Clement J. McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016. doi: 10.1093/jamia/ocv080. URL https://doi.org/10.1093/jamia/ocv080.

Mark Endo, Rayan Krishnan, Viswesh Krishna, Andrew Y. Ng, and Pranav Rajpurkar. Retrieval-based chest x-ray report generation using a pre-trained contrastive language-image model. In *Proceedings of Machine Learning for Health*, volume 158 of *Proceedings of Machine Learning Research*, pp. 209–219. PMLR, 2021. URL https://proceedings.mlr.press/v158/endo21a.html.

Ziqing Fan, Cheng Liang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. Advancing radiology foundation models with reasoning through step-by-step verification from daily reports. *Communications Medicine*, 6(1), 2026. doi: 10.1038/s43856-026-01654-y. URL https://doi.org/10.1038/s43856-026-01654-y.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. RARR: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16477–16508. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.910. URL https://doi.org/10.18653/v1/2023.acl-long.910.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International*

Conference on Learning Representations, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Xinyue Hu, Lin Gu, Qiyuan An, Mengliang Zhang, Liangchen Liu, Kazuma Kobayashi, Tatsuya Harada, Ronald M. Summers, and Yingying Zhu. Expert knowledge-aware image difference graph representation learning for difference-aware medical visual question answering. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4156–4165. ACM, 2023. doi: 10.1145/3580305.3599819. URL https://doi.org/10.1145/3580305.3599819.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Q. H. Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P. Lungren, Andrew Y. Ng, Curtis P. Langlotz, and Pranav Rajpurkar. RadGraph: Extracting clinical entities and relations from radiology reports. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://openreview.net/forum?id=pMWtc5NKd7V.

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.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pp. 611–626. ACM, 2023. doi: 10.1145/3600006.3613165. URL https://doi.org/10.1145/3600006.3613165.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.153. URL https://doi.org/10.18653/v1/2023.emnlp-main.153.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.741. URL https://doi.org/10.18653/v1/2023.emnlp-main.741.

James Munkres. Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics*, 5(1):32–38, 1957. doi: 10.1137/0105003. URL https://doi.org/10.1137/0105003.

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. GREEN: Generative radiology report evaluation and error notation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 374–390. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.21. URL https://doi.org/10.18653/v1/2024.findings-emnlp.21.

Ankit Pal, Jung-Oh Lee, Xiaoman Zhang, Malaikannan Sankarasubbu, Seunghyeon Roh, Won Jung Kim, Meesun Lee, and Pranav Rajpurkar. ReXVQA: A large-scale visual question answering benchmark for generalist chest x-ray understanding. In *Biocomputing 2026*, pp. 251–264. World Scientific, 2025. doi: 10.1142/9789819824755_0018. URL https://doi.org/10.1142/9789819824755_0018.

Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. *MedVLM-R1: Incentivizing Medical Reasoning Capability of Vision-Language Models via Reinforcement Learning*, pp. 337–347. Springer Nature Switzerland, 2025. doi: 10.1007/978-3-032-04981-0_32. URL https://doi.org/10.1007/978-3-032-04981-0_32.

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*, pp. 311–318. Association for Computational Linguistics, 2002. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

Qwen Team. Qwen3.5-2B model card, 2026a. URL https://huggingface.co/Qwen/Qwen3.5-2B.

Qwen Team. Qwen3.5-4B model card, 2026b. URL https://huggingface.co/Qwen/Qwen3.5-4B.

Qwen Team. Qwen3.8-27B model card, 2026c. URL https://huggingface.co/Qwen/Qwen3.8-27B.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, et al. MedGemma technical report, 2025. URL https://doi.org/10.48550/arXiv.2507.05201.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Ng, and Matthew Lungren. 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)*, pp. 1500–1519. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.117. URL https://doi.org/10.18653/v1/2020.emnlp-main.117.

Stanford AIMI. CheXagent-2-3b model card, 2024. URL https://huggingface.co/StanfordAIMI/CheXagent-2-3b.

Joonho Yang, Seunghyun Yoon, ByeongJeong Kim, and Hwanhee Lee. FIZZ: Factual inconsistency detection by zoom-in summary and zoom-out document. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 30–45. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.3. URL https://doi.org/10.18653/v1/2024.emnlp-main.3.

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. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, 4(9):100802, 2023. doi: 10.1016/j.patter.2023.100802. URL https://doi.org/10.1016/j.patter.2023.100802.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

Xiaoman Zhang, Hong-Yu Zhou, Xiaoli Yang, Oishi Banerjee, Julián N. Acosta, Josh Miller, Ouwen Huang, and Pranav Rajpurkar. 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*, pp. 90–99. PMLR, 2025. URL https://proceedings.mlr.press/v281/zhang25b.html.

Weike Zhao, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Ratescore: A metric for radiology report generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 15004–15019. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.836. URL https://doi.org/10.18653/v1/2024.emnlp-main.836.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36*, pp. 46595–46623. Neural Information Processing Systems Foundation, 2023. doi: 10.52202/075280-2020. URL https://doi.org/10.52202/075280-2020.

APPENDIX

A ADDITIONAL EXPERIMENTAL DETAILS AND REGISTRY ARCHITECTURE

Runtime Registry. The production registry is the frozen v0.4 runtime artifact `runtime_registry.v04.v2.json`. It contains 12,183 active entities, 25 attribute axes, and 358 redirect entries. These counts describe the deployed configuration rather than a separate algorithmic component.

Attribute Distance Formulation. The normalized attribute distance $d_k(v_g, v_p) \in [0, 1]$ used in Equation 5 is defined across nominal and ordinal axes using frozen domain tables:

- **Nominal Axes**: For discrete nominal axes without inherent ordering (e.g., laterality), any mismatch incurs the maximum normalized distance $d_k(v_g, v_p) = 1.0$.
- **Ordinal Axes**: For ordered clinical axes with discrete progression degrees (e.g., severity degrees: mild, moderate, severe), d_k is the normalized rank difference:

$$d_k(v_g, v_p) = \frac{|\mathrm{rank}(v_g) - \mathrm{rank}(v_p)|}{K_{\max} - 1} \tag{9}$$

  where $K_{\max}$ denotes the total number of ordered levels for axis k.

Exact Assignment Solver. To solve the minimum-cost bipartite matching $L(G, P)$ subject to $(g, p) \in M \implies \mathrm{Eligible}(g, p)$, the problem is decomposed into disjoint connected components induced by candidate eligibility edges. For small components with $\max(|G_c|, |P_c|) \leq 5$, the exact minimum is computed via brute-force search over valid injective allocations. For components exceeding this threshold, the exact minimum is solved via the Kuhn–Munkres (Hungarian) algorithm over the cost matrix augmented with omission ($D(g)$) and insertion ($I(p)$) slack variables.

A.1 SCORER-LEVEL PROPERTIES: STATEMENTS AND PROOFS

Theorem 1 (Expected cost of uncertain attribute guessing)**.** *For frozen $\kappa > 0$, $w_k > 0$, and an active axis $k \in A_g \cap \mathcal{O}$ on a matched eligible pair (g, p), if a guessed value $v_p \in \mathcal{V}_k$ is correct with probability $p \in [0, 1]$, honest omission ($k \notin A_p$, cost w_k) yields strictly lower expected attribute loss than guessing ($k \in A_p$) in the zero-confidence limit ($p \to 0$), uniformly over $d_k \in [0, 1]$ including the infimum $d_k \to 0$, if and only if*

$$\tau_k > \frac{1}{1 + \kappa}. \tag{10}$$

Corollary 2. *For $\kappa = 1.0$, the condition requires $\tau_k > 0.5$. Setting $\tau_k = 0.60$ yields a minimum mismatch penalty of $1.20\, w_k > w_k$.*

Proof of Theorem 1. Under Equation 5, omitting attribute k incurs deterministic cost $\mathrm{Cost}_{\mathrm{omit}} = w_k$. Guessing candidate $v_p \in \mathcal{V}_k$ incurs zero loss with probability p and mismatch loss $\ell_k(d_k) = (1 + \kappa) w_k [\tau_k + (1 - \tau_k) d_k]$ with probability $1 - p$. The expected guessing cost is $\mathbb{E}[\mathrm{Cost}_{\mathrm{guess}}] = (1 - p)(1 + \kappa) w_k [\tau_k + (1 - \tau_k) d_k]$.

As $p \to 0$, $\mathbb{E}[\mathrm{Cost}_{\mathrm{guess}}] \to (1 + \kappa) w_k [\tau_k + (1 - \tau_k) d_k]$. A uniform guarantee over the normalized distance range must hold at its infimum as $d_k \to 0$, whether or not a discrete axis attains zero for a mismatch. This gives $(1 + \kappa) w_k \tau_k > w_k \iff \tau_k > \frac{1}{1+\kappa}$. Conversely, this threshold makes the limiting infimum $(1 + \kappa) w_k \tau_k$ strictly greater than w_k, so omission has lower loss throughout the range. □

Theorem 3 (Cost of an assertion with inverted state)**.** *Let $G = \{g\}$ with $g = \langle a, s_g, \emptyset \rangle$, $\kappa \geq 1.0$, $b_a > 0$, and zero auxiliary penalties ($H = U = S = 0$). For prediction $P = \{p\}$ with $p = \langle a, s_p, \emptyset \rangle$ and $s_p \neq s_g$, eligibility fails and*

$$R_{\mathrm{fact}}(\{p\}, G, \mathcal{O}) - R_{\mathrm{fact}}(\emptyset, G, \mathcal{O}) = -\frac{\kappa b_a}{\max(1, b_a)} < 0. \tag{11}$$

Under $b_a = 1.0$, predicting the opposite assertion state yields $R_{\mathrm{fact}} = -\kappa < 0$, strictly below the omission reward of 0.

Proof of Theorem 3. For $G = \{g\}$ with $g = \langle a, s_g, \emptyset \rangle$, the denominator is $Z_G = \max(1, D(g)) = \max(1, b_a)$. Total omission ($P = \emptyset$) leaves g unmatched, yielding $L(G, \emptyset) = D(g) = b_a$ and $R_{\text{fact}}(\emptyset) = 1 - \frac{b_a}{\max(1, b_a)}$.

For candidate $P = \{p\}$ with $p = \langle a, s_p, \emptyset \rangle$ and $s_p \neq s_g$ (opposing assertion state), Equation 4 prevents (g, p) from forming an eligible edge. Thus both commitments remain unmatched, giving $L(G, \{p\}) = D(g) + I(p) = (1 + \kappa) b_a$ and $R_{\text{fact}}(\{p\}) = 1 - \frac{(1+\kappa) b_a}{\max(1, b_a)}$. Since $\kappa \geq 1.0$ and $b_a > 0$,

$$R_{\text{fact}}(\{p\}) - R_{\text{fact}}(\emptyset) = -\frac{\kappa b_a}{\max(1, b_a)} < 0. \tag{12}$$

When $b_a = 1.0$, $Z_G = \max(1, 1.0) = 1.0$, reducing the rewards to $R_{\text{fact}}(\emptyset) = 1 - 1.0 = 0$ and $R_{\text{fact}}(\{p\}) = 1 - (1 + \kappa) = -\kappa < 0$. □

Theorem 4 (Penalty for spurious commitments)**.** *Let $p_{\text{spurious}} \notin P$ be a commitment that cannot form an eligible edge with any $g \in G$, and suppose auxiliary penalties are non-decreasing under insertion (H, U, S each $\geq$ their value on P). Then*

$$R_{\text{fact}}(P \cup \{p_{\text{spurious}}\}, G, \mathcal{O}) < R_{\text{fact}}(P, G, \mathcal{O}). \tag{13}$$

Proof of Theorem 4. Let $p_{\text{spurious}} \notin P$ and $P' = P \cup \{p_{\text{spurious}}\}$. Since p_{spurious} cannot form an eligible edge with any commitment in G under Equation 4, no augmenting path or edge exchange involving p_{spurious} can cover any $g \in G$. The optimal matching cost on G is therefore unchanged, and p_{spurious} is charged as an unmatched insertion:

$$L(G, P') = L(G, P) + I(p_{\text{spurious}}), \quad I(p_{\text{spurious}}) \geq \kappa b_a > 0. \tag{14}$$

With auxiliary penalties satisfying $H(P') \geq H(P)$, $U(P') \geq U(P)$, and $S(P') \geq S(P)$, the numerator in Equation 7 increases by at least $I(p_{\text{spurious}})$. Because $Z_G = \max(1, \sum_{g \in G} D(g))$ depends solely on G, we have

$$R_{\text{fact}}(P', G, \mathcal{O}) - R_{\text{fact}}(P, G, \mathcal{O}) \leq -\frac{I(p_{\text{spurious}})}{Z_G} \leq -\frac{\kappa b_a}{Z_G} < 0. \tag{15}$$

□

B CFP CONSTRUCTION AND POLICY OPTIMIZATION

B.1 CLINICAL FACT PARSING AND DETERMINISTIC COMPILATION

The transformation from free-form text y to canonical commitments P proceeds in two stages: CFP parsing followed by deterministic compilation.

Stage 1: Clinical Fact Parser (Φ_ϕ). The compact, schema-constrained CFP Φ_ϕ parses input text y into an un-normalized candidate set $\tilde{P} = \Phi_\phi(y) = \{\tilde{p}_1, \ldots, \tilde{p}_{|\tilde{P}|}\}$ adhering to the fact-commitment schema (Appendix A).

Stage 2: Deterministic Compiler ($\mathcal{N}$). A stateless deterministic compiler $\mathcal{N}$ compiles candidate set $\tilde{P}$ into canonical commitments P. First, entity names and attribute values are validated against the frozen runtime registry, redirecting aliases and merged concepts to active canonical identifiers. Second, frozen relational tables map certified equivalent entities to canonical representatives, losslessly decompose composite entities into base entities with qualifying attributes (e.g., qualifying an anatomical entity with explicit laterality), and enforce attribute binding compatibility. Finally, duplicate commitments are deduplicated. Any candidate with out-of-vocabulary entities, illegal values, unresolvable ambiguities, or binding conflicts fails closed into an unresolved record; along with extracted inter-entity relations, these auxiliary records are retained for downstream contradiction, unresolved, and relational penalty terms.

The CFP is initialized from Qwen3.5-2B (Qwen Team, 2026a) and fine-tuned on offline text–commitment pairs, then frozen during policy optimization (Figure 2).

B.2 REINFORCEMENT LEARNING INTEGRATION VIA GRPO

We optimize the policy π_θ using Group Relative Policy Optimization (GRPO). For each query $x = (q, G, \mathcal{O})$, the policy generates a group of $K = 8$ completions $\{y_1, \ldots, y_K\} \sim \pi_{\theta_{\text{old}}}(y \mid q)$. Each completion y_i contains a reasoning trace and a final answer. The fact reward $r_i = R_{\text{fact}}(P_i, G, \mathcal{O})$ is computed strictly on the commitments P_i obtained by processing the final answer through the CFP Φ_ϕ and deterministic compiler $\mathcal{N}$; the reasoning trace is retained for audit inspection and does not enter the fact reward computation.

Group-relative advantages are computed over the $K = 8$ scalar rewards:

$$\hat{A}_i = \frac{r_i - \mu_{\{r\}}}{\max\left(\sigma_{\{r\}}, \sigma_{\text{floor}}\right) + \epsilon_{\text{num}}} \tag{16}$$

where $\mu_{\{r\}}$ and $\sigma_{\{r\}}$ denote the empirical mean and standard deviation of $\{r_i\}_{i=1}^K$, with $\sigma_{\text{floor}} = 0.25$ and $\epsilon_{\text{num}} = 10^{-4}$. When the within-group reward range satisfies $\max_j r_j - \min_j r_j < 0.15$, advantages for the entire group are set to zero ($\hat{A}_i = 0$).

Policy parameters θ are updated via the clipped surrogate objective:

$$\mathcal{L}_{\text{GRPO}}(\theta) = \mathbb{E}_{x, \{y_i\}_{i=1}^K \sim \pi_{\theta_{\text{old}}}} \left[\frac{1}{K} \sum_{i=1}^{K} \min\left(\rho_i(\theta)\hat{A}_i, \text{clip}\left(\rho_i(\theta), 1-\epsilon, 1+\epsilon\right)\hat{A}_i \right) \right] \tag{17}$$

where $\rho_i(\theta) = \frac{\pi_\theta(y_i|q)}{\pi_{\theta_{\text{old}}}(y_i|q)}$ denotes the completion-level importance ratio, ϵ is the clipping threshold, and the reference KL penalty is set to zero ($\beta = 0.0$).

Figure 3 illustrates parsing, commitment matching, reward calculation, and policy update for an eight-completion group.

C Extended Experimental Protocols and Results

C.1 Clinical fact parsing and reward reliability

C.1.1 Evaluation setup

We evaluate the Clinical Fact Parser (CFP) on 600 recorded responses to 80 clinical tasks: 20 multiple-choice questions, 20 short-answer questions, 20 multi-fact questions, and 20 report-generation tasks, with eight, eight, five, and nine candidate responses per task, respectively. Responses come from general-purpose and medical models, VeraRad, and intermediate policy checkpoints. The panel contains 472 distinct context–response inputs. Question-answering inputs include the question and options; report inputs contain only the report text.

A local 27B model prepares entity, assertion-state, and attribute annotations, which are reviewed and corrected by a human annotator. The response annotation $A(y)$ records what an answer asserts, including incorrect assertions. The task reference G is annotated separately from the benchmark answer or source report. This shared panel supports both fact-extraction evaluation and evaluation of how CFP parsing affects reward.

C.1.2 Fact extraction and reward preservation

We compare CFP with its Qwen3.5-2B base model using the same prompt, output schema, and decoding settings. Both models are scored against all 600 reviewed responses, containing 3,994 reference facts. We retain complete, schema-valid facts from truncated outputs, discard unfinished objects, and remove repeated fact tuples within each prediction. Unrecoverable outputs contribute no predicted facts, while their reference facts remain in the recall denominator. Entity–assertion precision, recall, and F1 are micro-averaged across responses; bound-attribute F1 uses one-to-one entity-instance matching shared across all attribute axes. Complete-tuple F1 additionally requires matching the full attribute set.

Table 1 shows that CFP improves entity–assertion F1 by 61.50 percentage points over its base model, with gains in both precision and recall. Attribute binding and complete-fact recovery also improve substantially. Its 97.88% precision and 61.22% recall indicate that missed commitments remain the larger entity-extraction gap. The base model reaches the output limit in 45 responses; 38 still yield complete facts under the shared extraction rule.

For reward evaluation, let $P_{\mathrm{ref}} = \mathcal{N}(A(y))$ denote the compiled human-reviewed response commitments and $P = \mathcal{N}(\Phi_\phi(y))$ the compiled CFP commitments. We hold the reference commitments G for task q, the observability mask $\mathcal{O}$, and the scoring rule fixed, and compare $R_{\mathrm{fact}}(P_{\mathrm{ref}}, G, \mathcal{O})$ with $R_{\mathrm{fact}}(P, G, \mathcal{O})$. Pairwise concordance measures preserved ordering among candidates with unequal reference rewards; a projected tie does not count as preserved ordering. We also measure agreement in the sign of the candidate-group advantage.

Across 80 candidate groups, projected rewards preserve 1,035 of 1,422 non-tied reference orderings (72.78%) and 463 of 600 advantage signs (77.17%). Using only entities and assertion states in both reward paths yields 75.97% concordance. Removing the contradiction penalty, relation penalty, mismatch penalty floor, or unresolved-commitment penalty individually yields concordance between 72.57% and 72.91%. These comparisons locate attribute-related discrepancies that remain after entity extraction improves.

C.1.3 Controlled changes to fact commitments

We select 100 reviewed responses from 77 tasks and construct controlled changes to their fact commitments. The structured path scores the edited commitments directly against the original task reference, isolating the scoring rule from CFP parsing. The natural-language path applies corresponding edits to the response text before CFP parsing.

A separate terminology-normalization probe replaces one entity name with a vocabulary alias accepted at the compiler input. Of the 100 seeds, 97 admit such a replacement; all preserve the canonical commitments and rewards across six scoring configurations (100% invariance). This verifies the structural normalization step.

Figure 4a shows that direct scoring is unchanged after duplicating a fact in all 100 cases and reordering the fact list in all 68 applicable cases. Adding an entity absent from both the original commitments and the task reference reduces reward in all 98 applicable cases, with a mean change of -0.6251. After excluding identified qualifier and temporal-scope conflicts, adding an opposite-state commitment lowers reward in 88 of 92 cases and leaves four unchanged. The mean change is -1.5365 with the complete reward and -0.8890 without the contradiction component, isolating the additional penalty from explicit contradiction handling.

To test whether duplication invariance survives CFP parsing, we append an exact copy of each of 37 report or multi-fact answers. The intended commitments retain the same reward in all cases; after CFP parsing, reward remains unchanged in 22 of 37, with a mean paired change of -0.0687. Entity–assertion F1 after compilation changes from 70.03% to 64.47%. The contrast between the zero-change reference and the observed changes in Figure 4b separates the scorer's duplication invariance from CFP's sensitivity to repeated text.

C.2 Clinical visual question answering

C.2.1 Experimental setup

We evaluate nine model configurations on 3,000 chest radiograph questions: 1,200 from ReXVQA (Pal et al., 2025), 1,200 from MIMIC-Ext-VQA (Bae et al., 2024), and 600 from Medical-Diff-VQA (Hu et al., 2023). ReXVQA uses four-option questions about clinical findings, differential diagnosis, anatomy, and geometry. MIMIC-Ext-VQA tests yes/no verification, including conjunctions and disjunctions of findings. Medical-Diff-VQA includes projection classification (270 questions), finding presence (30), and temporal change verification (300). The temporal questions were converted from original open-ended questions and paired reports; their labels underwent automated review.

The comparison includes VeraRad-2B, Qwen3.5-2B and Qwen3.5-4B (Qwen Team, 2026a;b), MedGemma-4B-it and MedGemma-27B-it (Sellergren et al., 2025), and CheXagent-2-3b (Chen et al., 2024); VeraRad and both Qwen models are tested with thinking enabled and disabled. The models receive the same questions and images without reference labels or reports.

Sampling and decoding. Source and category quotas were set during development comparisons. ReXVQA and Medical-Diff-VQA contribute at most one question per patient; MIMIC-Ext-VQA contributes at most three. Patients may appear in both MIMIC-derived sources. Each yes/no category is label-balanced, as are ReXVQA option positions (300 per letter) and AP/PA projections (135 each).

All questions come from source test partitions except 95 derived temporal questions from validation. The temporal questions use original question-answer pairs and current and prior reports, with automated label review. Temporal category names describe the reference change in the image pair; negative questions can ask about a different proposed change.

The eight configurations using the common prompt are asked for one choice letter, yes/no, or an AP/PA label. ReXVQA retains all supplied views, and Medical-Diff-VQA presents the current image before the prior image. Decoding uses temperature 0, a 4,096-token output limit, and a 1,024-token thinking budget when enabled. Failed or incomplete requests allow up to three retries; the first normally completed answer is scored, regardless of its correctness or format.

CheXagent uses the authors' published chat template (Stanford AIMI, 2024), adapted here with bracketed choices and constrained option decoding; its choices are mapped to the common labels. MedGemma-4B-it reuses verified predictions on the same questions and protocol.

C.2.2 Metrics

Strict accuracy is the primary metric: a question counts as correct only when the model follows the requested single-label format and matches the reference answer. Accepted labels are A-D, yes/no, or an AP/PA projection; case, surrounding whitespace, and predefined projection aliases are normalized. Responses that do not complete normally or do not give a valid label count as incorrect. Overall accuracy divides correct answers by 3,000, while each source score uses its own question count.

Models sometimes supply a clear answer inside a sentence despite the instruction to output one label. We therefore also report answer-extraction accuracy on the same responses. A deterministic extractor

keeps all strictly valid answers and reads explicit answer declarations, standalone labels, leading yes/no answers, projection declarations, or exact option text from otherwise noncompliant completed responses; for example, "The answer is B." yields B. It receives no reference labels. Only final-answer text outside thinking blocks is evaluated. Conflicting answers, unclosed thinking tags, incomplete responses, and unresolved outputs remain incorrect, with the same 3,000-question denominator.

C.2.3 RESULTS

Table 2 reports accuracy in percent. Bold marks the highest value in each column; underline marks the second-highest distinct value. N/A indicates models without a separate thinking mode.

With thinking enabled, VeraRad-2B has the highest overall strict accuracy (75.53%), ahead of Qwen3.5-4B (74.73%) and CheXagent (73.63%). It exceeds its Qwen3.5-2B base model in the same mode by 8.83 percentage points; enabling thinking raises VeraRad's score by 5.83 points. The three sources have different leaders: MedGemma-4B-it on ReXVQA, CheXagent on MIMIC-Ext-VQA, and Qwen3.5-4B with thinking on Medical-Diff-VQA. With thinking enabled in both models, VeraRad ranks second on every source and outperforms its Qwen3.5-2B base model by 5.50 to 14.67 points.

Compared with its Qwen3.5-2B base model, VeraRad with thinking gains 12.00 points on differential diagnosis and 6.04 points on conjunction questions (Tables 4 and 5). These tasks probe competing diagnoses and co-occurring findings, closely matching the Fact-Commitment Space reward's treatment of assertion state, attributes, and unsupported commitments. VeraRad's score changes little under answer extraction (75.53% to 75.73%); Qwen3.5-4B with thinking gains 4.40 points to 79.13% and leads on that secondary measure as explicit answers outside the requested single-label format are credited.

C.2.4 FINE-GRAINED STRICT ACCURACY

Tables 4–6 report strict accuracy by category in percent. Bold and underline mark the best and second-best distinct scores; ties share the same mark.

Table 4: ReXVQA: four-option clinical questions.

Model	Thinking	Differential diagnosis ($n = 400$)	Finding exclusion ($n = 450$)	Finding presence ($n = 200$)	Localization ($n = 120$)	Geometry ($n = 30$)
Qwen3.5-2B	Off	53.25	65.11	75.50	64.17	53.33
Qwen3.5-2B	On	67.00	80.00	64.50	60.83	30.00
Qwen3.5-4B	Off	74.50	85.33	<u>77.50</u>	70.00	63.33
Qwen3.5-4B	On	73.00	<u>88.22</u>	76.00	72.50	<u>66.67</u>
MedGemma-4B-it	N/A	<u>78.00</u>	**90.67**	**80.00**	**75.83**	60.00
MedGemma-27B-it	N/A	6.75	23.33	10.50	14.17	0.00
CheXagent-2-3b	N/A	63.25	80.89	66.50	70.00	53.33
VeraRad-2B	Off	74.25	79.78	**80.00**	<u>74.17</u>	**73.33**
VeraRad-2B	On	**79.00**	85.78	74.00	70.00	53.33

Table 5: MIMIC-Ext-VQA: clinical yes/no verification.

Model	Thinking	Conjunction ($n = 480$)	Disjunction ($n = 600$)	Other clinical verification ($n = 120$)
Qwen3.5-2B	Off	55.62	69.83	60.83
Qwen3.5-2B	On	65.42	68.33	60.00
Qwen3.5-4B	Off	68.96	64.67	71.67
Qwen3.5-4B	On	67.50	70.33	66.67
MedGemma-4B-it	N/A	61.46	69.00	<u>73.33</u>
MedGemma-27B-it	N/A	41.46	44.00	50.00
CheXagent-2-3b	N/A	**84.38**	**80.00**	**79.17**
VeraRad-2B	Off	57.50	70.67	63.33
VeraRad-2B	On	<u>71.46</u>	<u>72.50</u>	70.00

Table 6: Medical-Diff-VQA: projection, finding presence, and temporal verification.

Model	Thinking	AP/PA ($n = 270$)	Finding presence ($n = 30$)	Appearance ($n = 50$)	Resolution ($n = 50$)
Qwen3.5-2B	Off	0.00	43.33	10.00	0.00
Qwen3.5-2B	On	61.85	73.33	30.00	50.00
Qwen3.5-4B	Off	0.74	20.00	12.00	6.00
Qwen3.5-4B	On	**85.93**	76.67	**66.00**	**72.00**
MedGemma-4B-it	N/A	66.67	**86.67**	<u>48.00</u>	50.00
MedGemma-27B-it	N/A	0.00	80.00	40.00	48.00
CheXagent-2-3b	N/A	74.44	<u>83.33</u>	38.00	32.00
VeraRad-2B	Off	<u>79.63</u>	73.33	42.00	56.00
VeraRad-2B	On	**85.93**	76.67	44.00	<u>62.00</u>

Model	Thinking	Persistence without change ($n = 160$)	Worsening ($n = 20$)	Improvement ($n = 20$)
Qwen3.5-2B	Off	30.63	35.00	25.00
Qwen3.5-2B	On	75.00	45.00	40.00
Qwen3.5-4B	Off	14.37	0.00	0.00
Qwen3.5-4B	On	<u>76.25</u>	<u>60.00</u>	<u>50.00</u>
MedGemma-4B-it	N/A	50.62	45.00	<u>50.00</u>
MedGemma-27B-it	N/A	58.13	30.00	40.00
CheXagent-2-3b	N/A	56.25	**65.00**	**75.00**
VeraRad-2B	Off	51.25	50.00	<u>50.00</u>
VeraRad-2B	On	**78.12**	<u>60.00</u>	45.00

C.3 Radiology report generation: report quality and sensitivity to image correspondence

C.3.1 Experimental setup

We evaluate report quality on 2,962 studies from the Stanford ReXRank (Zhang et al., 2025) test sets after excluding studies overlapping the training reference library or question bank: 2,186 MIMIC-CXR (Johnson et al., 2019), 590 IU X-Ray (Demner-Fushman et al., 2016), and 186 CheXpert Plus (Chambon et al., 2024) studies. Of these, 2,834 have Findings references (2,186, 590, and 58, respectively). A stratified panel of 300 studies from this cohort (150/100/50) tests sensitivity to correct, shuffled, and absent images. Panel studies have distinct patients where patient identifiers are available; IU X-Ray uses distinct case groups because patient linkage is unavailable.

VeraRad-2B is the Qwen3.5-2B policy after 800 steps of Fact-Commitment Space GRPO, using rank-32 LoRA (Hu et al., 2022) and the checkpoint-2434 CFP. GRPO used image-question tasks, with no complete-report generation task. We evaluate the final policy with thinking enabled (On) and disabled (Off). Comparators are Qwen3.5-2B in both modes, Qwen3.5-4B with thinking enabled, MedGemma-4B-it, MedGemma-27B-it, and CheXagent-2-3b. On uses a 1,024-token thinking budget.

All models receive one selected chest radiograph per study and the following prompt. Reference reports and clinical context stored in the manifests are not included in this prompt.

> You are an expert radiologist. Based on the provided chest X-ray image(s), write a complete radiology report with Findings and Impression sections. Output only the report text.

The full-cohort comparison uses temperature zero, a 4,096-token output budget for vLLM (Kwon et al., 2023) models, and CheXagent's native 512-token limit. For thinking-enabled Qwen and VeraRad configurations, reasoning trace and final text share the 4,096-token limit. We regenerated 27 length-limited, reference-eligible reports at each model's original limit with repetition penalty 1.1; all other outputs remain as first generated. Models retain their native image processors and chat templates; CheXagent uses its dedicated inference adapter. Scoring uses final-answer text.

Scoring and aggregation. We use ReXRank's Findings evaluation and its primary ranking metric, inverse mean RadCliQ-v1. RadCliQ combines report-similarity measures and was developed to track radiologist-rated errors more closely than individual metrics (Yu et al., 2023); ReXRank reports its

reciprocal so that higher is better (Zhang et al., 2025). For a scored cohort S,

$$Q(S) = \left[\frac{1}{|S|} \sum_{i \in S} \text{RadCliQ-v1}_i \right]^{-1}.$$

We also report RadGraph Combined F1, RaTEScore, BLEU (Papineni et al., 2002), SembScore (Endo et al., 2021), and BERTScore (Zhang et al., 2020). All six metrics are higher-is-better; datasets are reported separately.

ReXRank compares the complete generated report with the reference Findings text. The evaluation metrics are independent of the training CFP. All 2,834 studies with Findings references contribute to quality scoring for every model.

C.3.2 Full-cohort report quality

Table 7 reports full-cohort report quality. The primary metric Q is inverse mean RadCliQ-v1. Scores use the 2,186 MIMIC-CXR, 590 IU X-Ray, and 58 CheXpert Plus studies with Findings references. Bold marks the highest and underline the second-highest distinct value in each column (ties share the mark).

Table 7: Full-cohort report quality across three datasets.

Model	Q ↑	RadGraph Combined F1 ↑	RaTEScore ↑	BLEU ↑	SembScore ↑	BERTScore ↑
MIMIC-CXR (2,186 studies)						
Qwen3.5-2B (Off)	0.752	**0.187**	0.570	**0.118**	0.370	0.315
Qwen3.5-2B (On)	0.739	0.179	<u>0.572</u>	0.111	0.373	0.304
Qwen3.5-4B (On)	0.735	<u>0.183</u>	**0.578**	0.111	<u>0.379</u>	0.294
MedGemma-4B-it	0.725	0.163	0.557	0.109	0.338	0.322
MedGemma-27B-it	0.588	0.140	0.525	0.083	0.296	0.170
CheXagent-2-3b	0.613	0.118	0.454	0.043	0.296	0.228
VeraRad-2B (Off)	<u>0.768</u>	0.173	0.542	<u>0.116</u>	0.348	**0.354**
VeraRad-2B (On)	**0.792**	0.179	0.561	0.110	**0.392**	<u>0.348</u>
IU X-Ray (590 studies)						
Qwen3.5-2B (Off)	**1.529**	0.263	0.667	<u>0.226</u>	**0.642**	<u>0.503</u>
Qwen3.5-2B (On)	1.258	<u>0.270</u>	0.669	0.202	0.599	0.438
Qwen3.5-4B (On)	1.202	0.257	0.650	0.193	0.594	0.428
MedGemma-4B-it	1.234	0.245	<u>0.678</u>	0.203	<u>0.628</u>	0.430
MedGemma-27B-it	0.888	0.244	0.638	0.142	0.620	0.245
CheXagent-2-3b	0.657	0.118	0.537	0.087	0.388	0.240
VeraRad-2B (Off)	<u>1.455</u>	0.238	0.639	**0.244**	0.606	**0.523**
VeraRad-2B (On)	1.378	**0.282**	**0.683**	0.223	0.617	0.460
CheXpert Plus (58 studies)						
Qwen3.5-2B (Off)	0.657	0.168	0.540	<u>0.118</u>	0.328	0.238
Qwen3.5-2B (On)	<u>0.680</u>	**0.186**	<u>0.546</u>	<u>0.118</u>	0.352	0.242
Qwen3.5-4B (On)	0.648	0.168	**0.558**	0.109	0.337	0.220
MedGemma-4B-it	0.628	0.138	0.519	0.105	0.294	0.238
MedGemma-27B-it	0.548	0.147	0.498	0.098	0.287	0.095
CheXagent-2-3b	0.629	0.178	0.503	0.065	<u>0.360</u>	0.170
VeraRad-2B (Off)	0.670	0.169	0.517	**0.122**	0.283	**0.282**
VeraRad-2B (On)	**0.724**	<u>0.184</u>	0.545	0.117	**0.382**	<u>0.278</u>

VeraRad-2B (On) leads the primary metric on MIMIC-CXR and CheXpert Plus and exceeds the matched Qwen3.5-2B (On) baseline on all three datasets. On IU X-Ray, Qwen3.5-2B (Off) leads the primary metric, while VeraRad-2B (On) leads RadGraph Combined F1 and RaTEScore. On MIMIC-CXR, Qwen3.5-2B (Off) leads RadGraph Combined F1 and Qwen3.5-4B (On) leads RaTEScore.

Report length and generation cost. Table 8 reports paired means over 2,957 studies completed by all four configurations under the initial decoding settings. Completion tokens include reasoning trace and final text.

VeraRad's final reports are 43.7% shorter with thinking On and 44.2% shorter with thinking Off than the corresponding base-model reports. On uses 35.4% fewer completion tokens. The reward

Table 8: Report length and generation cost on the original paired 2,957 studies. SD denotes the standard deviation. Completion tokens include reasoning trace and final text.

Policy	Report characters, mean (SD)	Reasoning characters, mean	Completion tokens, mean
Qwen3.5-2B, thinking Off	577.3 (208.6)	—	133.1
VeraRad-2B, thinking Off	322.3 (113.7)	—	75.8
Qwen3.5-2B, thinking On	613.0 (253.1)	3,677.4	1,092.7
VeraRad-2B, thinking On	345.0 (108.2)	2,356.6	705.4

penalizes unsupported commitments; favoring well-supported assertions over speculative elaboration is one explanation consistent with the shorter reports and improved composite scores on this untrained report format. Appendix C.3.4 examines report changes across checkpoints.

Five full-cohort responses originally hit the generation limit. Repeating those five with `repetition_penalty=1.1` and the same 4,096-token budget gave five normal completions. The resulting 2,962-study mean report lengths are 577.6/322.5 characters for Qwen/VeraRad-2B (Off) and 613.0/345.0 for On. These five regenerated responses complete the length analysis. The quality scores above follow the separate evaluation protocol in Appendix C.3, which regenerated 27 length-limited, reference-eligible reports across the full model comparison before scoring. The primary length table retains the original paired 2,957-study predictions.

Effect of enabling thinking. With thinking enabled, VeraRad improves RadGraph Combined F1, RaTEScore, and SembScore on all three datasets. Its primary score rises on MIMIC-CXR and CheXpert Plus but falls on IU X-Ray; BLEU and BERTScore favor Off throughout. Enabling thinking therefore changes the report profile rather than uniformly improving every metric.

C.3.3 Sensitivity to image correspondence

To test whether report quality depends on the image belonging to the reported study, we evaluate each panel case with its own radiograph (Correct), a radiograph from another case in the same dataset (Shuffled), or without an image (No image). The prompt and reference remain fixed. Donors come from different patients where identifiers are available and match projection where recorded; they are not selected by diagnosis. For example, a case whose reference mentions cardiomegaly receives another case's radiograph, which may also show cardiomegaly.

For each dataset, let S be the intersection of studies with valid scored outputs across every model and condition. Using the same S, we compute

$$\Delta_{\text{specific}} = Q_{\text{correct}}(S) - Q_{\text{shuffled}}(S), \qquad \Delta_{\text{total}} = Q_{\text{correct}}(S) - Q_{\text{no image}}(S),$$

$$\Delta_{\text{presence}} = Q_{\text{shuffled}}(S) - Q_{\text{no image}}(S), \qquad \Delta_{\text{total}} = \Delta_{\text{specific}} + \Delta_{\text{presence}}.$$

A positive Δ_{specific} means cohort-level report agreement with the target references is higher with the studies' own images than with donor images. It measures dependence on study-image pairing, not on a deliberately reversed finding. Δ_{total} compares the correct image with no image; Δ_{presence} separates the contribution of an in-domain but unmatched image.

The panel uses the first complete response under a bounded retry protocol (4,096 tokens, then up to 8,192 tokens with repetition penalty 1.0 or 1.1). This yields valid outputs for every model and condition on the same 150/100/50 studies; 94 of 7,200 outputs required a retry.

In Table 9, VeraRad-2B (On) has the highest correct-image score on MIMIC-CXR and CheXpert Plus, and replacing its image with a donor lowers reference agreement. Its report fidelity thus depends in part on study-specific image correspondence, beyond image presence alone. On IU X-Ray, correct and donor images yield similar scores while both exceed the no-image condition: image presence matters more than the particular pairing on this panel. VeraRad-2B (On) also has a larger pairing contrast than Qwen3.5-2B (On) on MIMIC-CXR and CheXpert Plus, consistent with Fact-Commitment Space training's reward for supported clinical assertions and penalty for unsupported ones in image-question tasks.

Table 9: Report quality under image interventions across three datasets. Correct, Shuffled, and No image are inverse mean RadCliQ-v1 scores on the same studies. Differences use unrounded scores. Bold and underline mark the two highest distinct values for Correct and the two contrasts.

Model	Correct	Shuffled	No image	Δ_{specific}	Δ_{total}
MIMIC-CXR (150 shared studies)					
Qwen3.5-2B (Off)	<u>0.7768</u>	0.6125	0.6124	<u>+0.1644</u>	+0.1645
Qwen3.5-2B (On)	0.7651	0.6094	0.4989	+0.1556	+0.2662
Qwen3.5-4B (On)	0.7492	0.5929	0.4528	+0.1564	+0.2964
MedGemma-4B-it	0.7545	0.6266	0.4571	+0.1279	<u>+0.2974</u>
MedGemma-27B-it	0.6022	0.5563	0.4537	+0.0460	+0.1485
CheXagent-2-3b	0.6025	0.5183	0.5075	+0.0842	+0.0950
VeraRad-2B (Off)	0.7720	0.6605	0.6887	+0.1115	+0.0833
VeraRad-2B (On)	**0.8098**	0.6442	0.4754	**+0.1656**	**+0.3344**
IU X-Ray (100 shared studies)					
Qwen3.5-2B (Off)	**1.4410**	1.4854	0.8789	-0.0445	+0.5620
Qwen3.5-2B (On)	1.2404	1.1567	0.5454	**+0.0837**	+0.6950
Qwen3.5-4B (On)	1.1942	1.1132	0.4877	<u>+0.0810</u>	+0.7065
MedGemma-4B-it	1.2339	1.1625	0.5063	+0.0714	<u>+0.7275</u>
MedGemma-27B-it	0.9076	0.8616	0.5050	+0.0459	+0.4026
CheXagent-2-3b	0.6624	0.6129	0.4672	+0.0496	+0.1953
VeraRad-2B (Off)	1.3420	1.4581	1.4458	-0.1162	-0.1039
VeraRad-2B (On)	<u>1.3589</u>	1.3315	0.5091	+0.0274	**+0.8498**
CheXpert Plus (50 shared studies)					
Qwen3.5-2B (Off)	0.6595	0.5536	0.5439	+0.1060	+0.1156
Qwen3.5-2B (On)	<u>0.6598</u>	0.5545	0.4531	+0.1053	<u>+0.2066</u>
Qwen3.5-4B (On)	0.6298	0.5431	0.4390	+0.0868	+0.1908
MedGemma-4B-it	0.6264	0.5547	0.4488	+0.0717	+0.1776
MedGemma-27B-it	0.5402	0.4943	0.4470	+0.0459	+0.0932
CheXagent-2-3b	0.6369	0.5010	0.4704	**+0.1359**	+0.1665
VeraRad-2B (Off)	0.6516	0.5648	0.5745	+0.0868	+0.0771
VeraRad-2B (On)	**0.7102**	0.5794	0.4418	<u>+0.1308</u>	**+0.2684**

C.3.4 Policy evolution in radiology report generation

Fact-Commitment Space Group Relative Policy Optimization (GRPO) trains VeraRad-2B on 8,500 image-question items: single-choice questions, focused questions, open enumeration, and verification. The training bank contains no complete-report generation task. We examine whether the resulting policy changes its reporting behavior on a separate chest-radiograph task, comparing it with the Qwen3.5-2B base model and tracing four intermediate checkpoints.

A fixed-case checkpoint trace We then generated reports for a frozen 32-case panel using the base policy and checkpoints 200, 400, 600, and 800. Every case kept the same image and report prompt across checkpoints, with thinking On, temperature zero, 4,096 total output tokens, 1,024 thinking tokens, and repetition penalty 1.1. All 160 generations completed normally. The panel was selected for an earlier visual-grounding study: eight cases each from MIMIC-CXR, IU X-Ray, and CheXpert Plus, and eight MIMIC/MS-CXR overlap cases. One case shares a patient with a GRPO training study. This exploratory panel is used to inspect policy evolution, not to estimate cohort-wide clinical quality.

Mean final-report length falls from 788.8 characters at the base policy to 444.4 at checkpoint 200 and 420.4 at checkpoint 800. At step 200, 31 of 32 reports are already shorter than their base counterparts; from step 400 onward, all 32 are shorter. The reasoning trace follows a different trajectory: its mean length falls to 2,397.6 characters at step 600, then rises to 2,599.8 at step 800. The same fixed panel therefore shows early compression of the final report without a monotonic shortening of reasoning.

In study 51288835 (Figure 6), the base report contains 969 characters; checkpoints 200, 400, 600, and 800 contain 387, 369, 485, and 462. All five reports retain cardiomegaly, small effusions, and interstitial edema. The final two checkpoints describe the edema as **mild-to-moderate**, compared with **mild** at the first three stages; the reference suspects moderate-to-severe edema. The selected reasoning excerpts continue to discuss the same findings across checkpoints. The visible shift is a



Reference findings — MIMIC-CXR report text (the report prompt contained the image only)

- There are ___ B lines and small effusions are still present.
- In combination with the obvious cardiomegaly, moderate-to-severe interstitial pulmonary edema is to be suspected.
- This is mainly caused by an increase in interstitial structures and subtle alveolar opacities.

	Thinking (CoT) — verbatim excerpts	Final report — verbatim excerpts
Base Qwen3.5-2B 969 chars	There is diffuse interstitial prominence throughout both lung fields. This suggests pulmonary edema or chronic interstitial disease. Given the heart size, edema is more likely. […] The cardiac silhouette appears enlarged (cardiomegaly). However, this is an AP portable film, which magnifies the heart.	There is evidence of cardiomegaly. The cardiac silhouette is enlarged. […] There are diffuse interstitial markings throughout the lung fields, suggestive of mild pulmonary edema and vascular congestion. […] small bilateral pleural effusions. **Editor note:** *969-char template-style report also covers technique, sternal wires, aortic knob and shoulder joints; the CoT calls cardiomegaly 'possible but borderline' on an AP film, while the reference calls it 'obvious'.*
Checkpoint 200 387 chars	There are diffuse interstitial markings throughout both lung fields. This suggests pulmonary edema or chronic interstitial disease. […] There is increased opacity in the right lower lung zone, obscuring the right hemidiaphragm. This could be consolidation or pleural effusion.	The cardiac silhouette is enlarged. […] There are small bilateral pleural effusions. […] There is mild interstitial prominence suggestive of mild pulmonary edema. **Editor note:** *The CoT weighs a right-lower-zone opacity ('consolidation or pleural effusion'); the final report states 'No focal airspace opacities are identified.'*
Checkpoint 400 369 chars	There are diffuse interstitial markings throughout both lungs. […] The costophrenic angles are obscured, suggesting pleural effusions.	The cardiac silhouette is enlarged. […] Small bilateral pleural effusions are present. […] 2. Mild pulmonary edema. **Editor note:** *The reference confirms small effusions but does not specify laterality; 'bilateral' is shown neutral (not counted as an error) in every row.*
Checkpoint 600 485 chars	There are diffuse interstitial markings throughout both lung fields. This suggests pulmonary edema or chronic interstitial disease. […] Pulmonary vascular congestion and interstitial prominence consistent with mild-to-moderate pulmonary edema.	The cardiac silhouette is enlarged. […] There are diffuse interstitial opacities and perihilar haze consistent with mild-to-moderate pulmonary edema. […] Small bilateral pleural effusions are present. **Editor note:** *Edema wording shifts 'mild' → 'mild-to-moderate': closer to the reference's moderate-to-severe, but still below it.*
Checkpoint 800 462 chars	There are diffuse interstitial markings throughout both lung fields. This suggests pulmonary edema or chronic interstitial disease. […] Pulmonary vascular congestion and interstitial prominence consistent with mild-to-moderate pulmonary edema.	The cardiac silhouette is enlarged. […] There is pulmonary vascular congestion and interstitial prominence suggestive of mild-to-moderate pulmonary edema. […] Small bilateral pleural effusions are present. **Editor note:** *The two CoT excerpts quoted here repeat checkpoint 600 word-for-word (the diffuse-markings sentence is already identical at 200).*

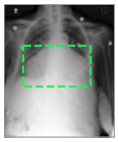
■ reference-aligned ■ beyond / against reference […] omitted text *gray italic* = editor note (not model text)

Figure 6: A five-stage report trace for MIMIC-CXR study 51288835. The base model and checkpoints 200, 400, 600, and 800 receive the same radiograph and generation settings. Each row pairs verbatim reasoning and final-report excerpts; highlights identify reference-aligned findings and the reported edema severity. The reference describes cardiomegaly, small effusions, and suspected moderate-to-severe interstitial edema.

more compact final report with its principal findings preserved, accompanied by a partial adjustment in severity wording.

The second five-stage trace (Figure 7) for study 59067739 shows a different trajectory. The base model already reports cardiomegaly and mild vascular congestion. Checkpoints 200 and 600 add bilateral effusions despite the reference's explicit absence of a right effusion; 400 and 600 turn congestion into a pulmonary-edema diagnosis, a stronger commitment than the reference's "without overt edema." Checkpoint 800 drops those claims but states **"No focal airspace opacity"** and omits a possible left lower-lobe pneumonia and small left effusion. This case records the tradeoff between avoiding stronger unsupported commitments and retaining uncertain findings.

What the trajectory supports The factual GRPO reward projects the **final answer** into structured commitments and charges for unsupported additions and omitted reference facts. The report-generation task was absent from the training bank. The full-cohort comparison and fixed-case traces show that a more concise final-answer style appears early and persists when the policy is applied to complete reports. Study 51288835 shows substantial compression while retaining its principal findings; study 59067739 shows that the selection of findings can also introduce and remove errors across checkpoints. The reasoning traces do not show a monotonic sequence of improved internal fact review: the reward does not score reasoning text, and mean reasoning length rebounds at checkpoint 800.



Reference findings — MIMIC-CXR report text (the report prompt contained the image only)

- marked enlargement of the heart with mild pulmonary vascular congestion without overt edema
- Retrocardiac opacity with subtle air bronchograms could reflect left lower lobe pneumonia.
- Small left pleural effusion cannot be excluded.
- There is no pneumothorax or right pleural effusion.

	Thinking (CoT) — verbatim excerpts	Final report — verbatim excerpts
Base Qwen3.5-2B 994 chars	The left heart border is somewhat silhouetted, suggesting possible left lower lobe pathology or just mediastinal width. […] There is mild interstitial prominence, perhaps early pulmonary edema or just normal variation. […] The cardiac silhouette is significantly enlarged (cardiomegaly).	The cardiac silhouette is markedly enlarged, consistent with cardiomegaly. […] The lung fields demonstrate mild interstitial prominence, suggestive of mild pulmonary vascular congestion, but no focal consolidation, pneumothorax, or large pleural effusion is identified. **Editor note:** *CoT notes 'possible left lower lobe pathology', but the report states no focal consolidation; the reference's possible retrocardiac pneumonia and small left effusion are not addressed. Mediastinal widening, aortic calcification and oral contrast items are additionally reported, making this the longest report of the five (994 chars).*
Checkpoint 200 470 chars	The costophrenic angles are somewhat obscured, suggesting small pleural effusions or just atelectasis. […] The cardiac silhouette is significantly enlarged. It occupies more than 50% of the thoracic width.	There is marked cardiomegaly. The pulmonary vasculature appears congested. […] Small bilateral pleural effusions are suspected. **Editor note:** *The reference's possible retrocardiac pneumonia is not mentioned; the reference reports no right pleural effusion.*
Checkpoint 400 725 chars	There is a retrocardiac opacity obscuring the left heart border and diaphragm. […] Mild interstitial prominence, possibly mild pulmonary edema.	A retrocardiac opacity is noted, obscuring the left hemidiaphragm and left heart border, suggestive of left lower lobe atelectasis or consolidation. […] 3. Mild pulmonary edema. **Editor note:** *The reference's possible small left effusion is not addressed; the report excludes only 'No large pleural effusions'.*
Checkpoint 600 444 chars	There are diffuse interstitial markings, suggesting mild pulmonary edema or vascular congestion. […] The costophrenic angles are somewhat obscured but not completely blunted to the point of frank effusion, though there might be small pleural effusions.	The lungs demonstrate mild interstitial prominence and vascular congestion. […] 2. Mild pulmonary edema. […] 3. Small bilateral pleural effusions. **Editor note:** *The reference's possible left-lower-lobe pneumonia is omitted at this checkpoint.*
Checkpoint 800 357 chars	There are diffuse interstitial markings, suggesting pulmonary edema or vascular congestion.	There is marked cardiomegaly. The pulmonary vasculature demonstrates mild congestion. […] No focal airspace opacity or pneumothorax is identified. **Editor note:** *The CoT still weighs edema, but the final report drops it; 'No focal airspace opacity' rules out the reference's possible retrocardiac finding, and the possible small left effusion is omitted.*

■ reference-aligned ■ beyond / against reference […] omitted text *gray italic* = editor note (not model text)

Figure 7: A five-stage report trace for MIMIC-CXR study 59067739. The base model and checkpoints 200, 400, 600, and 800 receive the same radiograph and generation settings. Rows pair reasoning and final-report excerpts, showing changes in finding selection across checkpoints.

A paired, fact-level audit of existing reports would determine whether shorter answers remove unsupported commitments while retaining clinically relevant findings. Discordant cases need image review, and an independent extraction route would test whether the CFP favors either policy's writing style.

C.4 Visual evidence use in a cardiomegaly case

We examine how VeraRad-2B and its Qwen3.5-2B base model use image evidence for one report-generation case, MIMIC-CXR study 59067739. The radiograph has an independent MS-CXR annotation for "markedly enlarged cardiac silhouette." MS-CXR provides radiologist-verified image regions paired with finding descriptions. Boecking et al., MS-CXR v1.1.0 We selected this illustration after inspecting a frozen 32-case pilot; the full pilot is provided in the separately submitted supplementary attachment *Visual Grounding Casebook* (33 pages: one page describing the input conditions, followed by one page for each of the 32 cases). The casebook presents image-condition candidate scores, available attribution and intervention results, and correct-image versus no-image report excerpts from eight model configurations.

The analysis combines fixed-candidate attribution and image interventions with separately generated reports. The candidate probes disable thinking and use 238 visual tokens at reduced resolution; the report examples retain the original evaluation resolution and generation protocol. Both models share the frozen vision encoder and visual connector. VeraRad's language-model adapters were trained by Group Relative Policy Optimization (GRPO) with the Fact-Commitment Space reward described in

the main manuscript. Reference reports, candidate labels, and region annotations are excluded from the input prompt. Candidate sentences are supplied only as continuations for offline scoring.

C.4.1 Spatial attribution and local image interventions

We contrast the candidates $y^+ =$ "There is cardiomegaly." and $y^- =$ "There is no cardiomegaly.". Given image I, the common reporting prompt, and the assistant prefix `Findings:` , denoted together by x, we define

$$\ell_\theta(y \mid I, x) = \frac{1}{|y|} \sum_{t=1}^{|y|} \log p_\theta(y_t \mid I, x, y_{<t}),$$
$$C_\theta(I) = \ell_\theta(y^+ \mid I, x) - \ell_\theta(y^- \mid I, x). \tag{18}$$

Positive C favors the cardiomegaly candidate under this mean-token score; negative C favors its negation. We use the contrast to compare these two sentences under matched prompts and images.

VeraRad concentrates more positive attribution in the annotated cardiac region. With a blurred reference image, 40.6% of its positive attribution falls inside the box, compared with 31.5% for Qwen3.5-2B. The box covers 18.7% of the image, giving area-adjusted concentration ratios of 2.171 and 1.680, respectively. A gray reference gives the same direction (2.169 versus 1.980). We use the path-attribution method of Sundararajan et al.

Direct pixel occlusion shows a related spatial shift. VeraRad's largest score decrease (+0.130) comes from a block centered inside the annotation, with 75.3% of its area overlapping the box. Qwen3.5-2B's peak block yields a similar decrease (+0.131) and straddles the lower box boundary, with 47.6% overlap; its center lies 18 pixels below the box. On this 4×4 grid, VeraRad's strongest intervention thus falls more squarely on the annotated cardiac silhouette.

C.4.2 Response to masking and changes in the image input

We rank visual-grid positions by gray-reference integrated gradients, replace the corresponding image pixels, and recompute C. We plot $\Delta C_{\text{mask}} = C(I_{\text{masked}}) - C(I)$, so a downward response means that masking reduces support for the cardiomegaly candidate. We compare high-attribution, low-attribution, and five fixed random orders under both blur and gray replacement.

At 10.1% high-attribution masking, Qwen3.5-2B shows an initial reversal: its candidate contrast rises by +0.095 under blur and +0.100 under gray replacement. VeraRad's changes are smaller (+0.017 and -0.003), indicating a more stable local response to the first intervention. Once 25.2% of the highest-ranked cells are blurred, both contrasts fall (-0.353 for Qwen3.5-2B; -0.280 for VeraRad). Equal-area random masks instead raise the contrast by +0.079 and +0.040 on average. The high-attribution ranking therefore identifies regions that affect this cardiomegaly contrast in both models; VeraRad's distinctive result is the smaller initial reversal.

Both models allocate their highest measured image-token attention at layer 12: 64.6% for Qwen3.5-2B and 64.0% for VeraRad on the positive candidate. VeraRad retains a higher image-token share at layer 20 (61.5% versus 57.1%); both reach approximately 33.9% at layer 24. This deeper-layer difference complements the spatial and input-intervention results. The profiles cover the six full-attention decoder layers during candidate scoring; attention weights are an observational measure.

In Table 10, both models favor the cardiomegaly candidate with the correct radiograph ($C > 1.3$), and both scores fall when the image is exchanged for the donor radiograph whose report describes a normal cardiac silhouette. The drop is larger for VeraRad (1.111 versus 0.973), consistent with stronger reliance on the paired cardiac appearance in this comparison. With no visual input, VeraRad's candidate contrast is close to zero (-0.063), while Qwen3.5-2B favors the negative sentence (-0.511). VeraRad thus shows less text-only preference for a negative cardiomegaly claim in this case. Constant-color controls also move both scores toward zero; the blurred image retains coarse anatomical contours and produces a negative contrast in both models.

Table 10: Mean-token candidate contrasts under input controls, with thinking disabled. These scores concern the specified cardiomegaly pair only.

Input	Qwen3.5-2B	VeraRad-2B	Information retained
Correct	+1.3179	+1.3732	The original radiograph
Shuffled	+0.3445	+0.2619	Different patient; reference: unremarkable cardiac silhouette, chronic upper-lobe fibrosis
No image	−0.5107	−0.0631	Text prompt; no visual tokens
Black	−0.0905	+0.0587	Constant black image; visual grid retained
White	−0.2143	−0.1186	Constant white image; visual grid retained
Mean gray	−0.1450	+0.0230	Original mean brightness; visual grid retained
Gaussian blur	−0.4495	−0.1224	Coarse image structure and brightness

C.4.3 Generated reports: evidence selection and missing-image behavior

With the correct image and thinking enabled, both models report cardiomegaly. VeraRad delivers a compact Findings/Impression report, 274 stored characters against Qwen3.5-2B's 950. The reference describes mild vascular congestion without overt edema, possible retrocardiac pneumonia, and a possible small left effusion. Qwen3.5-2B escalates the congestion to likely pulmonary edema in its Impression. VeraRad avoids that assertion, while omitting the congestion and the possible pneumonia and effusion and concluding "No acute pulmonary process." Both models mention right-upper-quadrant oral contrast, which the reference does not discuss. VeraRad's concise report captures the annotated heart enlargement; its omissions remain clinically relevant to the full reference report.

Without an image, both outputs with thinking disabled assert normal findings. With thinking enabled in this case, Qwen3.5-2B asks the user to upload an image in conversational prose; VeraRad keeps the Findings/Impression structure and writes "The provided image is not available for review" and "Unable to evaluate."

The full 32-case pilot shows how often this behavior recurs. Each of the four configurations produced one saved no-image report per case. We classified the complete report text using the mutually exclusive categories in Table 11; the case-level audit in Appendix C.4.5 links every label to its original output.

Table 11: Report behavior without an input image, 32 paired cases per configuration. "Mixed" denotes a missing-image acknowledgment followed by patient-specific image findings; normal-finding examples explicitly labeled as templates are counted separately. A "No findings" Impression after a missing-image acknowledgment is kept as an ambiguous category.

Observed output	Qwen3.5-2B (Off)	Qwen3.5-2B (On)	VeraRad-2B (Off)	VeraRad-2B (On)
Findings/Impression refusal	0	8	0	9
Brief or conversational refusal	0	7	0	17
Image-based report without missing-image acknowledgment	32	10	32	5
Missing image acknowledged, then findings asserted	0	4	0	0
Normal-finding template marked as illustrative	0	2	0	0
Missing image acknowledged, then "No findings"	0	1	0	1
Total	**32**	**32**	**32**	**32**

With thinking disabled, both models write image-based findings in all 32 no-image cases. Enabling thinking raises explicit missing-image refusals to 15/32 (46.9%) for Qwen3.5-2B and 26/32 (81.3%) for VeraRad. Qwen3.5-2B (On) still gives 10 image-based reports without acknowledging the missing image and four responses that recognize the missing image before asserting patient-specific findings. Two further responses place normal-finding prose in examples labeled as templates, making 16/32 outputs with normal-finding prose overall. VeraRad-2B (On) gives five image-based reports without acknowledging the missing image and no patient-specific mixed responses; one additional Impression says "No radiological findings" despite noting the absent image. The paired outputs show a substantial shift toward evidence-aware refusal after training, with five direct unsupported VeraRad reports still present.

Across the seven eligible annotated cases in the pilot, both models place integrated-gradient peak centers inside the reference box in five. This case adds a finer view of differences in attribution density, pixel-occlusion response, and reporting style that a binary peak count misses. The supplementary

casebook contains all 32 records, including contrasting examples; Appendix C.4.4 documents the probes.

C.4.4 VISUAL PROBE METHODS AND NUMERICAL CHECKS

Cohort and model identity. The pilot contains 24 cases selected by a fixed SHA256 ordering, eight from each dataset in the frozen report triplet panel, plus all eight images in that panel with MS-CXR region annotations. These two groups do not overlap. The eight annotated cases comprise six cardiomegaly, one consolidation, and one atelectasis case. All eight receive spatial analysis; the other 24 receive candidate scores and retain historical reports. This is a small, unbalanced localization sample.

The compared models are Qwen3.5-2B and its VeraRad-2B policy after 800 GRPO steps with the Fact-Commitment Space reward, using the checkpoint-2434 Clinical Fact Parser (CFP). The visual encoder and visual connector are frozen; the policy changes are language-model LoRA adapters. In the new probes, the base and VeraRad configurations are obtained by disabling or enabling those adapters on the same loaded model. Encoder features were identical for the tested input. This visual connector is distinct from the CFP used during GRPO reward computation.

Exact image-path comparison with the actual GRPO bank found no overlap among the 32 selected images. One annotated case, mimic:50452688, shares a patient with a different training study; it remains visible but is excluded from the seven-case localization summary. The main-text case has no detected GRPO image or patient overlap. Pretraining and SFT exposure have not been ruled out.

Scoring target and input controls. The common input is a report-generation prompt asking for Findings and Impression, with one image where applicable. The stored clinical context, reference report, target label, and MS-CXR box are not given as prompt evidence. Each candidate is scored offline after the assistant prefix `Findings:`, with its preceding candidate tokens teacher-forced. The score is

The length-normalized candidate contrast is defined in Equation 18.

For the main case the pair is `There is cardiomegaly.` and `There is no cardiomegaly.` Natural logarithms are used. C is a difference of length-normalized log-likelihoods. It is not a log-odds ratio of whole-sentence probabilities, calibrated disease probability, clinical accuracy, or reward from the CFP reward. A near-zero contrast does not establish calibrated uncertainty. A negative cardiomegaly contrast does not imply an otherwise normal chest.

Seven conditions are retained: correct image; frozen shuffled donor image; no image; black; white; image-mean gray; and Gaussian blur with radius 0.08 times the shorter original image dimension. Constant images retain dimensions and visual-token positions but do not produce zero encoded features. Blur preserves low-frequency structure. A donor image can itself contain the target finding, so a shuffled score is not a diagnostic-sensitivity test without independent donor labels.

New candidate probes have thinking disabled and at most 262,144 image pixels. This case has a 17 $\times$ 14 grid of 238 merged visual tokens, corresponding to a 544 $\times$ 448 processed image. Its original image is 3,056 $\times$ 2,544 pixels (height $\times$ width). The input-control table uses BF16. Attribution, pixel occlusion, and guided deletion use FP32 with high matmul precision, allowing TF32 kernels. The unmasked contrast is therefore 1.317949 / 1.373237 (Qwen / VeraRad) in the BF16 table and 1.361213 / 1.393836 in the FP32 intervention run. Each decrease uses its own run's unmasked value.

Spatial attribution and localization metrics. Integrated gradients (IG) interpolates between the contextual post-connector visual features of a baseline image and the original image, while keeping the original token layout and text prefix fixed. It differentiates the candidate contrast and sums over feature channels at each visual position. Gray and blurred image baselines are both retained. Trapezoidal integration uses 16 intervals for both baselines in the main case; the wider pilot allows refinement up to 64 intervals. Gradient $\times$ feature difference is a separate first-order diagnostic. The feature positions already contain information mixed by the vision encoder, so their mapped locations are not isolated causal pixel regions. Integrated gradients reference

For an attribution array A and the union B of target-finding MS-CXR boxes, the metrics are:

$$\mathrm{Mass}(A, B) = \frac{\sum_{p \in B} \max(A_p, 0)}{\sum_p \max(A_p, 0)},$$

$$\mathrm{Concentration}(A, B) = \frac{\mathrm{Mass}(A, B)}{\mathrm{Area}(B)/\mathrm{Area}(I)}.$$

Peak-in-box records whether the center of the maximum-positive grid cell lies inside B.

Arrays are projected by nearest-neighbor resizing to the original image dimensions for mass calculation; coordinates use the original pixel frame. The concentration ratio is 1 for a uniform positive map. Negative attribution mass is excluded from this localization ratio but remains visible in the signed maps. Boxes are coarse clinical reference regions, not segmentations. MS-CXR annotations were verified by radiologists. Dataset documentation

Figure 5(a–f) and the original supplementary panels use a symmetric display scale per method, shared across the two models: the larger of their 99th percentiles of absolute map values. Color and opacity saturate beyond that threshold; reported map statistics use the original values. The original supplementary figure also displays attention and gradient $\times$ feature difference.

Coarse pixel occlusion. The original image is divided into a 4×4 grid. Each block is replaced in turn by the same coordinates from the globally blurred image, and the full image is re-encoded. The map stores C(original) minus C(altered). This directly intervenes on input pixels, though its response can include boundary artifacts and interactions with other regions.

For the main case, the MS-CXR box is x=[441,1991], y=[953,1892]. Qwen's maximum-drop cell spans x=[636,1272], y=[1528,2292], with center (954,1910), 18 pixels below the annotated box. VeraRad's maximum-drop cell spans x=[1272,1908], y=[764,1528], with center (1590,1146). The cells overlap the annotated region by approximately 47.6% and 75.3% of their own areas. The peak-in-box indicator is therefore sensitive to coarse cell placement and must be read with the continuous maps and overlap measurements.

Guided deletion. Grid positions are ranked by the original gray-reference IG map. We replace image pixels corresponding to the highest or lowest scores, or to five random permutations with seed 20260924. The same frozen orders are used at both mask sizes and for both blur and mean-gray replacement. Fractions round to 24/238=10.084% and 60/238=25.210% of cells. Random masks can intersect the annotated region, and high-attribution masks are not restricted to it. The original rankings are preserved when later IG numerical refinement affects another case; for this main case, the displayed and original IG arrays are identical.

Figure 5(g–h) plots $\Delta C_{\mathrm{mask}} = C(I_{\mathrm{masked}}) - C(I)$ on a common vertical scale, so masking-induced loss of positive-candidate support goes downward. The source JSON retains the opposite sign under its original `drop` field. Bands show ± 1 sample standard deviation across five random masks of this image. The lowest-ranked cells can have negative attribution; masking them can increase \$C\$. The curves sample two nonzero deletion fractions and show a smaller initial high-attribution reversal for VeraRad in this case.

Decoder attention. For each full-attention layer, we select the causal query immediately preceding each target token, average its attention weights over heads and target positions, and sum the weights assigned to image tokens. Positive and negative candidates are evaluated separately. The displayed layers are 4, 8, 12, 16, 20, and 24. The remaining 18 layers use linear attention and are not included in these attention matrices. This is attention within the joint multimodal decoder sequence, not a separate cross-attention module and not an all-layer attention rollout.

Table 12 gives the positive-candidate percentages.

Attention is an auxiliary descriptive measurement. These percentages do not represent the causal fraction of a prediction attributable to vision. The teacher-forced OFF candidate curves also do not explain the complete ON generation trajectory. Attention interpretation reference

Numerical checks and generated reports. The initial BF16 directional finite-difference check failed. Attribution was consequently evaluated with FP32/high-matmul precision; subsequent local

Table 12: Image-token attention weight (%) for the positive cardiomegaly candidate.

Model	L4	L8	L12	L16	L20	L24
Qwen3.5-2B	32.29	42.69	64.65	57.64	57.11	33.91
VeraRad-2B	32.34	42.20	64.04	59.35	61.53	33.87

directional checks passed for both models. Native forwarding and the feature-insertion scoring path agreed in the validation case. These checks verify local implementation behavior, not every possible input.

Final IG acceptance requires absolute completeness error $\leq$0.01 or relative error $\leq$10%, using a 0.05 denominator floor. For the main case, gray/blur absolute errors are 0.00185/0.01111 for Qwen and 0.00677/0.01321 for VeraRad; corresponding relative errors are 0.12%/0.61% and 0.49%/0.87%. All four maps pass. The complete pilot retains original maps, refinements, and failed checks.

Historical reports comprise eight inference configurations $\times$ three image conditions $\times$ 32 cases = 768 outputs, copied from the frozen 300-study evaluation. They use the original report-generation pipeline with a 4,096-token output cap and a 1,024-token thinking budget when enabled; native image processing retains substantially more visual tokens than the candidate probes. The twelve Qwen/VeraRad ON/OFF reports for the main case all ended with `stop`. The report panel in Figure 5 uses stored final-answer text; the casebook provides report excerpts and identifies omissions. Original responses and reasoning remain in the source records.

The missing-image report audit in Appendix C.4.5 classifies all 128 blind outputs from the four Qwen/VeraRad ON/OFF configurations. All ended with `stop`. Labels are assigned to the complete answer text, not the reasoning field. A patient-specific anatomy or disease assertion is an unsupported report; an explicit missing-image acknowledgment followed by such an assertion is mixed. Two normal-finding examples explicitly labeled as templates receive their own category. We separate the ambiguous "No findings" Impression from "No radiological findings can be documented," which states that assessment is unavailable. Refusals with Findings and Impression headings are structured; the rest are brief or conversational. The archived machine-readable audit retains every case label, report hash, and source-file hash.

The main case does not have a new low-resolution free-generation or actual-output replay experiment. Its heatmaps target the offline candidate, not the exact generated report sentence. Those additional experiments were run on three other fixed cases and are retained in the source records: 120 requests under a separate, smaller native-HF budget, of which 115 stopped normally and five reached the length limit. Their scores are not mixed with the original report benchmark.

Scope of the evidence. The original findings reference for study 59067739 also includes possible retrocardiac pneumonia and possible small left effusion. The concise VeraRad report omits these concerns and concludes that there is no acute pulmonary process. Cardiomegaly localization therefore cannot validate the entire report. Right-upper-quadrant oral contrast appears in both models' ON outputs but not in the reference text; its correctness was not independently adjudicated in this analysis.

The complete pilot includes counterexamples, including a VeraRad ON report that asserts normal findings without an image, and a new white-image response that generates normal findings despite rejecting a black image. Across the seven annotated cases after excluding known GRPO patient overlap, IG peak-in-box counts are 5/7 for both models.

C.4.5 Case-level audit of missing-image reports

All 128 stored no-image reports from four configurations on the 32 cases were read in full; all ended normally. Table 13 gives their mutually exclusive labels, and Table 11 gives aggregate counts. Labels concern the complete final answer, not its reasoning trace. The separate 33-page casebook provides the corresponding case records.

A patient-specific anatomy or disease statement makes a no-image report unsupported. An explicit missing-image acknowledgment followed by such a statement makes it mixed. Normal-finding examples explicitly marked as templates remain separate. "No radiological findings can be documented" describes inability to assess and counts as a refusal; "No findings" or "No radiological findings" in

the Impression is classified separately as ambiguous. Refusals with both Findings and Impression headings are structured; other refusals are brief or conversational.

Table 13: Case-level no-image report labels. S: structured refusal; R: brief or conversational refusal; U: image-based findings without missing-image acknowledgment; M: acknowledgment followed by findings; T: explicitly illustrative normal template; A: acknowledgment followed by ambiguous "No findings."

Source	Study/case	Qwen Off	Qwen On	VeraRad Off	VeraRad On
CheXpert Plus	patient64548_study1	U	R	U	R
CheXpert Plus	patient64576_study1	U	U	U	S
CheXpert Plus	patient64580_study1	U	U	U	S
CheXpert Plus	patient64624_study1	U	U	U	R
CheXpert Plus	patient64633_study1	U	R	U	S
CheXpert Plus	patient64637_study1	U	S	U	R
CheXpert Plus	patient64648_study1	U	R	U	R
CheXpert Plus	patient64681_study1	U	T	U	R
IU X-Ray	CXR1344_IM-0223	U	A	U	R
IU X-Ray	CXR1413_IM-0263	U	S	U	R
IU X-Ray	CXR1860_IM-0558	U	S	U	S
IU X-Ray	CXR2168_IM-0784	U	U	U	R
IU X-Ray	CXR2336_IM-0903	U	R	U	S
IU X-Ray	CXR2743_IM-1197	U	S	U	U
IU X-Ray	CXR2916_IM-1318	U	R	U	R
IU X-Ray	CXR3020_IM-1395	U	S	U	R
MIMIC-CXR	50452688	U	U	U	R
MIMIC-CXR	50535882	U	U	U	R
MIMIC-CXR	50545797	U	U	U	S
MIMIC-CXR	50910303	U	U	U	A
MIMIC-CXR	51288835	U	M	U	R
MIMIC-CXR	51467319	U	M	U	R
MIMIC-CXR	53225676	U	M	U	R
MIMIC-CXR	53887723	U	S	U	R
MIMIC-CXR	53913561	U	U	U	U
MIMIC-CXR	57544155	U	T	U	R
MIMIC-CXR	57586513	U	R	U	U
MIMIC-CXR	58068113	U	M	U	S
MIMIC-CXR	59067739	U	R	U	S
MIMIC-CXR	59249240	U	U	U	U
MIMIC-CXR	59284918	U	S	U	S
MIMIC-CXR	59748962	U	S	U	U